CHURN PREDICTION USING CUSTOMERS' IMPLICIT BEHAVIORAL PATTERNS AND DEEP LEARNING

by
ANEELA TANVEER

Submitted to the Graduate School of Business
in partial fullfilment of
the requirements for the degree of Master of Business Analytics

Sabancı University
July 2019

CHURN PREDICTION USING CUSTOMERS' IMPLICIT BEHAVIORAL PATTERNS AND DEEP LEARNING

Approved by:

Prof. Dr. Burçin Bozkaya
(Thesis Supervisor) .....

Assoc. Prof. Dr. Selim Balcısoy .....

Assoc. Prof. Dr. Enes Eryarsoy .....

Approval Date: July 19, 2019

CHURN PREDICTION USING CUSTOMERS' IMPLICIT BEHAVIORAL PATTERNS AND DEEP LEARNING

Approved by:

Prof. Dr. Burcin Bozkaya .....
(Thesis Supervisor)

Assoc. Prof. Dr. Selim Balcısoy .....

Assoc. Prof. Dr. Enes Eryarsoy .....

Date of Approval: July 19, 2019

ABSTRACT

CHURN PREDICTION USING CUSTOMERS' IMPLICIT BEHAVIORAL PATTERNS AND DEEP LEARNING

ANEELA TANVEER

Masters Business Analytics MSBA THESIS, MAY 2019

Thesis Supervisor: Prof. Burcin Bozkaya

Keywords: churn prediction, deep learning, behavioral patterns, sequence modeling, recurring neural network, Graph network

The processes of market globalization are rapidly changing the competitive conditions of the business and financial sectors. With the emergence of new competitors and increasing investments in the banking services, an environment of closer customer relationships is the demand of today's economics. In such a scenario, the concept of customer's willingness to change the service provider – i.e. churn, has become a competitive domain for organizations to work on. In the banking sector, the task to retain the valuable customers has forced management to preemptively work on customers data and devise strategies to engage the customers and thereby reducing the churn rate. Valuable information can be extracted and implicit behavior patterns can be derived from the customers' transaction and demographic data. Our prediction model, which is jointly using the time and location based sequence features has shown significant improvement in the customer churn prediction. Various supervised models had been developed in the past to predict churning customers; our model is using the features which are derived jointly from location and time stamped data. These sequenced based feature vectors are then used in the neural network for the churn prediction. In this study, we have found that time sequenced data used in a recurrent neural network based Long Short Term Memory (LSTM) model can predict with better precision and recall values when compared with baseline model. The feature vector output of our LSTM model combined with other demographic and computed behavioral features of customers gave better prediction results. We have also

proposed and developed a model to find out whether connection between the customers can assist in the churn prediction using Graph convolutional networks (GCN); which incorporate customer network connections defined over three dimensions.

ÖZET

ÜSTÜ KAPALI MÜŞTERİ DAVRANIŞ BİÇİMLERİNİ KULLANARAK KAYIP MÜŞTERİ TAHMİNİ VE DERİNLEMESİNE ÖĞRENME

ANEELA TANVEER

İş Analitiği Masterı MSBA Tezi, Mayıs 2019

Tez Danışmanı: Prof. Dr. Burcin Bozkaya

Anahtar Kelimeler: Kayıp Müşteri Tahmini, Davranış Biçimleri, Zaman Bazlı Sıralı Modelleme, Yinelenen Sinir Ağı, Grafik Ağı

Günümüz pazarının küreselleşme süreci, iş ve finans dünyasının rekabetçi koşullarına göre hızla değişmektedir. Banka hizmetlerine yapılan yatırımlar ve yeni rakiplerin ortaya çıkmasıyla beraber, yakın müşteri çevresi de günümüz ekonomisinin talep etmektedir. Böyle bir durumda, müşterinin hizmet sağlayıcısını değiştirme isteği kavramı, organizasyonlar için rekabetçi bir çalışma alanı haline gelmiştir. Bankacılık sektöründe mevcut müşterileri koruma görevi, yönetimlerin öncelikli olarak müşteri verileri üzerinde çalışmasını ve müşterileri bağlayacak ve kayıp müşteri oranını azaltacak projeler yaratmalarını mecbur hale getirmiştir. Müşteri işlemleri ve demografik veriler ile değerli bilgiler ortaya çıkarabilir ve aynı zamanda davranış biçimleri hakkında kanıksamalar yapılabilir. Bizim öngörü modelimizde birleşik olarak kullanılan zaman ve yer tabanlı önergeler, kayıp müşteriyi önceden kestirme konusunda önemli gelişmeler kaydetmiştir. Geçmişte gözleme dayalı çeşitli modeller geliştirilmiştir, bizim modelimizde ise mühürlü yer ve zaman verilerden bileşik olarak elde edilmiş özellikleri kullanmaktadır. Söz konusu dizgi tabanlı veriler, kayıp müşteri tahmininde kullanılmaz üzere, vektörler halinde sinirsel ağda kullanılmıştır. Bu çalışmada; tekrarlı sinirsel ağa dayalı Uzun Süreli Bellek (USB – İngilizce LSTM) modeli içerisinde kullanılan zaman sıralı verinin, ilk modellere kıyasla daha hassas tahminler yaptığı ve daha fazla değer ortaya çıkardığı bulunmuştur. USB modelinin vektörel çıktılarının, diğer demografik ve müşterilerin dijitalleştirilmiş davranışlar modelleri ile birleştirildiğinde, daha iyi tahmin sonuçları verdiği görülmüştür. Ayrıca bu çalışmada, müşteri ağı bağlantılarını üç boyutta kapsayan grafiksel evrişimli ağ (GEA – İngilizce GCN)

kullanarak yapılan kayıp analizlerinin, müşteriler arasında bir bağlantı olup olmadığını anlamada yardımcı olacak bir model geliştirip önerdik.

ACKNOWLEDGEMENTS

First and foremost, I am thankful to Allah for the all the Blessings and giving me the strength to continue my education after a long gap.

I would like to express my sincere gratitude to my advisor Prof. Burcin Bozkaya for his invaluable guidance and mentoring in my study. I am thankful for his support and giving me the opportunity to work at Fondazione Bruno Kessler (FBK) in Trento, Italy as a Visiting Researcher with the research team at the Mobile and Social Computing Lab (Mobs). I would like to acknowledge the suggestions and guidance given by Dr. Bruno Lepri and his research team especially Yahui Liu and Gianni Barlacchi at FBK for my thesis work.

I am grateful to Higher Education Commission, Pakistan for initiating a scholarship program for employees and award of scholarship to me. I hope that this program will continue in the future for the employees' support and development. My deepest gratitude to my mentor Mr. Anwar Amjad, who has been always a source of motivation and guidance for me to move forward and learn passionately.

I am sincerely grateful to my parents, Naseem Akhter and Ghulam Jilani for their love, endless prayers and guidance throughout in my education and profession. They are always a source of inspiration for me. I appreciate and thankful for the support of my dear husband Tanveer Ali during the study. I am also thankful to my brothers, Naveed and Adeel and my sisters Neelum and Saima for their support.

I have joyful memories of the time spent in Turkey and in Sabanci University. And I would like to say special thanks to my dear friends especially Atia Shafique, Sumeyye Cangal, Sumaiyah Najib, Aasmah malik, Sanaullah, and Shah for their continuous encouragement and support during my studies. I am blessed with your friendship and doing master's would not be enjoyable without you.

Dedication
to my parents

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Customers have always been a vital part of the service industry and their retention is one of the major organizational challenges in today's competitive economic environment. Organizations need to devise sustenance plans to develop their business stability. Churn is defined as the tendency for customers to defect or cease business with a company (Kamakura et al. [2005]). And with the passage of time, consumer churn is one of the major problems that the service providers are facing in rapidly changing markets that are extremely competitive. With the emergence of Big Data and Business Analytics, a new paradigm is apparent; which is how to use IDA (Intelligent Data Analysis) based decision making strategies targeting proactively to avoid customers' churn decisions. Hence, churn prediction may be possible by applying effective analytic tools on existing data and avoid churn before it happens.

Following the Pareto principle of economic theory, a large number of customers contribute less to the revenue in contrast to the small slice of customers who have a major contribution [Chiang et al., 2003]. Organizations work continuously on evolving procedures to help and retain their profitable customers [Lejeune, 2001] and also in building and sustaining loyalty at the individual customer level [Kumar and Shah, 2004]. The accessibility of data about customers acquired from multiple sources and availability of exponentially growing computational power has enabled us to efficiently process huge volumes of data and thus making it possible to predict a customer's next action. Discovery of knowledge from large sets of databases is defined as taking out valuable information from the large volume of exponentially growing data [Fayyad et al., 1996].

Generally, causes of churn can be classified as voluntary or involuntary [Spanoudes and Nguyen, 2017]. Voluntary churn is, when it is the customer's own decision to change a service provider for a particular service, which can be due to the unsatisfactory levels of technical or service support, agreed commitments, services rates, or competitive offers from other providers of the same business. On the other hand, churn can occur due to some uncontrolled or incidental reasons like customer's relocation to a place where the

current service provider is not operating or it may be due to the customer's own financial crash that he or she cannot continue to avail that service further long.

Data mining techniques are the main support to get valuable insights into patterns of customer behavior using customer related data, which are available and integrated from multiple channels. Data mining techniques employ data modeling approaches for problems such as association, classification, clustering, sequence discovery, and work with mining algorithms such as decision trees, random forests, neural networks, to extract the valuable information hidden in the massive data sets. Customer churn is clearly a major type of business problem tackled by such data mining approaches, which is faced in many domains such as telecom [Khan et al., 2015], banking [Xie et al., 2009], media and gaming [Kawale et al., 2009].

This thesis pertains to the churn prediction of bank customers using the customers' demographic and transaction data. The implicit features derived from demographics and banking transactions data are used to predict customer behavior and their upcoming decision to leave the service provider. In this thesis, we employ deep learning methods to improve the customer churn prediction rate reported in the literature and comparisons are made with the traditional classification methodologies. For this study, we use demographic data along with the one-year transaction record of customers including online or offline transactions. This data set has also the Bank's own classification labels for identification of those customers who are going to leave the bank, which are used in the development and confirmation of our model and its performance. The Bank has defined more than a dozen definitions for a churning customer based on their communication through the Bank's call center, interaction with the banking channels, customer response pattern and services acquisition. And for our study, we have taken customer status marked as inactive for three months as an indicator of a churning customer. To derive the behavioral features we used customers demographics and credit card transaction records in our prediction model. Two deep learning (DL) methodologies are used in this study. First, we use DL for the identification of transaction sequence patterns over the time and then in feature vectors along with other demographic and calculated behavioral features to predict the churning customers. Secondly, we employ a graph-based deep learning method that uses graph features, which describes the commonality and connections between customers, merchants, shopping spending categories and common proximity information due to money transfer activities. For the sequence prediction model, features are calculated while taking the quarterly transaction records over an year. Quarterly values are used to find the sequence pattern in the customers' transactions, and the internal state values are used in conjunction with the

demographic and behavioral features in a neural network for churn prediction.

The main contributions of this study are:

(i) Identification of the sequence patterns in the transactions made by customers over a time span and then using these patterns for churn prediction.

(ii) Use of graph features and a network model for churn prediction.

This thesis is organized as follows: Chapter 2 covers the literature review on churn and its management, implicit behavioral patterns derived from the data and the approaches used in churn prediction. In Chapter 3, we describe the details of data collection, data preprocessing and feature extraction. Chapter 4 includes details of the prediction modeling techniques used. A discussion and comparison of the results are presented in Chapter 5, which is followed by Chapter 6 with the concluding remarks, contributions and a summary of the work done.

2. Literature Review

In this chapter, a literature review on churn and churn management, behavioral patterns and features extracted from the data and the techniques used for churn prediction are presented.

2.1 Customer Churn and Churn Management

Churn is defined as "the tendency for customers to defect or cease business with a company" by Kamakura et al. [2005]. Glady et al. [2009] have labeled churn as a strategic action to be taken by marketing division to retain a customer, after knowing that a customer is going to leave the company in the near future. Churn rate can be explained with two points of view: first the number of customers leaving a company; and second the revenue amount a company is losing.

Generally, while considering the number a company is losing, the causes of churn can be classified as voluntary or involuntary [Spanoudes and Nguyen, 2017]. Voluntary churn is when the customer makes a decision to leave and/or change the service provider, which can be due to the unsatisfactory levels of service support, non-fulfillment of agreed commitments, abrupt changes in the rates of service, or lagging in the attractive offers as compared to the other players of the same business. On the other hand, churn behavior can occur due to some uncontrolled or incidental reasons like customer's relocation to a place where that particular service provider is not operating or due to the customer's own financial crunch such that he or she cannot continue to avail that service further long.

Customers have always been an imperative entity in the services domain and their re-

tention is one of the strategic tasks of an organization for its business stability and sustenance in today's competitive environment. Organizations functioning on life-long and financially solvent strategies apprehend the customer retention [Kamakura et al., 2005]. And focusing on this critical issue suggests that an organization should be well-aware and capable of determining and further managing these factors which cause churning. Organizations attempt to maintain their customer balance in the progressive market according to Lee et al. [2001], which varies with the entrance of emerging competitors who are equipped with innovative service offerings. In such an environment, one of the tactical points is to retain the customers with less investment than a new customer acquisition [De Chernatony, 2010], which is an indicator of reliance on the performance. The sustenance plans to retain customers bank and finding appropriate dimensions that can reduce customer turnover leads to churn management [Hadden et al., 2007]. Churn management requires an appropriate amount of related data and application of right analytics techniques to exploit information from the available (integrated or non-integrated) data sets [Kamakura et al., 2005][Lejeune, 2001].

Application of appropriate data analytics techniques enables firms in successfully retaining their consistent and valuable customers with a higher level of their satisfaction, and this practice is termed as churn management. The churn prediction efficacy and contribution in business sustenance doesn't depend only on the correct prediction of customers who are going to leave in a near future, but also on other aspects, like market saturation and competitiveness [Datta et al., 2000]. The advocates of customer retention policies termed churn management as a key financial gain for the firms rather than just struggling for the new customers in the market [Seng and Chen, 2010]. A similar point was emphasized by Kotler and Armstrong, that although customer pull is also vital, retention of customers is of prime importance, as it leads towards the lifetime loss from that particular customer for the firm [Hair Jr et al., 2010].

The accessibility of customer data from multiple sources and the availability of exponentially growing computational power has enabled us to handle and dive deep into big data. Big data analytics has made it convenient to predict a customer's next action, with the assistance of emerging data mining and machine learning tools. The discovery of knowledge from the large sets of databases is defined as extracting valuable information from the large volume of exponentially growing data [Fayyad et al., 1996].

In the past, research has been conducted to predict the customer and organization churn rate in different business domains such as the banking sector [Xie et al., 2009] [He et al., 2014] [Oyeniyi et al., 2015], telecommunication industry [Khan et al., 2015] [Verbeke et al., 2014] [Lemmens and Croux, 2006], gaming [Kawale et al., 2009], insurance companies

[Günther et al., 2014] and many others. As such many conventional techniques have been employed to predict the churning of customers using demographic and domain specific data.

2.2 Implicit Behavioral Patterns and Features

In-depth analysis of human behavior has gained much importance with the advent of more diverse and complicated learning algorithms; and this analysis has eventually revealed the causes of decisions which a person is up to and so it helps in the future business planing of organizations. The concept of behavior informatics or behavioral computing was presented by Cao [2010] in his paper, which attempts to present computational tools and technologies for the deep understanding of behavior from the related social networks data.

Researchers in the past have identified and used multiple attributes which are engineered from the accessible data to reveal the implicit behavioral attributes and patterns the customers' exhibit. Some of the behavioral aspects which are identified and proven to be the integral predictors for churning customers include recency, frequency and monetary commonly known as RFM and usually used and defined in combination, features calculated with the division on location and time (known as spatio-temporal features), customer segmentation or profiling and customer lifetime value.

For the extraction of behavioral features, mandatory step is the definition of features, which pertains to a specific service domain and then the identification of the sources of data and the possible linkages between the data sources (if available). Then comes the feature acquisition phase which is accomplished by the data transformation process to calculate the features from single or multiple data-space[Cao, 2010].

Below are some of the data dimensions which help to extract and define the human behavioral values from the underlying services data and their usage.

2.2.1 Recency, Frequency, Monetary (RFM) Based Features

Three key variables that are commonly stated together and are used to define and analyze the customer behavior; are recency, frequency and monetary value (RFM). Recency is the time interval since the last purchase or transaction is made; frequency is the number of purchases made in a specified time window and monetary value is the amount spent during a specified time window [Wang, 2010]. The patterns and/or the thresholds of these three attributes usually infers the predictions for customers' behavioral patterns. Organizations use thresholds against these attributes for churn management and definition of a customer's retention rate. These three variables work as the key observatory of the customers' drive or behavior towards a particular product or service, brand, service gains, and reliance levels [Wei et al., 2010]. Moreover, Hadden et al., in their research have discussed that exploration and selection of appropriate RFM variables from a transaction data set can lead to a confident definition of customer behavior and subsequently prediction of churning customers; which can be achieved by a thorough and methodical understanding of the data semantics and context [Hadden et al., 2007]. In a related research, Martens et al., have shown that the detailed attributes of transaction data contribute well to the prediction performance along with the variations of recency, frequency and monetary (RFM) attributes that are traditionally being used. The RFM attributes can be engineered and transformed to get the maximum information about the customers' behavior [Martens et al., 2016].

2.2.2 Spatio-temporal and Choice Patterns Based Features

Kaya et al. [2018] have analyzed and used financial transactions data to determine customers' behavioral patterns and traits, which leads to better churn prediction. They have defined new attributes i.e. diversity, loyalty, regularity and choices made by the customer for making financial transactions; which they have characterized by them as spatio-temporal features. These spatio-temporal features are based on the location and time stamp of transactions made by customers. These derived features help to measure and analyze the behavioral traits of a customer, as how regularity, loyalty and diversity traits of customers tend to shift when customers while using a services from a bank, churn or tends to churn. Customer purchase patterns tend to change with the passage of time

and this change is observed with the change in the shopping behavior while a customer is used to purchase from a specific place, specific brand or merchant. And this change can be observed from the change in the customers shopping time.
Spatio-temporal features with slightly different variations were also used by Singh et al. [Singh et al., 2015]; to illustrate the financial well-being of customers defined using three behavioral indicators, namely overspending, trouble and late payments, which can be suggested for the churn prediction as a future work. Xie et al. [2009] have employed customer demographics, account status and credit card usage details to gain information about a customer's behavior. Account and credit card usage status and details of various service(s) agreement/contracts with the bank were used in their research to make inference about the customer's behavior towards churn.

2.2.3 Features Based on Customer Similarity Scoring

Martens et al. [2016] have introduced behavioral similarity measures based on the payment(s) made by different customers to the same entity/entities. They have assigned weights are assigned to customers who share more entities, however, the entities or merchants who are more popular or of common services get less weight. The customer's behavioral score, which is the sum of weighted values provides the measure of similarity of the customers. And this similarity measure helps with the identification of groups of customers who are going to churn together. For future research, they have identified that this similarity score measure can be used to categorize alike customers and predict their behavior, while extending the same measurement for the customers who make transactions using a credit card. Martens et al. in their research have identified that the use of the transaction details which were made to specific merchants by customers along with the structured data (i.e. demographics and transaction data of customers), can improve the prediction results significantly.

Behavioral scoring can also help the decision makers to recognize their customers' value and Hsieh [2004] has proposed a two-staged approach for the behavioral scoring by taking the transaction and bank account data of customers. The first stage identifies the grouping of customers based on recency, frequency and monetary (RFM) values and then the payment history of customers is used to profile the customers in different groups.

2.2.4 Understanding of Customer Behavior using Assessment Tools

Keramati and Ardabili [2011] have defined customer satisfaction as "an experience-based assessment that stems from the degree to which customer expectations about characteristics of the service have been fulfilled". Their study is based on the telecommunication services data and they have used customer demographic data, call detail records, length of duration since a customer is associated with the service provider, count of logged complaints (taken from call center data) for the churn analysis. Based on the study they conclude that the number of a customer's complaints has a major contribution in churn prediction along with the service failure incidents which are reported by the customers. Based on this study, it can be concluded that the call center and/or data from the customer relationship management (CRM) tools helps to get information about customer satisfaction level, which can be used along with other attributes for better analysis and prediction of churning customers. To assess customer satisfaction value and then linking it with customer retention, a mathematical model was developed by Rust and Zahorik [1993] for retail banking. They presented a concept where the retention is linked with customer satisfaction which is further linked to customer loyalty, their overall retention rate and the market share of an organization. They have carried out experiments to identify the elements which highly impact the customers' retention rates while considering their satisfaction level. They have suggested that an important component for the market share is the customer retention rate which can be assessed and controlled using the attributes and values of customer satisfaction.

Likewise, mining of customer demographics and transaction-based data can give valuable comprehensions and insights into customer behavior and their pattern variability to foretell their future working and association with an organization. The aim of the churn prediction models is to preemptively identify the customers, who have the churn tendency in the near future and then devise customer-specific campaigns and measures to retain them or to minimize the churning attitude of alike risky customers showing similar patterns.

2.3 Churn Prediction Methodologies

As discussed in section 2.1, customer retention or churn management is vital for the business sustenance plan. In the past, multiple statistical and data mining methods like logistic regression, decision trees, random forest, support vector machines, artificial neural networks, to name a few, have been used for churn prediction. These models have given reasonable results depending on the services domain and the problem definition, though the data set(s) and features that are derived and used in these prediction models vary.

Decision trees are one of the popular and powerful machine learning techniques and are used extensively for both classification and regression based problems. Datta et al. [2000] have used decision trees to predict customer churn for the telecommunication industry and developed a model called Churn Analysis Modelling and Prediction. Euler [2005] has also used decision trees for the churn analysis and predicting customer churn behavior using five months of call-data of customers. They have derived features from the temporal data and also used aggregated features for each month which help them in developing a better model to predict when and which type of customer is going to churn.

Xie et al. [2009] have employed improved balanced random forest (IBRF) learning technique to predict churning customers of the bank while using three data descriptors of customers which are demographic (including age, education, income, family status), account level (account type, loan details) and customer behavior (measured using account and credit status). Balanced random forest via over-sampling the minority class helps to deal effectively with the imbalanced data. For their experiments, they have proposed integrated sampling to maintain sample distribution and randomization taken from the balanced random forest, and assigning weights to the minority class observations. Lift curve and top-decile lift were used as the evaluation criteria in their experiments and results were compared with artificial neural network (ANN), decision tree (DT) and class-weighted core support vector machines (CWC-SVM) showing that their work outperforms these other models.

Kaya et al. [2018] have developed a model using a random forest classification technique for the churn prediction of bank customers. Three features, namely diversity, loyalty and regularity were computed by them to assess the behavioral patterns of customer and then used along with the demographic attributes to predict the churning customers. Prediction results were convincing while using random forest for binary classification, proving that dynamic behavioral patterns contributed well in the churn prediction. They have defined customer behavior features while using the multiple scales for location and date/time stamped data from transaction records. Diversity, loyalty, regularity and choice pattern for fund transfer and purchase transactions are the featured attributes calculated from the transaction data of customers.

He et al. [2014] have applied support vector machine (SVM) for bank customer churn prediction, while using a random sampling method to deal with the imbalanced data. Three models which are logistic regression, linear SVM and SVM with radial basis kernel function (RBF) were compared with different class ratios. Experiments results reveal that SVM with RBF comes out as a better prediction model based on recall and precision values as the evaluation metrics. The capability of timely prediction of customer attrition by using their model has facilitated the bank to take proper and well-timed measures against churning customers. Their work described that SVM can be a better alternative to logistic regression.

Goal oriented sequential pattern algorithm was proposed by Chiang et al. [2003] for the identification of customers who are going to churn in the near future. The definitions of their association rules come from the sequential patterns which are observed over a time period. This research has mined the sequential patterns to find deviations in the behavioral patterns of churning customers using the association rules.

Burez and Van den Poel [2009] have worked on improving the prediction accuracy of churn prediction while focusing on the class imbalance issue in the customer data and using gradient boosting and weighted random forest algorithms. They have addressed this highly impacting rare event of churn by using multiple variants of the data sampling techniques (random and under-sampling) and suggested that the models with cost-sensitive learning give better performance. They performed experiments with six data sets including two banks, telecommunication, newspaper, television subscription services and a supermarket. Mutanen et al. [2006] have predicted customer churn in the retail bank sector with logistic regression while using the under-sampling method to handle the class-imbalance. Along with churning customer prediction they have calculated the customer value, which helped them in taking a decision whether to retain a customer or not. They have also studied how much data-duration is suitable to study a customer before making churn prediction.

Prasad and Madhavi [2012] have used classification and regression tree (CART) and C5.0 classification techniques for churn prediction by modeling the purchasing behavior of customers who are savings account holders. And they have concluded CART as a better prediction for the churn class, that eventually help bank managers to devise a strategy to retain their valuable customers rather than losing revenue in the future. Binary classification using the linear discriminant Boosting algorithm was proposed by Xie and Li [2008] for the churn prediction with an excessively imbalanced data set. With the usage of linear discriminant technique, discriminative features are computed in each iteration and a heavier penalty is imposed for each misclassification of the minority class

while using the boosting technique, eventually giving more precise results. They have compared their work with other methods like Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machine (SVM) and the classical Adaboost by measuring accuracy from the top-decile lift and lift curve evaluation method.

Genetic algorithm based neural network has been developed and used by Pendharkar [2009] for predicting the customers who are availing wireless cellular services and are most likely to churn in the near future. Pendharker has carried out the experiments with a medium sized neural network architecture, and used the ROC curve as the evaluation criteria for prediction. He has shown that the genetic algorithm based neural network outperforms the zscore based prediction model.

Most of the prediction problems have data with the date and time stamp and location details; like data from telecommunication industry having call records and balance pre or post-payment information, banking data is with the transaction date and time along with the mode of payment such as online or offline and/or services taken of bank used, medical diagnostics having multiple test results record (count of tests) spanned over the time. Prediction of such time series-based sequential data has been in great focus since the emergence of machine learning algorithms. In this setting data from an observation window can be used to make a prediction for the subsequent windows while keeping track of multiple data dimensions in the observation window. The definition and duration of the observation window and prediction window depends on the nature of the problem and data availability. Single or multiple activities are monitored during the observation window, and the resultant feature vectors are used in a suitable algorithm for prediction.

Mining data over time pertaining to customers has increased the prediction accuracy while using the behavioral features, which are calculated using the time based data. The changes in the patterns can be identified by measuring the similarities or deviations in behaviors at different time spans as proposed by Chen et al. [2005] who propose measuring these changes using the definition of the association rules. Deviation in the customer behavioral profile and services acquired leads to the future prediction of customer behavior. The measure of similarity and the unexpected event occurrence can help to gauge the deviations in the customer behavior over some defined time span. Hybrid data mining techniques have also been employed by researchers for better churn prediction. Tsai and Lu [2009] have put forward two hybrid models for better prediction performance. Both of their proposed models, which are Artificial Neural Networks (ANN) and Self Organizing Map (SOM) (ANN + SOM) and two artificial neural network (ANN + ANN) models, outperform a single algorithm based model and the baseline model here is the single neural network based model. However, the selection and usage of a single mining algorithm

or combination of techniques depend on the size of available data and the problem domain. With the hybrid data mining techniques used in their study, hidden patterns and data relationship are discovered using a clustering technique (which can be considered as data preprocessing step) and then the resultant vector are used for prediction.

Extraction and use of time series data with the similarity based classification using similarity forest method, as proposed by Óskarsdóttir et al. [2018], has given competitive prediction accuracy in comparison to other traditional classification methods. The dynamic temporal feature based networks are presumably more representative of the real world activities and they help in prompt planing and decision making. Mainly there are two approaches to deal with dynamic temporal networks the first deals with time dependent network structures where the same features are calculated against different time marks. And the second is about building a time-based network using the aggregated values of features calculated over a specific time span. In their work, María et al. have adopted similarity forest methodology while using time series data for the behavior patterns representation of telecommunication industry customers and early detection of potential churners. Similarity forest is the extension of random forest, in a way that it constructs multiple decision trees and node splits are done based on the similarity between the node objects. The similarity between the objects is marked and at each split point, this similarity flag is taken into account. Considering a binary classification problem, each observation is also labeled as class 0 or 1. Trees are constructed recursively in a way until the leaf node labels are pure. The area under curve (AUC), top decile lift and expected maximum profit (EMP) methods are used for the model's evaluation. Their results have shown that similarity forest gives better performance when the analysis intention is to do future prediction.

Mallya et al. [2019]) have used a recurrent neural network based model (with LSTM) for the prediction of congested heart failure. LSTM has given promising results for sparse irregular and high dimensional features data calculated over 24 months. Prediction is carried out using the observation and prediction windows of 18 and 6 months, respectively. Condition frequencies from the test results were calculated and aggregated for the patients for each 6-months time slice. The demographic data of patients were encoded and used with feature vectors of LSTM in a fully connected network for the binary classification of diagnosis of congested heart failure (CHF). Though this study is not directly related to the churn rate prediction, it has inspired our study to devise and use temporal features from customer transaction data to predict future churning customers.

A comparative study of conventional machine learning algorithms with the deep learning technique for churn prediction in the telecommunication industry has been carried out

by Prashanth et al. [2017]. From call data records features were calculated for churn prediction using the random forest, logistic regression, artificial neural network, recurrent neural network; and out of these models random forest and the recurrent neural network gave the best results. In this study, various feature values are consecutive and sequenced with time frames (for month1, month2, month3). The deep learning based models such as LSTM and RNN gave comparable results as they have the capability of using internal memory for sequence data prediction.

In the literature, the spatio-temporal based behavioral features are used in churn prediction, however, the purchasing sequence patterns over time are not studied and used for customer churn analysis. Our study have used the customers' transaction record data which gives signal of the changing purchasing pattern over regular time interval. The deviations in the regular patterns are studied and used with other behavioral and the demographic attributes for a prediction model. Next, we have explored customers' network which is defined by employing three dimensions of virtual connections that exist among customers. This graph network is formed using the extracted features from customers' transactional and the demographic data. We build deep learning models with these graph network features to predict the churning customers. In our findings from deep learning experiments, we obtain prediction results which are comparable or even better than the other conventional methods.

3. Data Collection and Exploration

In this chapter, we cover the details of the data used in our prediction methods, and how the data set is prepared and processed. Location and time based features which were used in our baseline model (i.e. Random forest) the extraction of features based on time and location which were used in our deep learning models (i.e. Recurrent Neural Network and Long Short Term Memory - LSTM) and the definitions of graph based features which were used in the graph network for prediction are presented in this chapter.

3.1 Data Source

We have data of one year from July 2014 to June 2015 of the bank customers, which is donated by a major bank of an OECD country. The bank has shared data over 20 topics, which encompass customer demographics, account balance, credit card information, credit card transaction details, ATM transaction details, bank campaigns and scoring details, funds transfers and call center record details. For our analysis, we have used demographics, credit card transaction detail (about 45 million records) and money transfer records of over 60 thousand customers. The bank has also provided its own monthly segmentation information of the customers which defines customer churn while customers were having different services from the bank. The demographic and transaction information of customers is anonymized by the bank by assigning a unique pseudo-identifier to the customers and their matching transaction and money transfer records.

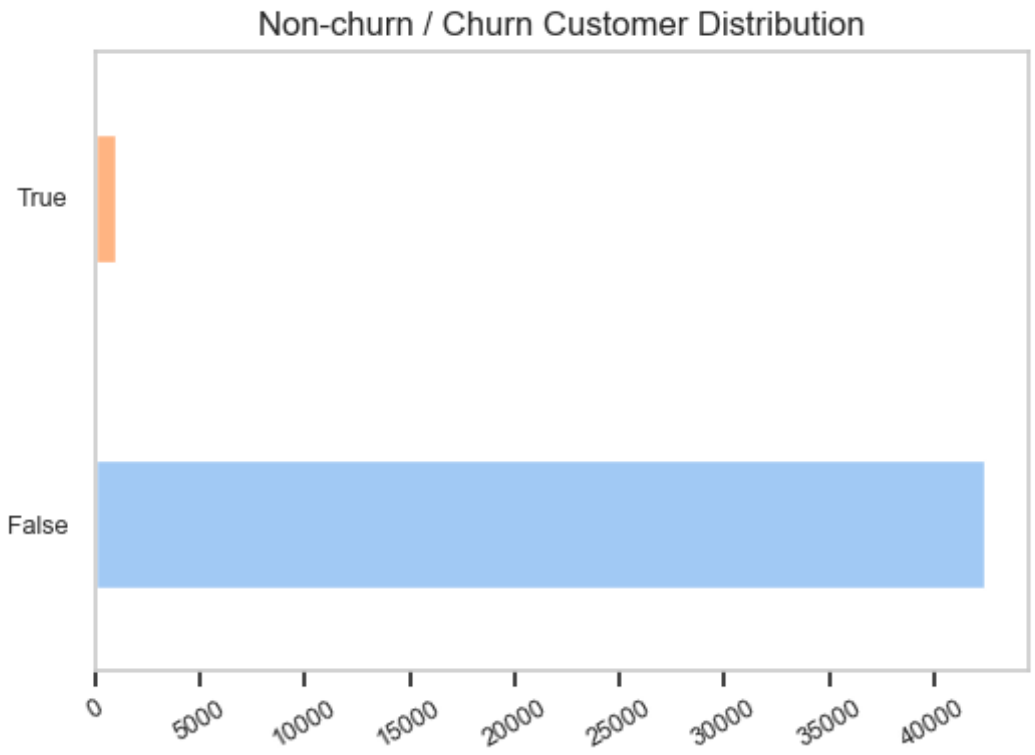
In the literature, numerous definitions of customer churn are used depending on the type of service a customer acquire and use. However, these definitions pertains to specific product or service domain and the rules to mark customers as churn are illustrated by

the industry experts. We have data segmentation information of customers for a 23-months period, which also covers the time period of transaction data (one year). In our study, we have used the definitions given by bank officials to build the prediction model for churning customers and used customers' segmentation information and transactions data of their credit card and online and offline payments record. Based on the banks' segmentation rules, we have used the definition where the customer is tagged as inactive in all of the months in the labeling window and eventually was considered as churned.

3.1.1 Customers Distribution

The distribution of churner and non-churner customers in our data-set is quite imbalance which is shown in figure 3.1. This imbalance practically exists in other real-life data sets [Burez and Van den Poel, 2009].

Figure 3.1 Churning and non-churning Customer's Distribution



The correct classification of the minority class is of great importance as compared to the

majority class in such problems. Multiple techniques are designed and tested while using machine learning algorithms to reduce the bias response of learning models and to gather maximum vital information from the minority class for better prediction results. In this study, we have also worked with multiple techniques to get desirable prediction results for the minority class by using different evaluation metrics that curtail the effect of the majority class.

3.2 Data Preparation

3.2.1 Demographic Features

The demographic features of the customers including gender, marital status, educational status, job type, income, and customer's age with the bank are available in the data set. Home and work address location points of customers and merchants (which are typically the shopping points) are also provided by the bank, which were used in the computation of location based features. The age (in years) since the customer is first engaged with the bank is also used in our prediction models. Except for income and work location address for house-wives, all the demographic attributes of the customers were complete. The missing part of income information was imputed with the mean income value of other customers. The work addresses of the house wives were filled with their home addresses. In the data preparation process, data records with missing home or work location, and the transaction records with missing merchant details are discarded. Data of accounts where job status was children and the records of customers who are marked as working abroad are also excluded from the final data set. These data pre-processing steps have reduced the transaction data set to 25M records for 43k customers.

3.2.2 Calculation of Features

Before we can implement prediction models further data pre-processing steps are performed. Categorical features including gender, education, marital status, job type are encoded and numerical features are normalized / scaled. Missing values are imputed with the mean value of the corresponding data column(s). After the data pre-processing step, behavioral features are computed form the customer's transaction data set. The section below gives details of the feature extraction steps in the context of each prediction method.

For our study, we have replicated the Random Forest classification model for the prediction of churning customers proposed by Kaya et al. [2018] and used it as our baseline model. Next we have developed a gradient boosting tree model (XGboost) and then a recurrent neural network prediction model which is based on time sequenced data (Long Short Term Memory - LSTM) and finally a graph network based model (Graph convolutional network - GCN).

3.2.2.1 Random Forest and Gradient Boosted Trees

For the base-line model i.e. Random forest; three behavioral features i.e. diversity, loyalty and regularity collectively called as spatio-temporal behavioral features and choice patterns opted to make payments are used. For the spatio-temporal features the credit card transaction which were made at the point of sales (POS) by the customers are used. And the choice patterns of making payments offline or online are taken from the money transfer records of customers. For these spatio-temporal and choice patterns based features, our study benefited from the definition and formulations given by Kaya et al. [2018] and Singh et al. [2015] work. For the calculations of behavioral features, location distance and time based bins were defined and used in the same pattern as in the referenced study. For the location dimension, square grids with 0.1 degree units and radial concentric areas of 0.5, 1, 2, 3, 4, 5, 10, 15, 30, 50, 100, 150, 300 and 500 kilometers are used. Home and work addresses of customers are taken for the calculation of these distance values. And for the time dimension, 24-hour values and 7 days of the week are used for the calculation of feature values from the transaction set. These three features are calculated against five variants of location and time based bins as grid (g), radial with home location (rh), radial with work location (rw), hourly (ho) and weekly (we). Following behavioral features of customers are calculated from the customers' transactions

over one year:

(i) Diversity: The diverse behavior of a customer is defined as the customer having a tendency to spend at diverse locations and in different spans of time. The higher the value of diversity, the more diverse a customer is marked in the context of his or her purchasing behavior. Multiple time and location-based bins are defined (as mentioned above) to measure the customer's diversity values. Mathematically, the portion of a transaction which lies in spatio or temporal bin j for customer i is calculated as p_{ij}; and then this value is normalized over all transactions using the total number of bins. The diversity score value lies between 0 and 1 and the customer whose transactions are spanned over a large number of bins will get a higher diversity score as compared to the less diverse customer.

$$D_i = \frac{-\sum_{j=1}^{N} p_{ij} log p_{ij}}{log M} \tag{3.1}$$

(ii) Loyalty: A customers' loyalty is defined as a fraction of all of his or her transactions f made in the top k most frequented bins. In our study, the top three bins ($k = 3$) are considered to calculate the loyalty score of a customer. Loyalty score also lies between 0 and 1, where higher values depict more loyal customers with the most frequented bins.

$$L_i = \frac{f_i}{\sum_{j=1}^{N} p_{ij}} \tag{3.2}$$

(iii) Regularity: This trait measures similarity / homogeneity in the customer behavior calculated over a short and long period of time. In this study, the short term is taken as one-third of the year which is referred to as the observation window by Kaya et al. [2018] and the long term as the full year. Regularity values approaching 1 show customer maintain his/her diversity and loyalty scores both in the short and long terms.

$$R_i = 1 - \sqrt{(D_i^s - D_i^l)^2 + (L_i^s - L_i^l)^2/2} \tag{3.3}$$

(iv) Choice Entropy: Some features which do not incorporate location or time based data were computed from the customer's money transfer data records, which represent the customer's choice patterns while purchasing from or transferring funds

to merchants or peers. Six features depicting a customer's money transfer or payment patterns are money transfer entropy (transe), electronic fund transfer entropy (efte), credit card transactions with respect to merchants (ecctmer), credit card transactions with respect to merchant category (ecctmcc), offline credit card transaction to merchants (efcctmer) and offline credit card transaction with respect to merchant category (efcctmcc).

3.2.2.2 Recurrent Neural Network Model

The dynamic behavioral patterns of customers based on the time series data are calculated using their transaction records. We have calculated time based features for each quarter from the customers' credit card transaction records made over year duration. These time based features have some sequence depicting the change or deviation in the behavioral patterns observed for each customer over a year and thus augment the prediction accuracy while used with other demographic and behavioral attributes. For the calculations, the location and time based bin values are used as were defined for computing the behavioral features.

- diversity_radial_hm: Diversity of customer based on radial distance between home and merchant location.
- diversity_grid_hm: Diversity of customer based on grid distance between home and merchant location.
- regularity_radial_hm: Regularity of customer spending based on radial distance between home and merchant location
- regularity_grid_hm: Regularity of customer spending based on grid distance between home and merchant location
- trans_freq: Transaction frequency of customer in each quarter
- trans_amnt: Transaction amount spent by customer in each quarter
- trans_cntprop_d: Proportion of transaction counts which were made on the weekdays
- trans_amntprop_d: Proportion of transaction amounts spent on the weekdays

- trans_cntprop_wd: Proportion of transaction counts which were made on the weekend
- trans_amntprop_wd: Proportion of transaction amounts spent on the weekend

3.2.2.3 Graph Network Model

Features that incorporate and define the social connection between customers may potentially helps in the churn prediction of the connected people. Co-churn or a single customer effecting the socially connected customers in some way and subsequently help the prediction of churn is illustrated in the study by Óskarsdóttir et al. [2017]. We propose three ways of defining the connection between the customers based on the transaction record data. In all three definitions, customers are taken as nodes and their relation attributes define the network edges.

(i) Proximity connection: The proximity connection uses the work and home locations values of customers and the closest distance between combinations of workplace and home addresses is calculated. The connection is said to exist if the distance is within a threshold value (which we take as 2.0 km). A binary flag is used to mark the pairs of records when their home or work place lies in the same district or region. This information is later used to mark the similarity connections between customers.

(ii) Money transfer connection: We define a connection between customers, when a customer i transfers money to customer j, irrespective of the amount. The total transfer frequency between the two customers either is counted and also the total transfer amount between customers i and j customer (irrespective of the transfer direction) is noted.

(iii) Common visited merchant connection: In our data set, many customers transact with the same merchants; so we calculate the count of transaction made with common merchants by a pair of customers. This information defines a connection between the pair of customers that share a common merchant(s) while making payments.

3.3 Descriptive Statistics

In this section, we describe the statistics of categorical and the numerical data of approximately 43k customers used in predictive modeling. The values for our behavioral features are also covered here.

3.3.1 Demographic Features

In our data set, the ratio of male and female customers is 69.9% and 30.1% respectively. And the churn ratio of female customers is slightly higher than that of males.

Table 3.1 Gender wise Customer Distribution

Gender	Churn Status	Count	Percentage
Female	False	12693	97.27
(30.1%)	True	356	2.73
Male	False	29691	97.91
(69.9%)	True	632	2.08

The statistics of customer with respect to marital status, education and job type including the percentage of churn and non-churn customers) are listed below. We observe that the churn rate is higher among customers with status as single which is then followed by the customers with divorced and unknown status. Married customers have a major representation of 73.64% in the data but their churn rate is lower than the customers' who are single or divorced.

Table 3.2 Marital Status wise Customers' Distribution

Marital status	Churn Status	Count	Percentage
Married	False	31318	98.05
(73.64%)	True	622	1.94
Single	False	7918	96.29
(18.96%)	True	305	3.71
Divorced	False	1803	97.99
(4.24%)	True	37	2.01
Unknown	False	1123	97.99
(2.64%)	True	23	2.00
Widow	False	222	99.55
(0.51%)	True	1	0.45

Customer's with college-level education are the major chunk in our data set, with 2.25% churn rate, which is lower than the customers with no education or level below high school. Undergraduate education level customers is the second biggest group with 2.02% churn rate.

Table 3.3 Education level wise Customers' Distribution

Education	Churn Status	Count	Percentage
College	False	18704	97.74
(44.12%)	True	432	2.25
Doctorate	False	99	98.02
(0.23%)	True	2	1.98
Graduate	False	1388	98.65
(3.24%)	True	19	1.35
High School	False	3536	97.79
(8.34%)	True	80	2.21
Middle school	False	3484	97.15
(8.27%)	True	102	2.84
No Education	False	559	94.58
(1.36%)	True	32	5.41
Primary school	False	2851	97.34
(6.75%)	True	78	2.66
Undergraduate	False	11721	97.98
(27.58%)	True	242	2.02
Unknown	False	42	97.67
(0.10%)	True	1	2.33

Customer's who are doing job in the private sector are the major chunk in our data set, with 2.35% churn rate, whereas their churn rate is lower than the customers whose status is mentioned as student where the churn rate is highest (6.67%). This seems logical as the students' used to open a bank account during their study period and may not continue using the same account during their career. Housewife, unemployed and public sector employed customers are the other major churning groups in our data set.

Table 3.4 Job-type wise Customers' Distribution

Job type	Churn Status	Count	Percentage
Housewife	False	484	95.27
(1.17%)	True	24	4.72
Not working	False	320	94.67
(0.78%)	True	18	5.32
Other	False	293	97.34
(0.69%)	True	8	2.66
Retired	False	2267	98.61
(5.30%)	True	32	1.39
Retired Employee (self-employed)	False	320	100.00
(0.74%)	True	0	0
Retired Employee (wage)	False	829	98.34
(1.94%)	True	14	1.66
Self-employed	False	5994	98.16
(14.08%)	True	112	1.83
Student	False	56	93.33
(0.14%)	True	4	6.67
Undefined	False	113	98.26
(0.27%)	True	2	1.74
Wage (Private)	False	28821	97.65
(68.05%)	True	693	2.35
Wage (Public)	False	2887	97.27
(6.84%)	True	81	2.73

Table 3.5 Descriptive Statistics of Age, Income, Bank-age Attributes

	Age	Income	Bank-age
Mean	39	4715	8
Standard Deviation	9	73257	4
Minimum	19	0	1
25%	32	1195	4
Median	38	2000	8
75%	46	3600	12
Maximum	85	9500000	39

3.3.2 Location and Time based Features

In this part, we report the descriptive statistics of our behavioral features including their mean, standard deviation (Std Dev), minimum (Min), maximum (Max), median (50%), first quartile (25%) and third quartile (75%).

Table 3.6 Descriptive Statistics of Loyalty based Features

	Loyalty-g	**Loyalty-rh**	**Loyalty-rw**	**Loyalty-h**	**Loyalty-w**
Mean	0.898032	0.817840	0.867687	0.524096	0.657123
Std dev	0.108829	0.128597	0.120691	0.118589	0.105712
Min	0.200000	0.375000	0.365385	0.226667	0.428571
Q1-25%	0.840000	0.725806	0.788965	0.437500	0.576923
Median-50%	0.928571	0.828571	0.894737	0.507937	0.641791
Q3-75%	1.000000	0.923077	0.973684	0.600000	0.727273
Max	1.000000	1.000000	1.000000	1.000000	1.000000

Table 3.7 Descriptive Statistics of Diversity based Features

	Diversity-g	**Diversity-rh**	**Diversity-rw**	**Diversity-h**	**Diversity-w**
Mean	0.000027	0.000095	0.000093	0.000087	0.000137
Std Dev	0.000027	0.000094	0.000092	0.000090	0.000140
Min	0.000006	0.000021	0.000021	0.000018	0.000029
Q1-25%	0.000011	0.000038	0.000038	0.000034	0.000055
Median-50%	0.000018	0.000064	0.000063	0.000058	0.000092
Q3-75%	0.000032	0.000115	0.000113	0.000106	0.000166
Max	0.000860	0.003060	0.002893	0.003064	0.004658

Table 3.8 Descriptive Statistics of Regularity based Features

	Regularity-g	Regularity-rh	Regularity-rw	Regularity-h	Regularity-w
Mean	0.960387	0.938436	0.952240	0.887785	0.907957
Std dev	0.050298	0.063878	0.056467	0.104169	0.079160
Min	0.434315	0.602252	0.595939	0.485741	0.595939
Q1-25%	0.942667	0.910640	0.932009	0.836822	0.865295
Median-50%	0.977143	0.957910	0.970537	0.922206	0.931254
Q3-75%	0.999989	0.987026	0.996250	0.967932	0.970884
Max	1.000000	1.000000	1.000000	1.000000	1.000000

Table 3.9 Descriptive Statistics of Choice Pattern based Features

	efte	ecctmer	ecctmcc	efcctmer	efcctmcc
Mean	0.571763	0.845711	0.756351	0.773285	0.748247
Std dev	0.394641	0.149508	0.141813	0.113643	0.147597
Min	0.000000	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.000000	0.788662	0.687177	0.709771	0.674139
Median-50%	0.757151	0.894994	0.788508	0.781017	0.780384
Q3-75%	0.898609	0.951310	0.857912	0.850995	0.854572
Max	1.000000	1.000000	0.994030	1.000000	1.000000

3.3.3 Sequence based features

Below are the statistics of the sequence based (time series) quarter wise features which were computed for the Recurrent Neural Network (LSTM) model, with the plots showing deviation in the mean values for churning and non-churning customers separately.

Table 3.10 Descriptive Statistics of Diversity(grid) Feature

	q1 diversity-g	q2 diversity-g	q3 diversity-g	q4 diversity-g
Mean	0.386340	0.354704	0.343280	0.340052
Std dev	0.270420	0.264521	0.271665	0.270893
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.181188	0.141182	0.091427	0.078963
Median-50%	0.394633	0.351464	0.336900	0.336844
Q3-75%	0.579380	0.530616	0.528159	0.522960
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.2 shows decreasing pattern in the mean of diversity(grid) score calculated over four quarters for the churning customers as compared to the not-churning customers.

Figure 3.2 Mean values plot of Diversity(grid)

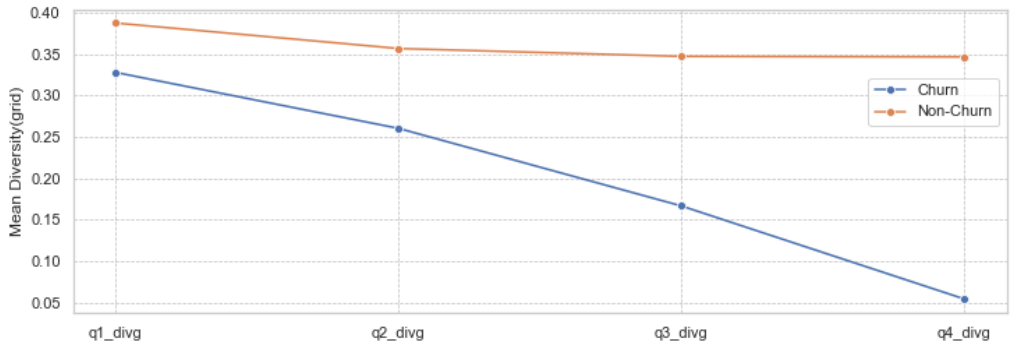


Figure 3.3 illustrates a pattern that the customers which are labeled as churning have a decreasing trend of mean diversity score when compared to those who are not-churning.

Table 3.11 Descriptive Statistics of Diversity(radial) Feature

	q1 diversity-r	**q2 diversity-r**	**q3 diversity-r**	**q4 diversity-r**
Mean	0.395363	0.377351	0.359800	0.363030
Std dev	0.224256	0.223145	0.229653	0.237693
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.259825	0.244219	0.210674	0.195676
Median-50%	0.433724	0.412697	0.390976	0.406885
Q3-75%	0.571956	0.554609	0.542338	0.554609
Max	0.961667	0.917052	0.928098	0.932705

Figure 3.3 Mean values plot of Diversity(radial)

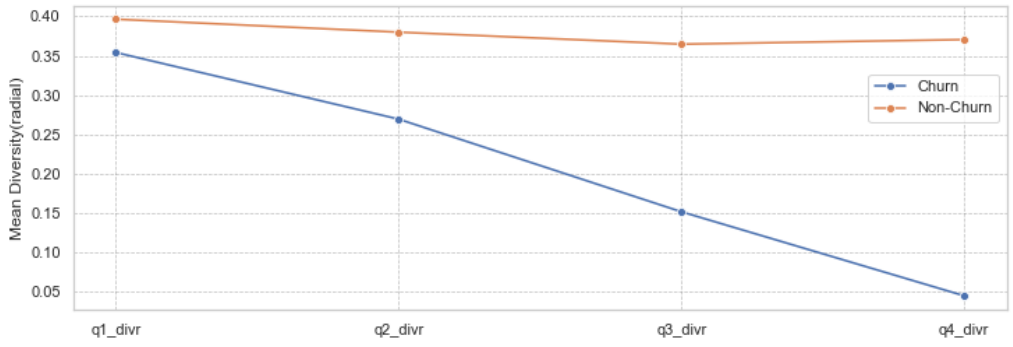


Table 3.12 lists the summary statistics of the transaction amount spend by the customers in four quarters and Figure 3.4 illustrates the decreasing pattern in the amount spent by customers (in weekdays) who are going to churn in near future while compared to the

retained customers transaction trend which shows similar pattern in all the quarter values. Similar pattern is observed and shown in Figure 3.5 while considering the transaction pattern of the customers making purchases at the weekend.

Table 3.12 Descriptive Statistics of Transaction amount(weekdays) Feature

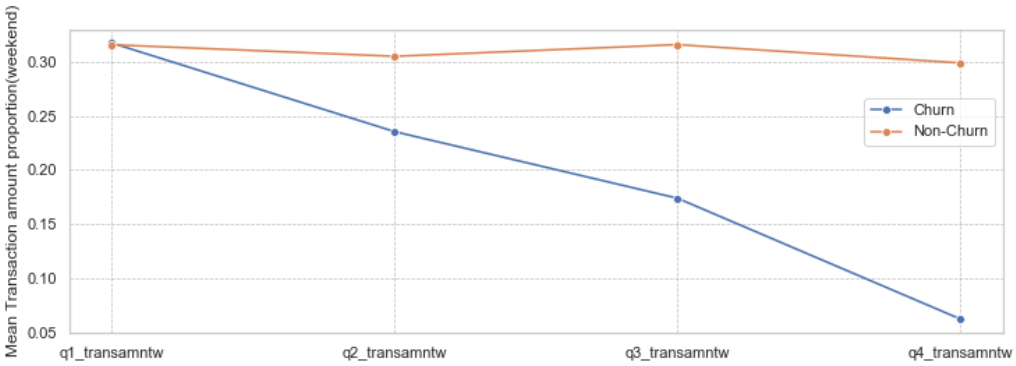
	q1 tamount-d	q2 tamount-d	q3 tamount-d	q4 tamount-d
Mean	0.658036	0.655410	0.618507	0.612399
Std dev	0.254920	0.263861	0.289924	0.299434
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.500000	0.500000	0.500000	0.500000
Median-50%	0.692308	0.700000	0.666667	0.666667
Q3-75%	0.833333	0.833333	0.821429	0.823529
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.4 Mean values plot of Transaction amount(weekdays)

Table 3.13 Descriptive Statistics of Transaction amount(weekend) Feature

	q1 tamount-w	q2 tamount-w	q3 tamount-w	q4 tamount-w
Mean	0.315633	0.303296	0.312462	0.293393
Std dev	0.236576	0.234696	0.250857	0.244146
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.146341	0.133333	0.125000	0.090909
Median-50%	0.291667	0.282051	0.285714	0.272727
Q3-75%	0.454545	0.434783	0.464286	0.437500
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.5 Mean values plot of Transaction amount (weekend)



To observe the deviations in the purchasing behavior of the customer, we have plotted mean values of the transaction frequency (at weekend and weekdays separately); and Figure 3.6 and 3.7 have confirm the declining trend in number of transactions along with the decrements in the amount spend by the churning customers.

Table 3.14 Descriptive Statistics of Transaction frequency (weekdays) feature

	q1__transfreqd	q2__transfreqd	q3__transfreqd	q4__transfreqd
Mean	0.658036	0.655410	0.618507	0.612399
Std dev	0.254920	0.263861	0.289924	0.299434
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.500000	0.500000	0.500000	0.500000
Median-50%	0.692308	0.700000	0.666667	0.666667
Q3-75%	0.833333	0.833333	0.821429	0.823529
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.6 Mean values plot of Transaction Frequency (weekdays)

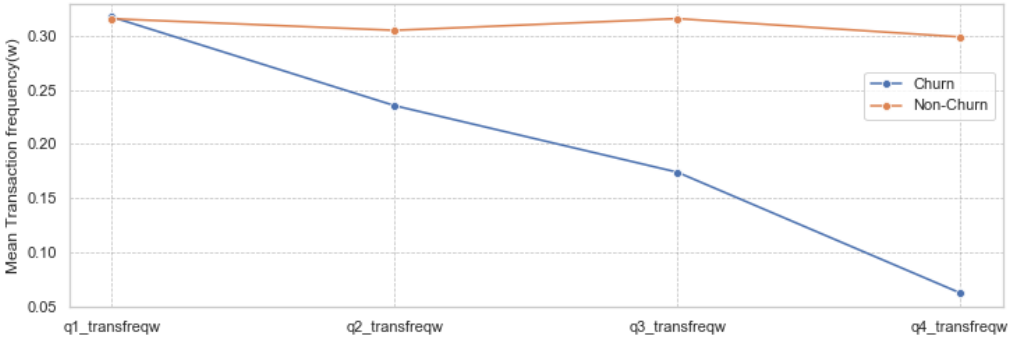


Table 3.15 Descriptive statistics of Transaction frequency (weekend) feature

	q1_transfreqw	q2_transfreqw	q3_transfreqw	q4_transfreqw
Mean	0.303296	0.312462	0.293393	0.612399
Std dev	0.236576	0.234696	0.250857	0.244146
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	0.146341	0.133333	0.125000	0.090909
Median-50%	0.291667	0.282051	0.285714	0.272727
Q3-75%	0.454545	0.434783	0.464286	0.437500
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.7 Mean values plot of Transaction Frequency (weekend)

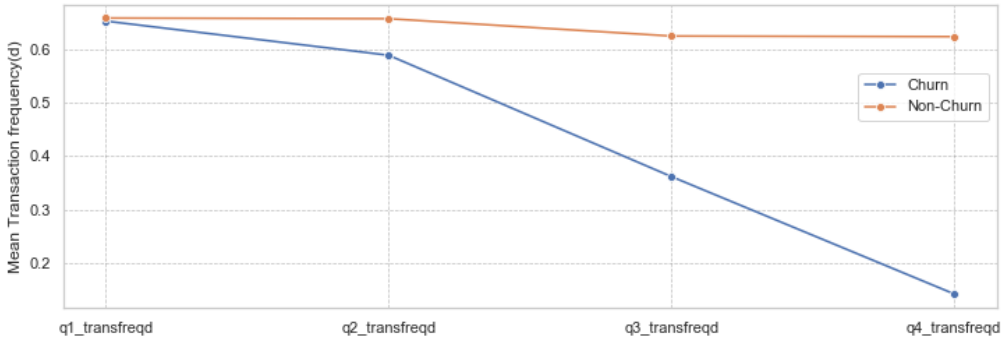


Table 3.16 Descriptive statistics of Transaction frequency feature

	q1 tfreq	q2 tfreq	q3 tfreq	q4 tfreq
Mean	15.327954	15.151411	14.478350	15.065895
Std dev	19.962317	19.638072	19.609779	20.624965
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	5.000000	5.000000	4.000000	4.000000
Median-50%	10.000000	10.000000	9.000000	9.000000
Q3-75%	19.000000	19.000000	18.000000	19.000000
Max	790.000000	846.000000	705.000000	912.000000

Figure 3.8 Mean values plot of Transaction frequency

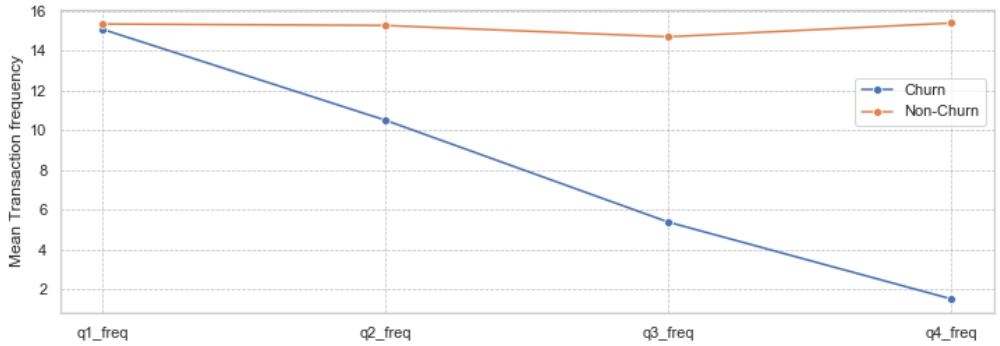
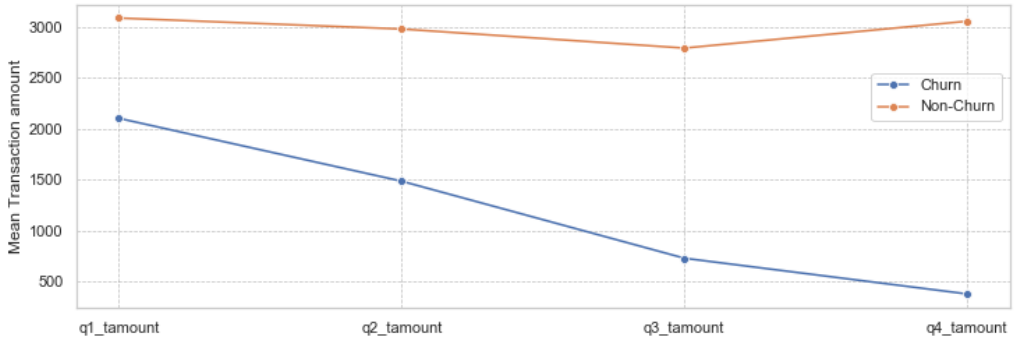


Figure 3.8 and 3.9 shows the transaction frequency and amount trend of the churning and non-churning customers, irrespective of considering segregation in the weekdays and weekend purchases.

Table 3.17 Descriptive statistics of Transaction amount Feature

	q1 tamount	q2 tamount	q3 tamount	q4 tamount
Mean	3065.259565	2945.955806	2746.228657	2995.664647
Std dev	7106.315760	6817.795993	6440.625401	6940.597065
Min	0.000000	0.000000	0.000000	0.000000
Q1-25%	521.115000	512.410000	429.907500	434.747500
Median-50%	1225.040000	1205.395000	1096.750000	1182.205000
Q3-75%	2880.275000	2765.822500	2590.290000	2847.057500
Max	285463.250000	224793.750000	248313.930000	258966.140000

Figure 3.9 Mean values plot of Transaction amount



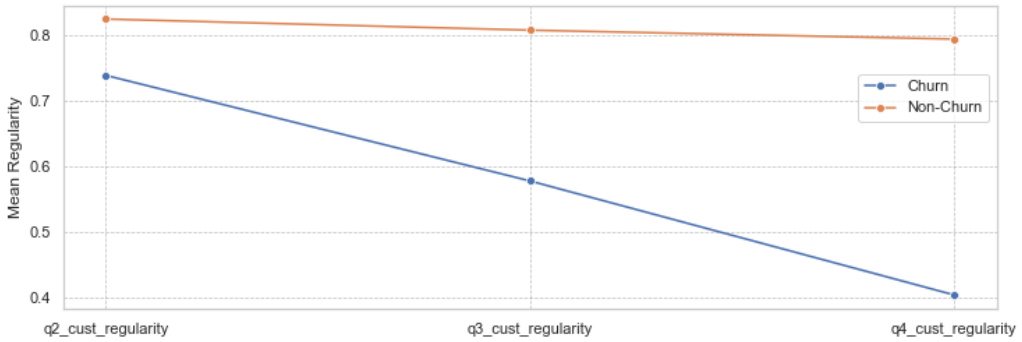
The regularity feature takes both the diversity and loyalty values of the customer to measure the homogeneity in the customer behavior over a short and long interval of

time. Figure 3.10 shows the decreasing mean value of regularity score for the churning customer. The irregular or declining value confirms that the customer is going to churn.

Table 3.18 Descriptive statistics of Regularity feature

	q1 regularity	q2 regularity	q3 regularity	q4 regularity
Mean	0.207313	0.822807	0.802511	0.785200
Std dev	0.115514	0.183589	0.202292	0.216307
Min	0.029850	0.216532	0.216532	0.216532
Q1-25%	0.143377	0.779740	0.753692	0.726635
Median-50%	0.183960	0.876401	0.867722	0.858785
Q3-75%	0.237232	0.946275	0.940847	0.936266
Max	1.000000	1.000000	1.000000	1.000000

Figure 3.10 Mean values plot of Regularity score



3.3.4 Graph Network based features

To form a virtual network of customers, we have defined connectivity based features by considering three dimensions. Customers' demographics and transaction data is used to compute these features.

3.3.4.1 Proximity Connection

To find the proximity connections between customers, the distance between home and workplace of each customer pair is calculated and a binary flag is used if they have home and workplace in the same district. The minimum value of 0.2 kilometer is considered to measure the distance between the customer's work place and home. This way approximately 1,5 million connections are defined among the customers. Mean distance between the customer pairs comes out to be 7.5 kilometer while considering their home distance and 4.1 kilometer with respect to their work distance.

Table 3.19 Descriptive Statistics of Proximity Connection Dataset

	Home distance	**Work distance**
Mean	7.4601	4.0944
Std dev	9.6179	7.7462
Min	0	0
Q1-25%	0.1535	0.0968
Mean-50%	3.6319	0.1715
Q3-75%	11.1927	5.2901
Max	105.5496	131.5519

3.3.4.2 Money Transfer Connections

Here we identify the customer pairs who have transferred amounts between them, and recorded the count of transfers made by either customer along with the total transfer amount. This data set is quite small as compared to the proximity information, having only 1400 connections. Average four (4) transactions are made between the customer pairs with mean amount of 11395. The statistics of this dataset is as below:

Table 3.20 Descriptive Statistics of Money Transfer Dataset

	Transfer count	**Total amount**
mean	3.9623	11395.2801
std	6.2128	42622.4054
min	1.0000	0.0000
25%	1.0000	200.0000
50%	2.0000	1005.0000
75%	4.0000	6138.0000
max	70.0000	732500.0000

3.3.4.3 Common Visited Merchant's Connection

Connections between a customer is marked to exist if they have make a transaction to a common merchant. 11 million connections are defined while considering common merchant information between the customer pairs. For this connectivity dimension, we have customer's ids defining the network nodes and the number of common merchants describe the connectivity information (edge value) in the network.

Table 3.21 Descriptive Statistics of Common visited Merchants' Dataset

	Visited Merchant
Mean	1.324312
Std dev	0.828186
Min	1.0
Q1-25%	1.0
Median-50%	1.0
Q3-75%	1.0
Max	35.0

4. Methodology

4.1 Modeling Approach

Various linear and non-linear techniques have been in practice in the past by researchers for various prediction models. This study tends to gauge the predictive performance of churning customers models by using two deep learning algorithms i.e. Recurrent Neural Network (RNN)-Long Short Term Memory (LSTM) and Graph Network (Graph Convolutional GCN) against the Random Forest and Gradient Boosted decision tree (XGBoost) conventional algorithms.

4.1.1 Baseline model

We have replicated the Random Forest model for churn prediction with the same parameters and architecture as done by Kaya et al. [2018] using one of the four data sets acquired from the same data source. In addition, we have used gradient boosted tree (XGBoost), which is a computationally more flexible and efficient machine learning algorithm for binary classification. We take the Random forest model as the baseline model and further analyze and compare the performance of the other three models, namely the gradient boosted tree method and the two deep learning techniques.

4.1.1.1 Random Forest

Random forest is an ensemble algorithm used effectively for both the classification and regression based tasks. Random forest operates by constructing a multitude of decision trees and yields the mode of classes for classification and the mean prediction of the individual trees for regression. Random forest is able to predict better due to its ability to find strength and correlation between the individual and weak predictors as introduced by Breiman [2001] following the earlier works of Amit and Geman [1997]. Breiman has proposed the random forest methodology where each tree is constructed using a different data sample taken by bootstrapping and each node is split using a random subset of features. The use of weak predictors is vital in the prediction process, and an ensemble method is employed in a Random Forest which works on the principle of merging the weak learners to get a stable and more accurate prediction performance. Random Forest is a way of averaging multiple deep decision trees, trained on different attributes of the same data set. This way, the high variance resulted from over fitting of the trees is reduced while boosting the model's performance. Random forest employs bootstrap aggregating or bagging for the selection of best features from the features set and thus resulting in a diverse tree structure that yields better results.

Random forest algorithm works as follows

- M number of random trees are created by drawing n observations randomly with (or without) replacement from the original data set of size N. These n observations are used to build trees.
- k features are selected randomly from a total of L features, where $k < L$; and using these k features best split point for each node is calculated.
- Each tree is grown fully to its maximum; no pruning is applied.
- Final predictions are made by considering the majority votes for classification and average for regression based on the predictions made by each tree.

Kaya et al. [2018] has performed churn analysis with Random Forest while using the customer's demographic and transaction data. They have extracted location and time based features from transaction data (behavioral features named as spatio-temporal and choice patterns). The purchasing trend of customer illustrates the behavior of customer which were calculated as how diverse a customer is, the change in the loyalty pattern of customer and regularity measuring the diversity and loyalty of customers over short and

long term time intervals. These behavioral features of customers were calculated using the credit card transaction data over a year. Their model is replicated here by adopting the same parameter values, and the results of this model are taken as a benchmark to assess the prediction performance of recurrent neural network (RNN-LSTM) and the graph network model.

4.1.2 Gradient Boosted Decision Tree (XGBoost)

XGBoost is one of the fastest implementation of gradient boosted trees algorithm for large data sets with a relatively large number of features, and has provided state-of-the art results to the data scientists in many competitions. Extreme gradient boosting (commonly known as XGBoost) is an efficient and scale-able machine learning algorithm for tree boosting developed by Chen and Guestrin [2016]. It creates a new model by taking loss value of all possible splits (across all leaf points) and thus reducing the search space of the next possible split. XGBoost uses an objective function of the loss function and regularization term to get better predictions and to control model complexity in relation to over-fitting.
The XGboost algorithm works as follows:

- Fit an initial model ($F0$) to predict the target variable (Y), such that:
  $F0(X) = Y$
- Fit a model to the residuals from the first step:
  $R0(X) = Y - F0(X)$
- Create a new model $F1$ which will be the boosted version of $F0$:
  $F1(X) = F0(X) + R0(X)$
- The process of building new trees goes on till the residual is minimized as much as possible; $Fm(X) = F(m-1)(X) + R(m-1)(X)$
- As XGBoost uses gradient descent for the optimization of a loss function, each new model will be fitted on the gradient of loss generated from the previous step. The $R0(X)$ will be trained on the cost reduction for each sample.
- The multiplicative factor γ for each terminal node is calculated and the boosted model $Fm(x)$ is:

$$Fm(X) = F(m-1)(X) + \gamma R(m-1)(X)$$

In out XGBoost implementation, we have used the same set of demographic features along with behavioral features (i.e. diversity, loyalty, regularity) based on location and time (spatio-temporal) and choice patterns. To achieve an optimized performance, tuning of model parameters is done using a grid search mechanism. Details on the parameters selection are covered in section 4.2.

4.1.3 Deep Learning Algorithms

Among many other techniques developed under machine learning, an emerging branch is the Deep Learning (DL). DL has provided solutions to many complex problems including image and language processing, speech recognition and in medical research and diagnostics. The design of DL is inspired and analogous to the neural network of human brain. They are composed of number of processing layers that learns from the data with multiple level of representations while modeling complex relationship among data. We have opted two deep learning techniques, Recurrent Neural Network (RNN) and Graph Convolutional Network (GCN) for our study. Details of both techniques are as follows:

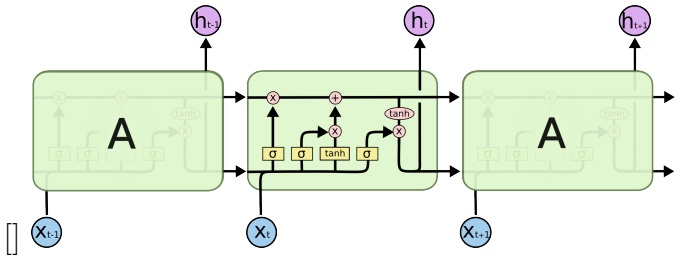
4.1.3.1 Recurrent Neural Network (RNN)

Recurrent neural network (RNN) is one of the robust neural networks which simulates the human neurons working, and has been extensively used in deep learning for the last two decades. In a traditional neural network all inputs (and outputs) are independent of each other, whereas the performance of a RNN is based on previous computations. Recurrent Neural Network are proposed for the tasks which involves sequential data as input and output is mapped and generated using the input data. RNN makes use of the sequential information or patterns present in the data and is well suited for the prediction problems where the given data has some sequence. RNN has internal memory capability and with this feature they can connect and learn the temporal connections of sequential data and predict the next sequence. The hidden state in the RNNs capture and store information

of the sequence. Practical implementations of RNN face the limitations of exploding and vanishing gradients, which stop them to learn from long range intervals. Bengio et al. [1994] have shown that the gradient descent of the error in RNNs is inadequate to train and learn from long-term dependent data.

Long Short Term Memory (LSTM) is the solution to this problem of RNN given by Hochreiter and Schmidhuber [1997], where the gradient is truncated while allowing LSTM to learn by enforcing constant error flow. Also LSTMs are able to handle the long term dependency issue by remembering the long time information. LSTM preserves memory and keeps it in their cell state, while these cells take input of the previous state h_{t-1} and current input (X_t). LSTM keeps or removes information in the cell states and these cell states are regulated and maintained by special structures called gates. LSTM has shown promising results in language modeling and text generation where the next sequence is dependent or connected with the previous data. LSTM works as follows (image source colah.github.io):

Figure 4.1 Architecture of LSTM



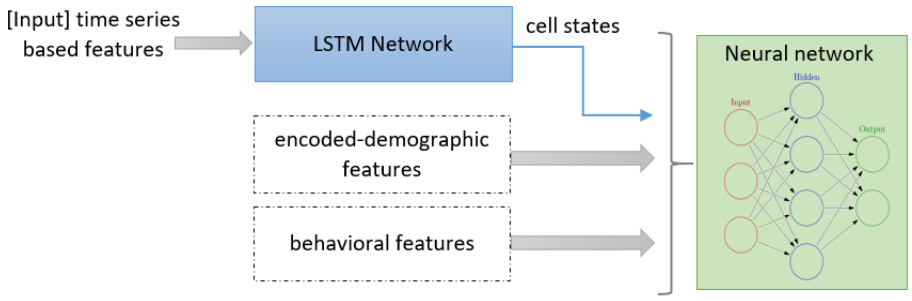
- Like recurrent neural networks, LSTM has a repeating module which is called as cell. The memory unit of LSTM is this cell. This cell is a four-layered network which works interactively in its structure.
- LSTM cell takes three inputs: X_t is the current input, h_{t-1} is the output from the previous LSTM unit and C_{t-1} is the memory of the previous unit.
- In the next step, the information will be stored in the cell state; after the sigmoid layer will decide which value to update and feature vector will be created at *tanh* layer, which will be added to the state.
- Old cell state C_{t-1} will be updated with the new state C_t; after multiplication of old state by f_t and then adding ($i_t \times \hat{C}_t$).
- The output is based on the cell state, which is passed through the sigmoid function first. Cell state is passed through *tanh* function and then it is multiplied with the

output of the sigmoid function. h_t is the hidden state output of the current network and C_t is the memory of the current unit.

LSTM handles sequential order and time variant events, therefore it is beneficial to use in the prediction problems which involve recurrent events. The activities of bank customer's are recurrent over a passage of time and show deviations which can be tracked from their transactions data. The deviations are not obvious in the transactions data of the customers but these variations in the purchasing pattern can be measured from the attributes calculated over the passage of time. The variations in the spending pattern will tell us whether the customer is thinking to leave the bank services. As the bank customers employ bank services in some regular fashion for their regular purchases or payments, we have proposed the recurrent neural networks (LSTM) methodology to predict their future action. A prediction model for the classification of churning and non-churning customers by learning the temporal patterns derived from the transactions data helps in making the competitive predictions.

To build a predictive sequence for the classification of churn and non-churn customers, we processed features from the transactions data based on credit card payments made at the point of sales (POS) by the customers. We have formulated sequenced features from the customers' transaction data to be used as an input in the RNN model. The features are calculated over three months (i.e. quarter wise) time duration over one year to predict thhe next set of values in the sequence as explained in the LSTM mechanism. To utilize the customer demographic data in the prediction, we have proposed a hybrid deep learning architecture with LSTM and neural networks, which takes both static and dynamic features (i.e. demographics and time series data) and the internal cell state from the LSTM model for the prediction.

Figure 4.2 Recurrent Neural Network (LSTM) architecture



4.1.3.2 Graph Convolutional Network (GCN)

Many real-world data sets typically have network based features defining user connections and other related attributes like in the cases of social networks, bio-chemistry, language domain. Some work has been done for the generalization of neural network techniques and convolutional neural networks for graph data to achieve better prediction capability. Defferrard et al. [2016] have introduced a neural network which operates on graph network data. Classical convolutional neural network (CNN) works efficiently for the recognition or classification of imagery, audio, video or speech data, where underlying data has a regular grid structure. Defferrard et al. [2016] have worked on generalizing the convolutional neural networks (CNN) for high dimensional irregular data, carried out experiments on a common data set (MNIST) and achieved competitive results. Their approach was intended for the solution to the problems where structured graph data exist (like in language semantics, social networks and sequence prediction in the field of chemistry). Li et al. [2015] have worked on defining graph features using graph neural networks with gates used as the recurrent units and then predicting sequences as an output. Kipf and Welling [2016] have proposed a graph convolutional network (GCN) based on CNN for graph structured data. With semi-supervised modeling, they have made the assumption that the connected nodes may share the same label in the graph and so it is possible to make a prediction for a data set where classification information is missing or not available. In GCN, the graph structure is encoded and trained in the neural network along with the labels, and the nodes having the same patterns for connectivity. The working of GCN is explained below:

(i) Graph network is defined as $G = (\nu, E)$ with n number of vertices (ν) and m edges (E).

(ii) Graph G takes input for every node in a $(N \times F)$ matrix format (N = no. of nodes; F = no. of features).

(iii) Graph description is contained in the adjacency matrix (A) which holds the information of the weighted edges. A_{ij} will be one (1) if there is an edge between i and j vertex, and 0 otherwise

(iv) The node level output is $N \times F$; where F is no. of output features for N nodes.

(v) Graph structure will be encoded in the neural network model layers as $f(X, A)$ and trained where labels information for nodes is present.

(vi) Hidden layer will be non-linear and formulated as: $H^k = f(H^{k-1}, A))$ where $H^0 = X$ and and $H^l = Z$ and k is the number of layers. Each of the layers H^k corresponds to an $N \times F^k$ feature matrix where each row is a feature representation of a node.

(vii) The features are aggregated at each layer and form the next layer's features using the propagation rule. In this way, features become more abstract at each consecutive layer. Different activation rules (like Relu) are used for layer wise propagation.

In our problem where we have defined the graph connections between customers, the prediction model using GCN has shown that similar or even more competitive results can be achieved using these connections and related attributes of these connections. Using the graph features, customers are marked as similar if any of the connectivity feature exists between the customer pair. The results can be further improved with the availability of more data which defines stronger connections between the customers including social and/or business connections. Integration of social media data with services data helps to define a network of people who are using the services from the same service provider and a change in this network can help to forecast the social trend.

4.2 Tuning Model Hyper-Parameters

For the implementation of machine learning algorithms, generally two types of parameters are trained and used in model building. Hyper-parameters which are passed to the model for better performance and others which are learned by the algorithm during the execution phase. Hyper-parameters can be randomly selected and tested for a better model, but it is computationally exhaustive and time consuming to search for and find the best combination of parameters using the hit and try method. Grid search and random search methods are employed in our study for the selection of the hyper-parameters. Every combination of specified hyper-parameters is used for building the model in the grid search method and then the model is evaluated against the criteria specified. In random search, however, random combinations of parameters are used to find the best model. The subset of parameter values in the grid and random search are taken from the related work done in the past for similar solutions that have produced competitive results. Grid search is computationally more expensive than random search as it builds a model for

each combination of parameters and it is largely a affected by the number of features defining the data. In grid search, the parameter values are defined in the matrix form and combinations of these parameters are evaluated for a model to achieve the best performance.

In our experiments, we have applied grid and random search methods on the training data set only. To get unbiased estimates of our models' predictive performance, the test data set was taken out as the first step with a 20-80 ratio, while keeping the balance of non-churn and churn customers in both data sets the same as in the full data set. Using the best parameters, the models were trained and evaluated with k-folds cross validation stratified sampling. To reproduce the experiments with the same results, a single seed value is used for all random factors where applicable.

Following tuned hyper-parameters were used in the experiments:

4.2.1 Random Forest

The following parameters are used for the binary classification of churner and non-churner customers using the random forest algorithm, which are replicated from Kaya et al. [2018] and taken as a base model:

- Number of Estimators (n_estimators): 500
  This value defines the number of trees in the forest.
- Split criteria (criterion): Gini
  To measure the quality of split criteria, Gini is used.
- Maximum depth of tree (max_depth): None
  Default value is used for the maximum depth of the tree, which means nodes will be expanded till all leaves are pure.
- Number of features (max_features): 2
  Two features are considered while considering each split.
- Minimum number of samples for internal split (min_samples_split): 2
  Minimum of two samples are required for internal node split.

- Minimum Number of Samples at Leaf Node (min_samples_leaf): 1
  Minimum one sample required to be at a leaf node.
- Limit on the Leaf Node Number (max_leaf_nodes): None
  Default value is used to grow trees with max leaf nodes.
- Minimum Impurity for Split (min_impurity_split): 10-7
  This is the threshold for early stopping in tree growth. This parameter is deprecated; minimum impurity decrease (min_impurity_decrease) is used here.

4.2.2 Xtreme Gradient Boosting (XGBoost)

Using the grid search method, the following parameters are selected for the gradient boosting tree classification method:

- Maximum Tree Depth (max_depth): 2
  This numerical value sets the maximum tree depth.
- Learning rate (learning_rate): 0.01
  Boosting learning rate of 0.01 is used.
- Number of trees (n_estimators): 200
  Number of trees to fit.
- Learning task (objective): binary:logistic
  Binary logistic is defined as the learning task for classification task of churner and non-churners.
- Booster (booster): gbtree
  In our case for tree based model gbtree is used.
- (min_child_weight): 1
  Minimum sum of instance weight needed in a child.
- (gamma): 0
  Default 0 value is used for gamma, which defines the minimum loss value required to make a further partition on a leaf node of the tree.

- (subsample): 0.5
  This value is used for the sub-sampling of training data prior to growing the tree.
- scale_pos_weight
  This value controls the balance of classes in the imbalance data set. We have used equal weights for both the classes.
- (reg_alpha): 100
  This is L1 regularization term on weights which is used after parameter tuning.

4.2.3 Recurrent Neural Networks (LSTM)

For the RNN model, the activation and optimizer functions which pertain to the design of the model were searched as a first step using the grid search method and then used for the tuning of the learning rate of model execution. For LSTM, a three-layered network is defined. The input layer of LSTM takes values of nine features which were calculated with a quarter time interval (i.e. four values calculated for an year). The cell-states are taken as the output from the LSTM network. Then three layered sequential neural networks takes the input of cell states (from LSTM) along with the demographic and calculated behavioral features. The neural network is designed with single hidden layer having 32 neurons. Dropout rate (of value 0.2) is used after the input and hidden layer, which reduces the network tendency to over-fit. Relu is used as the activation function at the input and hidden layer, and the sigmoid function is used at the output layer. The binary cross entropy loss function is used with adam function as an optimizer.

4.2.4 Graph Convolutional Network (GCN)

For the GCN model, a three-layered network with learning rate of 0.001 and drop out rate of 0.2 after the input layer is used for training the model with 200 epochs. Rectified linear unit (ReLU) is used as an activation function. This network is based on the graph network model defined by Kipf and Welling [2016].

4.3 Handling Imbalanced data

In the available data set, non-churners largely outweigh the churners, which is a common scenario in real time systems where the data is skewed and representation of classes is not evenly distributed. This imbalance in the data set adds bias to the machine learning models' output, by learning only the majority class and failing to detect the class of interest (which is the churn class in our study). For both multi and binary classification problems, various data engineering techniques are practiced to handle the imbalanced data. Data and algorithm level methodologies were identified by He and Garcia [2008] in their work to handle the imbalance data. At the data-level, different forms and combinations of data under-sampling, over-sampling, re-sampling with replacement, directed or random sampling are employed to make the data suitable for usage. However, both under and oversampling techniques have some consequences in their usage; for instance in under-sampling there is a danger of losing some important data samples, and in over-sampling meaningless samples may be drawn from the data. In algorithmic approaches, class ratios are adjusted to counter the class imbalance issue (with the option of imposing some cost values) or the adjustments are made in the decision threshold values for the classification problems.

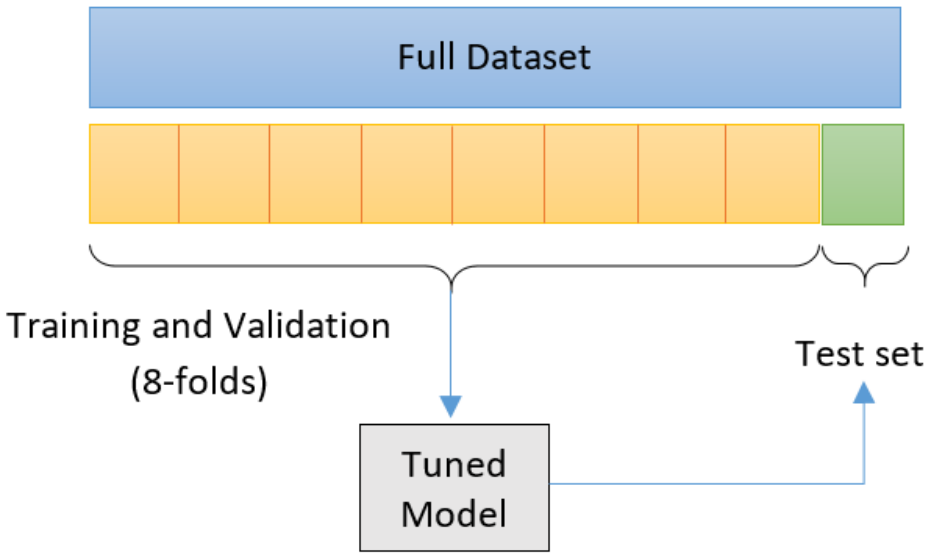
To handle the data imbalance in the Random Forest baseline model; oversampling is performed using SVM-SMOTE with the ratio 0.25:1 for churn and non-churn customers respectively. With the gradient boosted trees, recurrent neural network model and graph network model, class labels are weighted equally to balance the binary classes i.e. positive (churn=True) and negative (churn=False) class.

4.4 Data Splits - Cross Validation

For a reliable and unbiased performance, the models need to be evaluated on the unseen data which will determine how well a model has learned. The evaluation results will also arbitrate the model generalization and performance so that it is not over-fitting. In our study, for the unbiased validation of machine learning algorithm performance, training

and testing steps are performed on separate subsets of the data. As mentioned earlier, we have taken out the test data set separately even before the hyper-parameter selection so that the test data set remains unseen. After the first split of data as training and test (taking the ratio as 80-20), k-fold cross validation is applied on the training data set. In this step, the training data is divided into k-folds and at each step of model execution, one of the folds is held out for validation and the rest of k-folds is used for the training purpose. This technique strengthens the performance of the model against bias and variance.

Figure 4.3 Data Split - Cross Validation



4.5 Performance Evaluation Metrics

Model evaluation is an integral part of data mining. For the classification problem, where data is highly imbalanced; Confusion Matrix, Precision, Recall, the area under ROC curve (AUROC) and precision-recall curve (PR-curve) are the preferred evaluation matrices.

4.5.1 Confusion Matrix

The confusion matrix, also known as the contingency table is one of the tools to measure performance and assessment of the classification method, and is used for both binary or multi-class problems. This table shows the values from the actual class of input data versus the predicted values from the model output. The calculations of other performance measures like precision, recall, F1-value are done from the values in the confusion matrix. The output of the confusion matrix for binary classification is 2×2, or for the multi-class classification problems it is $N \times N$, where its cell values are the number of actual and predicted class. In our work, the negative class represents the non-churner (denoted by 0) and positive class represents the churning customer (denoted by 1). True positive and true negative values are targeted to be maximized so that maximum prediction accuracy can be achieved by the learning algorithm.

Figure 4.4 A 2×2 Confusion matrix

		Predicted class	
		Class 0	Class 1
Actual class	Class 0	C_{00}	C_{01}
	Class 1	C_{10}	C_{11}

Accuracy is one of the intuitive performance measure which is used for the classification tasks. It is a measure of correctly predicted observations out of the total observations. Using the confusion matrix values accuracy can be calculated as:

$$Accuracy = \frac{TP + TN}{(TP + FP + FN + TN)}$$

Accuracy can be misleading where the number of sample records are not in equal proportion for each class. For example, considering a data set where 90% of data belongs to a negative class and positive class has only 10% representation. The classifier will predict most of the samples in the negative class giving the 90% classification accuracy. However, practically this value is misleading for imbalanced data and not useful to evaluate the performance of a prediction model as being the majority class most of the samples will be marked in the negative class.

4.5.2 Precision and Recall

Precision (also known as positive predictive value (PPV)) is the evaluation metric preferably used in the case of imbalanced data. Precision measures what proportion of positive class predictions are actually positive and is calculated as:

$$Precision = \frac{TP}{(TP + FP)}$$

Recall is known as the true positive rate (TPR, also called sensitivity), and it represents the proportion of the actual positive class which is predicted correctly. Recall is calculated as:

$$Recall = \frac{TP}{(TP + FN)}$$

For the performance measurement of our models, both precision and recall values are considered. However, it is commonly observed that recall values increase while the precision value decrease, and vice versa. Consequently, a threshold value for classification should determine the acceptable values of recall and precision measures. Another metric called F1-score has been introduced, which gives the weighted average value of both recall and precision.

$$F1 - score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

In our work, we have used precision, recall and F1-score as performance measures. Adjustment of the threshold value for classification can be done to get the optimal results where the F1 value is higher. However this adjustment is a business decision and it depends on the problem domain where either recall or precision measure needs to be more emphasized than the other.

4.5.3 Area under the Receiver Operating Characteristic Curve (AUROC)

Provost et al. [1998] have recommended using the Receiver Operating Characteristics (ROC) for the evaluation instead of accuracy for the binary classification problems. ROC plots false positive rate (FPR) against true positive rate (TPR) along the x and y-axis, respectively. The area under the curve (AUC) score is calculated from the ROC curve

and its value lies between 0 and 1. The larger the value of area under the curve, the better the model performance is, while a model with 0.5 or below area under curve value is accepted as not performing well. Though AUC score is generally used to compare multiple models' prediction performance. However, the value for the area under ROC may significantly improve with the higher number of majority class (non-churner) records without improving the positive predictive value of the minority class (churners). ROC curves are insensitive to class imbalance and they represent an optimistic view of model performance in the skewed class problem [Davis and Goadrich, 2006]. Precision-recall curve (PR curve) is recommended to use in such cases.

4.5.4 Area under Precision - Recall Curve

ROC is commonly used for the performance evaluation and comparison of the learning models; however in the case of imbalanced data set where positive class (like churn in our case) is rare, an alternative measure, which is the precision recall curve is recommended to be used ([Craven and Bockhorst, 2005],[Bunescu et al., 2005]). The precision-recall curve (PR-curve) is plotted with Precision at y-axis and Recall at x-axis. PR-curve is the trade-off between the positive predictive value (i.e. precision) and the true positive rate (i.e. recall or sensitivity) using probability threshold values. So the PR-curve considers the correct prediction of the minority class (in our case, the churn class). The balance between the precision and recall can be attained by testing different threshold values, as lowering the threshold value will increase the recall value (true positive count) and decrease the precision value.

4.6 Software and Libraries

In this study, libraries and functions are used from Python, which is an open-source tool for the development. Following is the list of libraries we have used:

- Handling Imbalance data: using SMOTE (imbalanced learn),

- Data split: sklearn (StratifiedShuffleSplit)
- Grid and Random search: sklearn (model_selection)
- Data standardization: sklearn (preprocessing)
- Random forest: sklearn (RandomForestClassifier)
- Extreme Gradient Boosting: xgboost (XGBClassifier)
- Recurrent Neural Network: keras (sequential)
- Graph Convolution Network: keras (tensorflow)

5. Results and Discussion

In this chapter, we present the results of our experiments for the churn prediction of customers and discuss the results of different machine learning methodologies we have used including the base line model (i.e. random forest), gradient boosted trees, recurrent neural network (LSTM) and graph network (GCN). This chapter also assesses the performance of different methodologies evaluated against different measures. Our experiments covers:

- Replication of random forest model for the prediction of churning customers
- Development and comparison of the gradient boosted trees (XGBoost), and deep learning models including recurrent neural network (LSTM) and graph network (GCN) for the churn prediction.
- Performance analysis of deep learning models while employing behavioral features calculated from the transaction data of customers.

For the experimental setup, data of 43372 customers of one year as provided by a financial organization is analyzed. Location and time based features are calculated from the transaction data of customers which were used in the supervised learning model (random forest, gradient boosted trees and recurrent neural network) and semi-supervised method (graph network). The results show that the behavioral attributes and graph features are significant and contributes in the customers' classification while using supervised and deep learning classification methods. The behavioral attributes of customers are computed from the transaction data while using the variants of customer and merchants locations and transaction time-stamps which are available in the given data set. The sequences extracted from the transaction data employed in the recurrent neural networks (LSTM) and the graph features which show the connection between the customers are used in the graph network (GCN) for prediction.

The cumulative density functions of the behavioral features calculated from customer transaction data are discussed in the section below.

5.1.1 Analysis of Spatio-Temporal based Features

This section covers the location and time (spatio-temporal) based behavioral features explained in terms of cumulative density functions. For the three behavioral features diversity, loyalty and regularity; div, loy and reg abbreviations are used in the following figures. And for five bins definitions these abbreviations are used with grid (g), radial with home location (rh), radial with work location (rw), hourly (ho) and weekly (we) suffixes.

Figure 5.1 Cumulative density function - Diversity

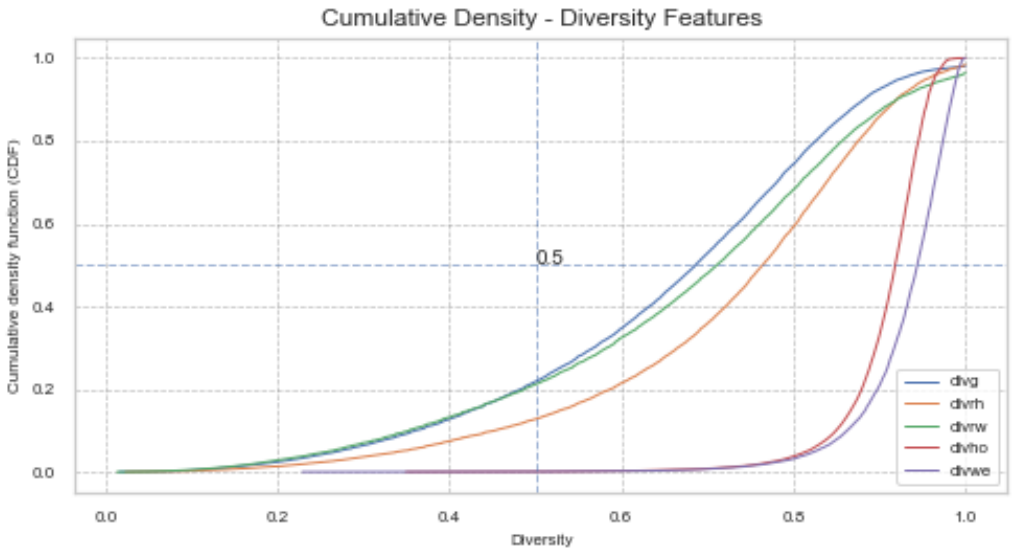
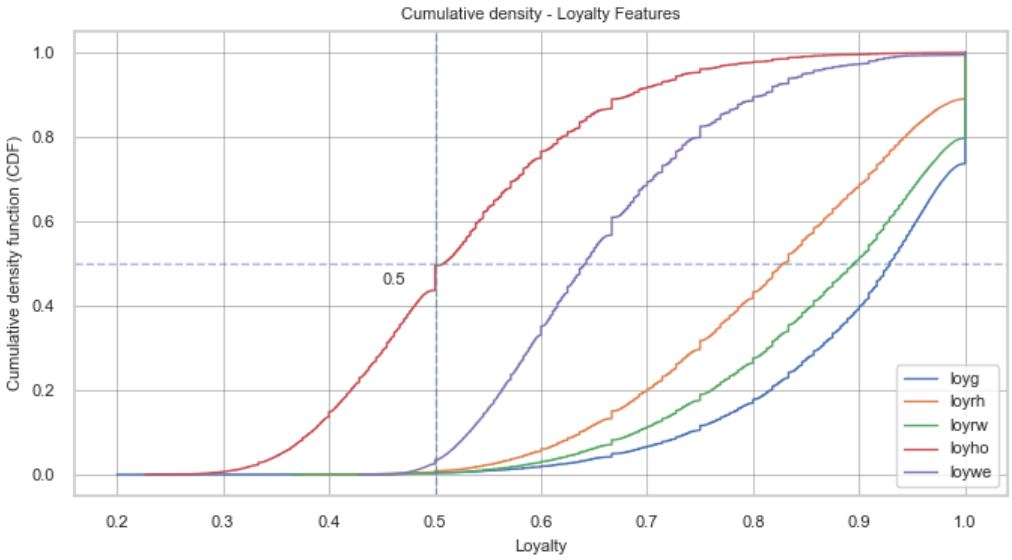


Figure 5.1 illustrates the cumulative density function values for the diversity related features calculated for the customers. Plot shows that the customers are relatively little more diverse when their shopping behavior is characterized using the grid-based location values than the radial based location values. The diversity is the entropy of transactions made by the customers which is calculated against the location and time-based bins.

Figure 5.2 Cumulative density function - Loyalty



The cumulative density function values for the loyalty features are shown in Figure 5.2. This plot shows that more than 50% of the customers prefer to shop, generally speaking at top three merchant preferences. Furthermore, 50% of the transactions are made in the top three most preferred shopping hours. Customers are in general more loyal with respect to the location of their purchase. The loyalty patterns vary in the context of transaction time but are consistent in the context of shopping locations.

Figure 5.3 Cumulative density function - Regularity

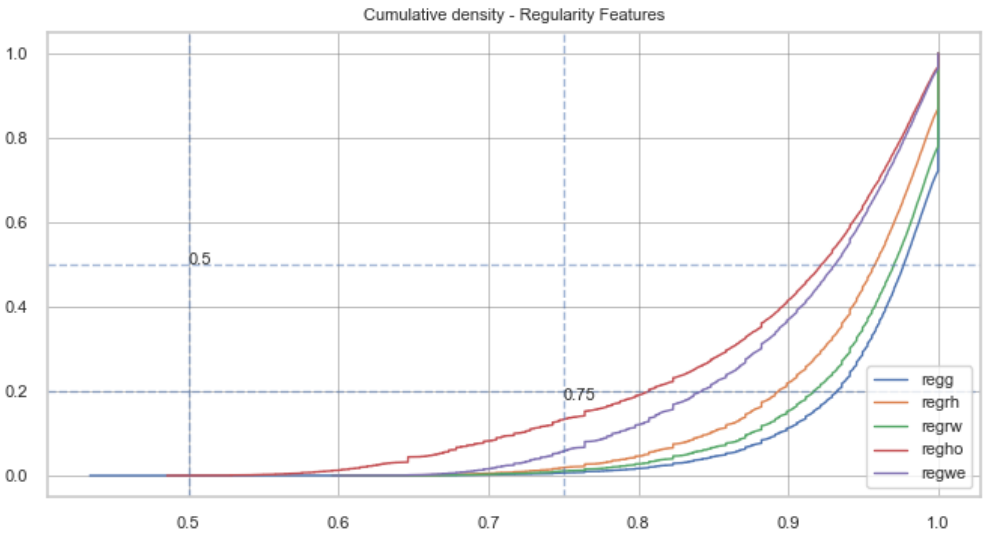


Figure 5.3 shows the cumulative density function values for the regularity behavior of the customers which are consistent in terms of location and time bins. This plot shows that the customers are more regular for shopping with respect to location. More than 80% of the customers have regularity score greater than 0.7. Customers are more regular in terms of the locations they shopped at, as compared to the time or day of shopping. Customers are regular in three of the choice patterns, i.e. they were regular in terms of making payment to different merchants and merchant categories, offline transactions via credit card to merchants.

Figure 5.4 Cumulative density function - Choice Pattern

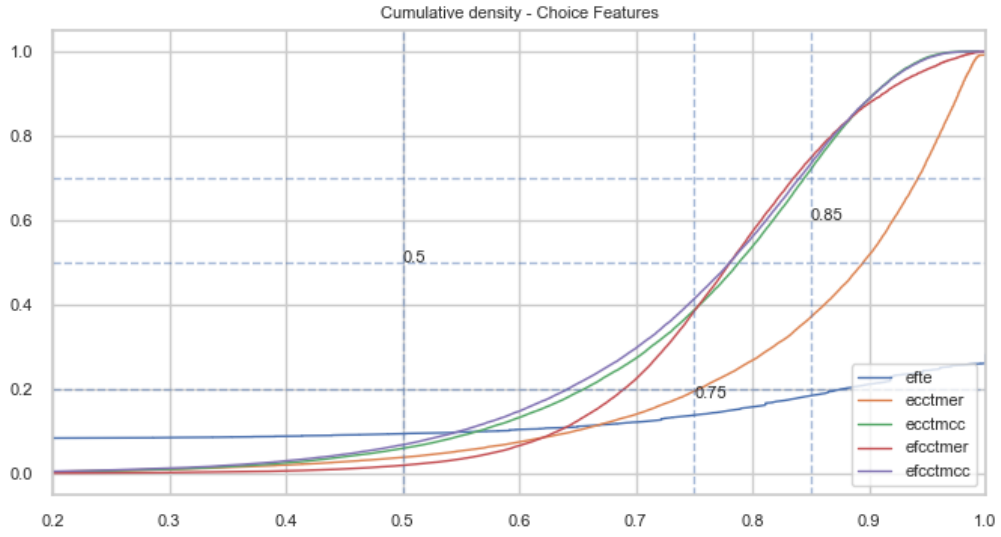


Figure 5.4 shows the cumulative density function values for the choices opted by the customers while making transactions. The plots of offline transactions via credit card with merchants (efcctmcc) and with respect to merchant category (ecctmcc) show that customers have evenly distributed transactions for these two payment modes. The customers are more diverse in the context of making payments to different merchants (ecctmer) while fewer customers use the EFT transaction mode (efte).

5.1.2 Sequence based Feature Analysis

This section covers the explanation of sequence based features (quarterly and over the course of a year) which are extracted from the credit card transaction data of the customers and are then used in the recurrent neural networks prediction model.

Figure 5.5 Cumulative density function - Diversity Radial Quarterly

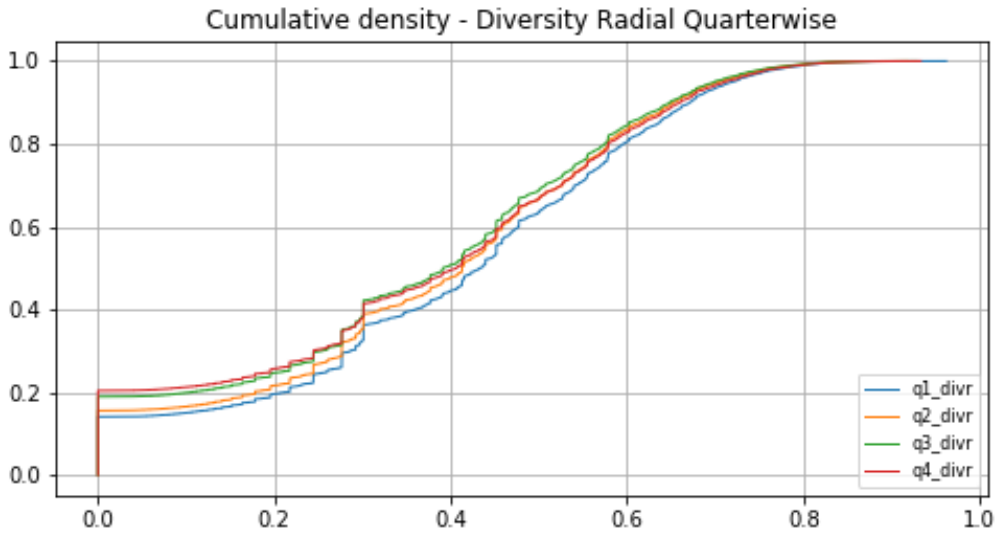
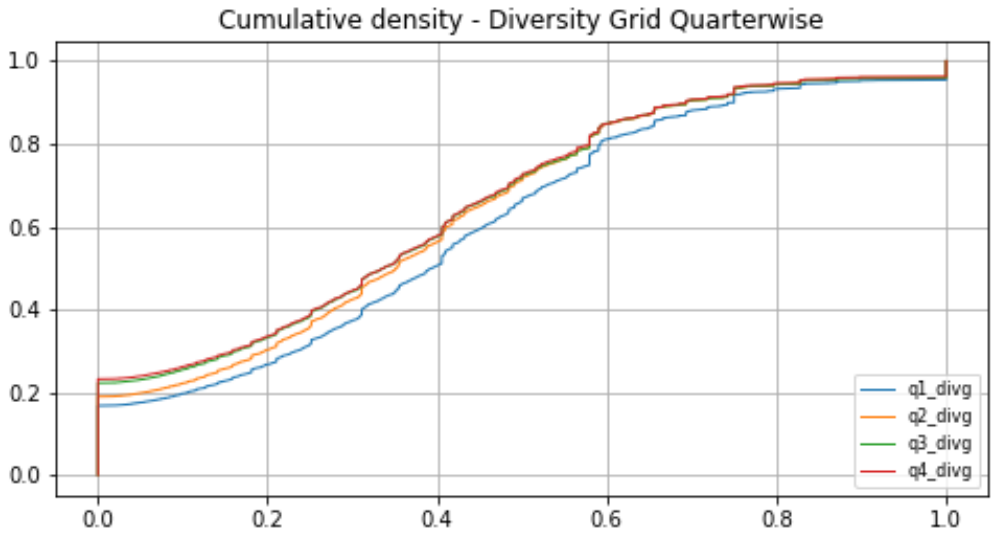


Figure 5.5 illustrates the pattern of customers diversity (radial) in four quarters. Less than 50% of the customers have diversity score greater than 50% and this pattern is almost similar in all quarters with minor deviation in quarter wise values. The deviation in the quarter wise values is not obvious in the quarter-wise plots as the count of churning customers is very low.

Figure 5.6 Cumulative density function - Diversity Grid Quarterly



The pattern of customers diversity (grid wise) in four quarters is shown in Figure 5.6. Less than 40% of the customers have entropy of diversity score greater than 0.5 and this pattern is almost similar in all quarters with the minor deviation observed in the each quarter.

Figure 5.7 Cumulative density function - Transaction Amount (weekdays) Quarterly

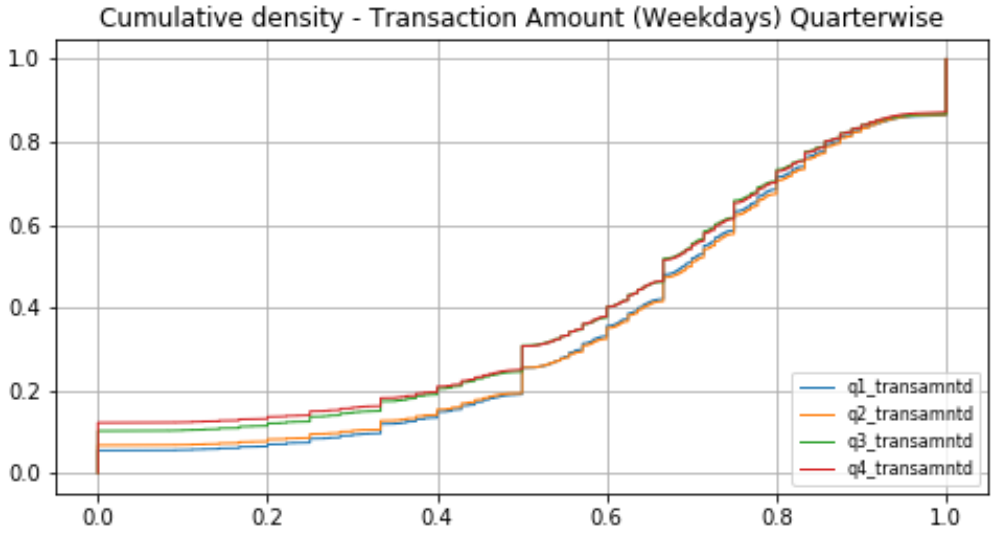


Figure 5.7 shows the cumulative density pattern of transactions made by customers during the week-days in four quarters. Identical pattern is observed in all quarters with more

than 60% of the customers having higher entropy value (more than 0.6).

Figure 5.8 Cumulative density function - Transaction Amount (weekend) Quarterly

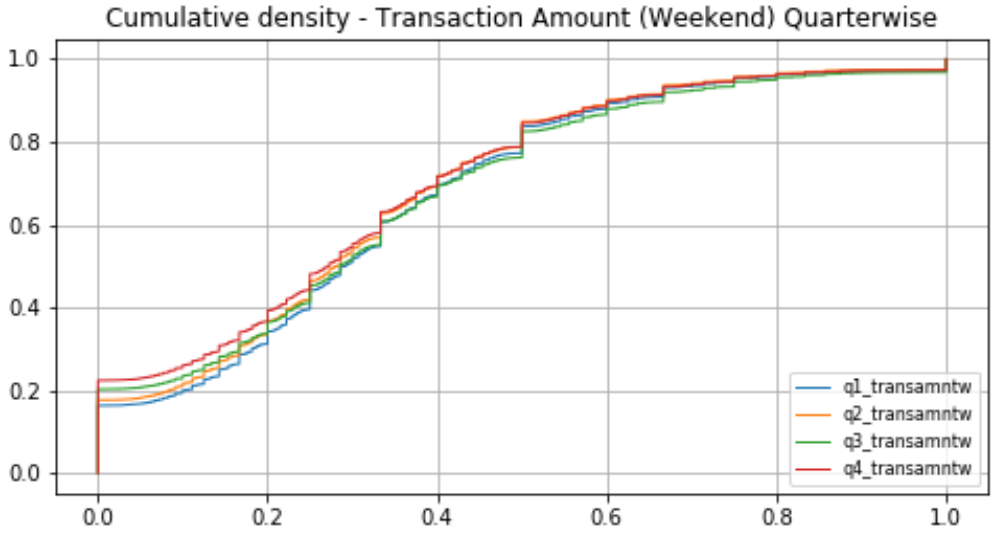


Figure 5.8 depicts the pattern of transactions made by customers at the weekends in four quarters. The plot shows that more than 50% of the customers have higher entropy values of making heavy purchases at the weekend.

Figure 5.9 Cumulative density function - Transaction Frequency(weekdays)

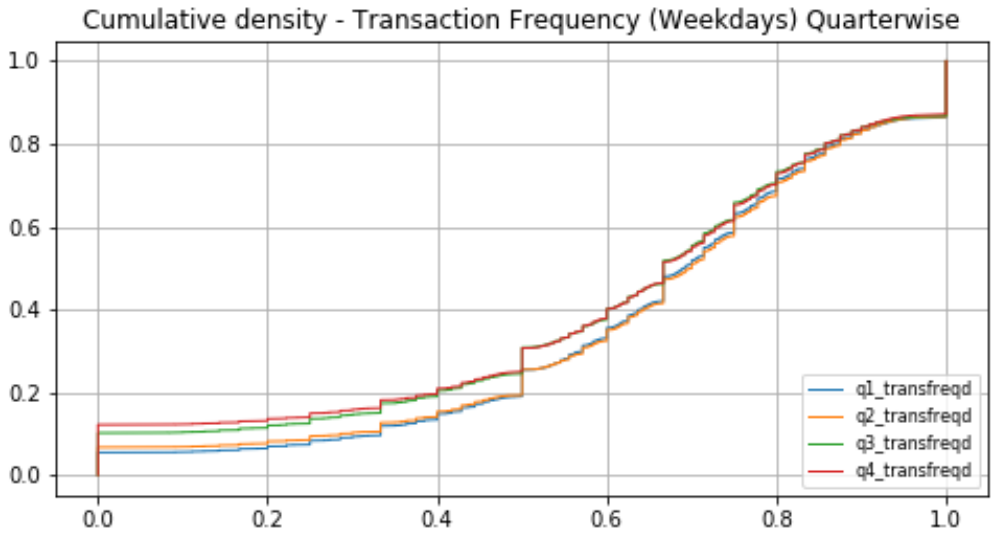


Figure 5.10 Cumulative density function - Transaction Frequency(weekend)

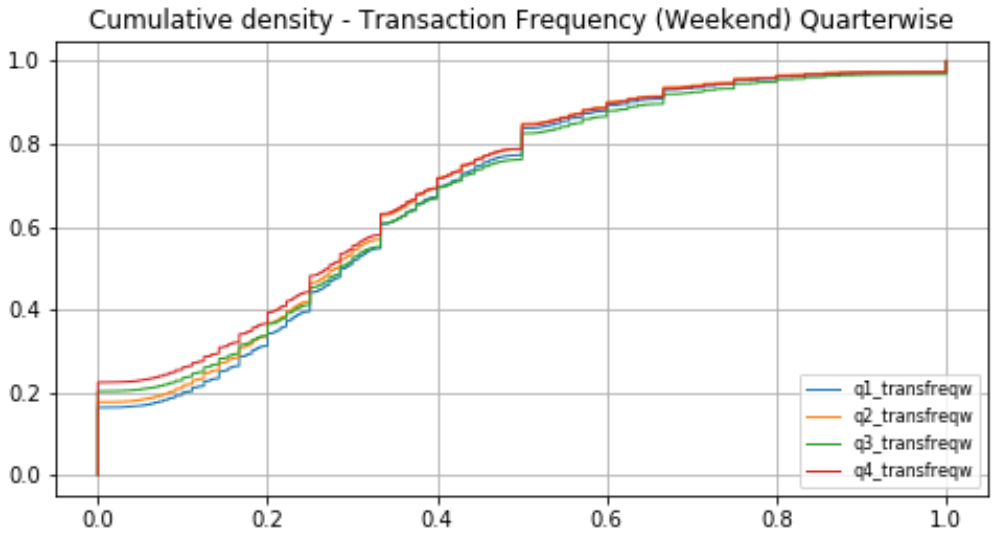


Figure 5.11 Cumulative density function - Transaction Frequency Quarterly

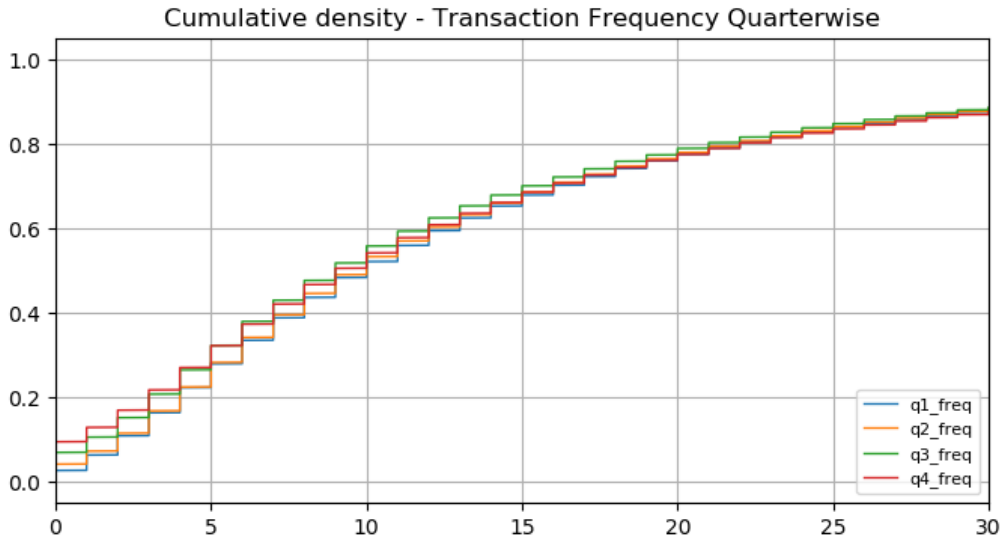


Figure 5.11 illustrates the frequency of transactions made by customers quarter wise. Transaction frequency is less than 50 for majority of the customers, while only a few the customers make large numbers of purchases.

Figure 5.12 Cumulative density function - Transaction Amount Quarterly

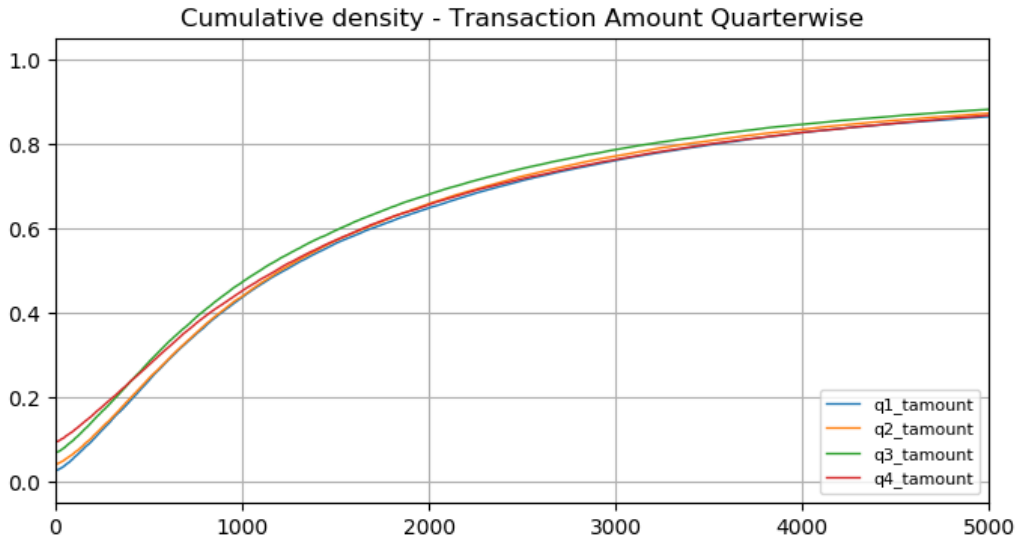


Figure 5.12 shows the total amount of transactions made by customers in each quarter. The transaction amount is below 5k for a majority of customers (more than 80%).

Figure 5.13 Cumulative density function - Customer Regularity Quarterly

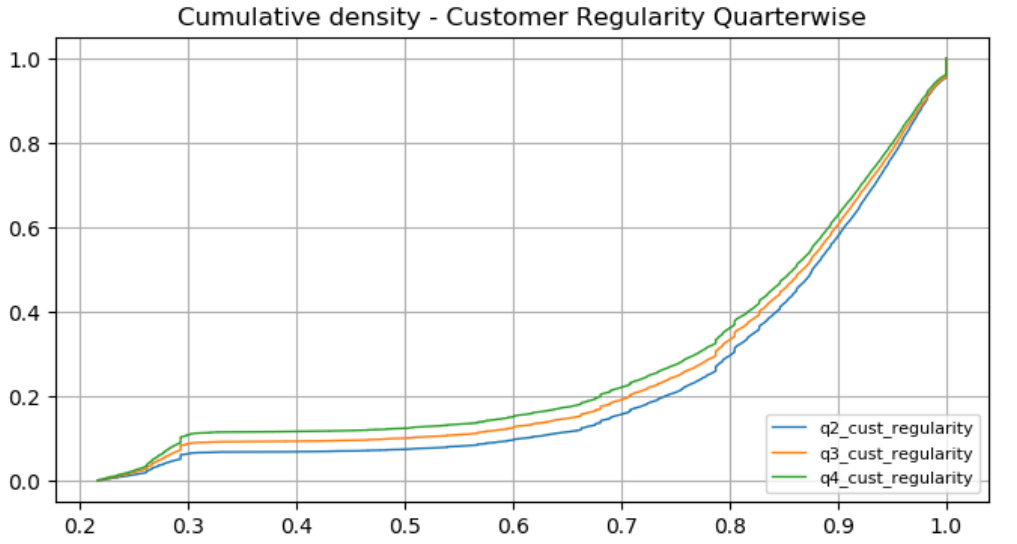


Figure 5.13 illustrates the customer regularity patterns which are based on the diversity and loyalty values of each quarter. The regularity pattern is calculated for the subsequent quarters (2,3,and 4) by using the quarter 1 regularity score as the base value. More than

10% of the customers have regularity score greater than 0.5 and this pattern is constant in all three quarters.

5.2 Churn Prediction Performance Analysis

The prediction performance of our supervised and semi-supervised models is evaluated, compared and presented in this section using various evaluation measures, which are recommended in the literature for use with the imbalanced data sets. For our experiments, 20% of the data is taken out in the beginning and is used for model testing, and the remaining 80% data is used (with k-fold cross validation) for training the models. Churn and non-churn classes are balanced (as 1:1) after using the class weight property for the training data set. Initial results are reported with 0.5 classification threshold value, whereas multiple threshold values are also plotted to evaluate the model performance. The random forest and gradient boosted tree methods employ demographic and calculated behavioral features for the churning customers prediction, whereas the recurring neural network model has used feature vectors calculated based on the time sequenced transaction data and the graph network model has additionally used customer connection data for prediction. The prediction results of Kaya et al. [2018] have been used as a baseline in our study using only one of their four test data sets.

Table 5.1. below lists the evaluation results obtained while using different supervised and semi-supervised algorithm in our experiments. Prediction results of the models are highly effected by the class imbalance ratio, where negative (non-churn) are outnumbered by the positive (churn) customers. In our study, recall value is considered as the evaluation metrics, which determines that the actual positive i.e. churned customers are predicted well by the models. Also the value of area under precision-recall curve is examined, which measures the correct predictions (of positive class) only.

While comparing the models performance, it is important to mention here that Kaya et al. [2018] have reported only the AUROC value from the Random Forest model, whereas we have calculated other evaluation measures as well, as it's recommended in the literature to compute Precision, Recall and area under precision-recall curve for the models' assessment in case of imbalanced data sets. Though the tree based algorithms handles the imbalanced

data but this issue impairs their performance (Kirui et al. [2013]). The results from deep learning models and gradient boosted trees as listed in the Table 5.1 shows that we have achieved improved results from our classification models if we consider the AUROC metric only.

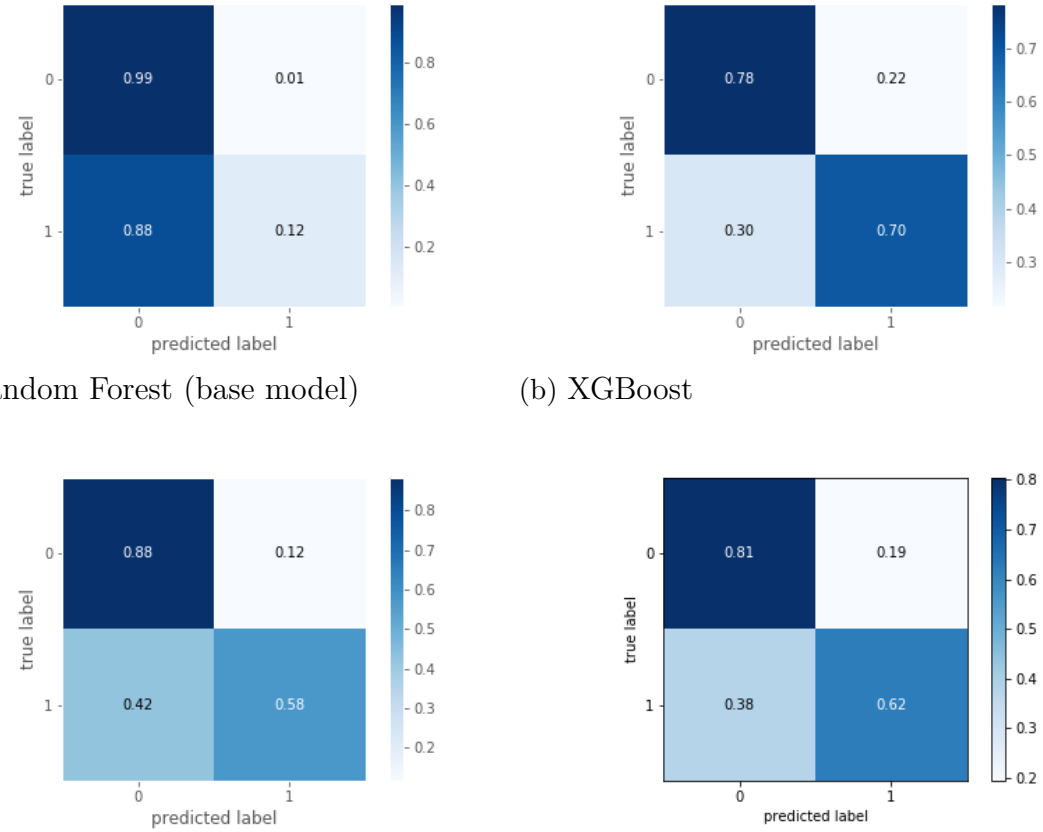
Table 5.1 Evaluation results

	Precision	**Recall**	**F1-Score**	**AUROC**	**PR-AUC**
Random Forest	99.29	90.97	17.02	79.20	13.69
XGBoost	78.16	**99.04**	12.27	**82.50**	**37.58**
RNN (LSTM)	84.54	**98.88**	14.40	**82.70**	**34.11**
GCN	83.21	**96.56**	13.21	**83.02**	**23.21**

5.2.1 Confusion Matrix

Figure 5.12 shows the confusion matrices with the normalized values, obtained from our four prediction models. From the confusion matrices, it can be concluded that both deep learning-based models and the gradient boosted trees have performed the prediction task quite well, while increasing the true positive rate (i.e. of the churning customers) in comparison to the base line model. The default threshold value (0.5) is used for the binary classification in our study for the churning and non-churning customers and to calculate the performance scores against multiple measures. Prediction of the minority class (i.e. churn positive) is not overlooked in the gradient boosted tree (XGBoost) and deep learning models as compared to the baseline model. As a result the prediction ratio of false positive records is less. The recall value of XGBoost model is better than the recurrent neural network and graph network prediction model, but the corresponding F1-score of RNN is better than the graph and gradient boosting methods. Area under the precision-recall curve is equal in the case of XGBoost and RNN models, meaning that both models are performing equally while predicting the positive class. We see considerable variation in the values obtained for the precision and recall measures of the baseline model and other models. We have high precision value and lower recall value in the case of the base model and a high false positive rate which shows that most of the churning customers are not identified correctly. The precision and recall values of the non-churning customers is high in base model. Recall value is less in the base model

as compared to other models, and this means that the ratio of correct prediction of the churned customers out of total churning customers is lower. The precision and recall values are different in case of gradient boosted tree (XGBoost) and deep learning models; where models were able to learn and predict the churning customers shows higher recall values. The proportion of false negatives is less in both RNN (0.12) and graph network (0.19) than XGBoost (0.22), while the proportion of false positives is higher (RNN = 0.42, GCN = 0.38) than XGBoost (0.30).



(a) Random Forest (base model)

(b) XGBoost

(c) Recurrent Neural Network (LSTM)

(d) Graph Network (GCN)

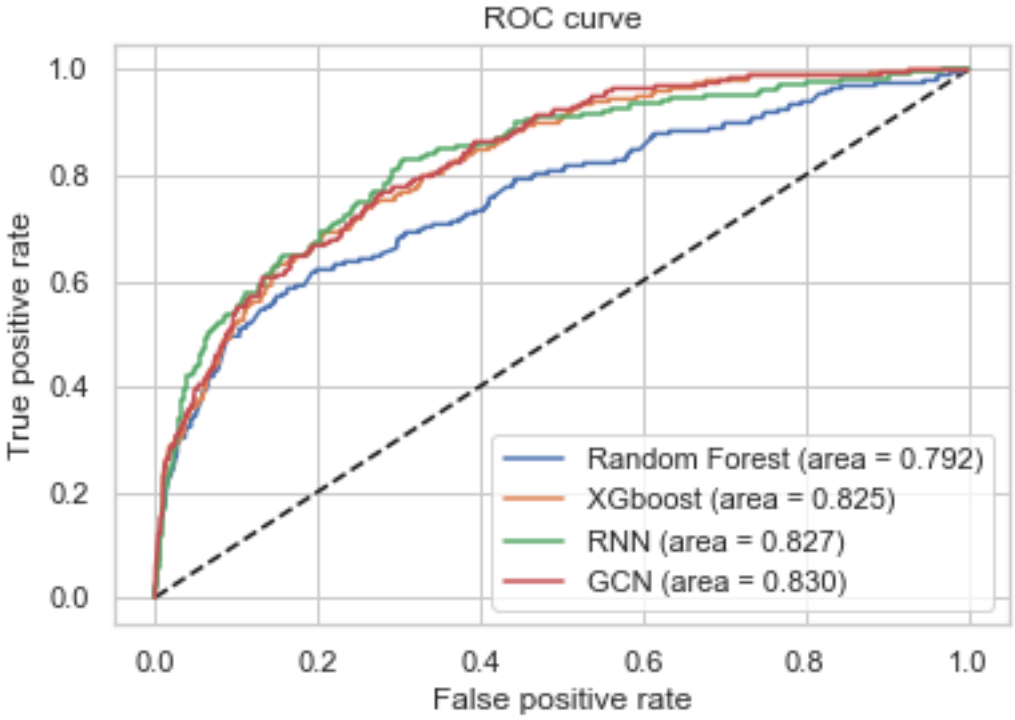
Figure 5.14 Confusion Matrix of four models

Table 5.1 lists the accuracy and ROC values calculated from the four prediction methods. Both of these measures are not recommended to be considered for performance evaluation in the case of an imbalanced data set; we are quoting these values here just for the comparison, as the results of the base model were reported in terms of area under ROC (AUROC). AUROC values obtained from other models show little improvement as compared to the base model. With the imbalanced class values or ratios, higher accuracy value can be obtained which may not give the correct predictions as the majority class (non-churn) will outweigh the minority class (churn).

5.2.2 Area under the ROC Curve

Figure 5.13 illustrates the ROC curves for the four prediction models including the baseline model. AUC values of XGboost and RNN models are same, depicting equal model performance, however the difference of predicting the churning customer is obvious in the confusion matrices and the precision recall curve values. XGBoost and the deep learning based models have predicted well while increasing the true positive values and decreasing the false positive rate.

Figure 5.15 Area under the ROC Curve



5.2.3 Area under the Precision-Recall Curve

In this section, we present the plot precision-recall curves for our four models. Precision-recall values for churned class are slightly better for RNN, xgboost than the base model.

Figure 5.14 plots the area under curve value for precision-recall curve for the base model

i.e. the random forest classifier.

For an imbalanced data set, to get the prediction of a rare event, the threshold value for the classification needs to be changed rather using the default value. Adjustment of the threshold value will affect the F1-score and increase or decrease either of the precision or recall value. To view the effect of varying threshold value on precision and recall values, the precision and recall curves are plotted separately in Figure 5.17 against multiple threshold values. This shows a trade-off over a range of threshold values while giving an option to the business analyst to select a threshold value for acceptable values of precision or recall, and prediction results. Figure 5.16 shows the area under precision-recall curve (PR-curve) calculated from the random forest model and 5.18 shows area PR-curve of the XGBoost model. In our study, our target is to improve the area under PR-curve using multiple prediction models, and the results show that XGBoost model has improved this value against the Random Forest model. Figures 5.17 and 5.19 illustrate the area under curves of precision and recall plotted (along y-axis) against multiple threshold values plotted at x-axis for random forest and XGBoost models respectively. In the case of random forest, the cross over point of threshold for precision and recall lines is 0.38 and for XGBoost it is about 0.82, which is approximately double the value from the random forest model. The difference in the values shows the difference in the tree formation criteria of these two models. The optimal threshold value is 0.85 in XGBoost where the F1-score will be maximum with 0.28. And the threshold for maximum F1-value (which is 0.25) for random forest comes out to be 0.42.

Figure 5.16 Precision-recall plot for Random forest

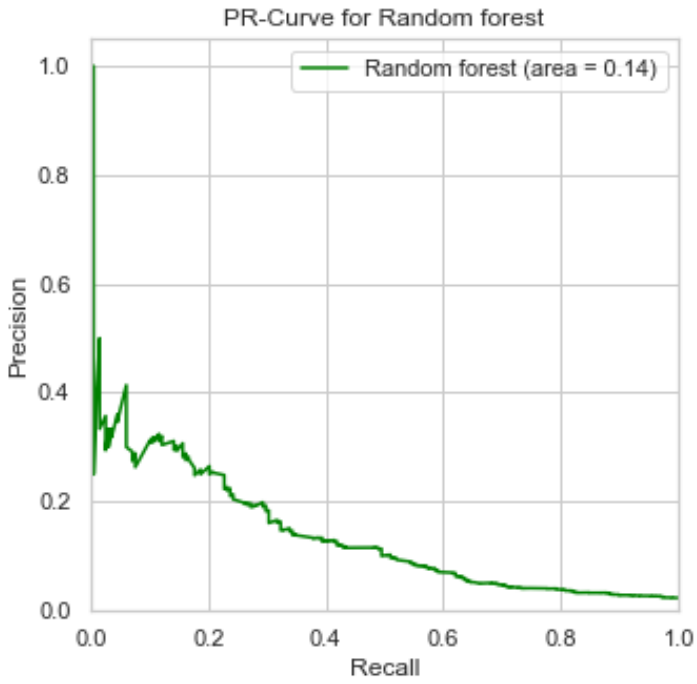


Figure 5.17 Precision-Recall threshold plot for Random Forest

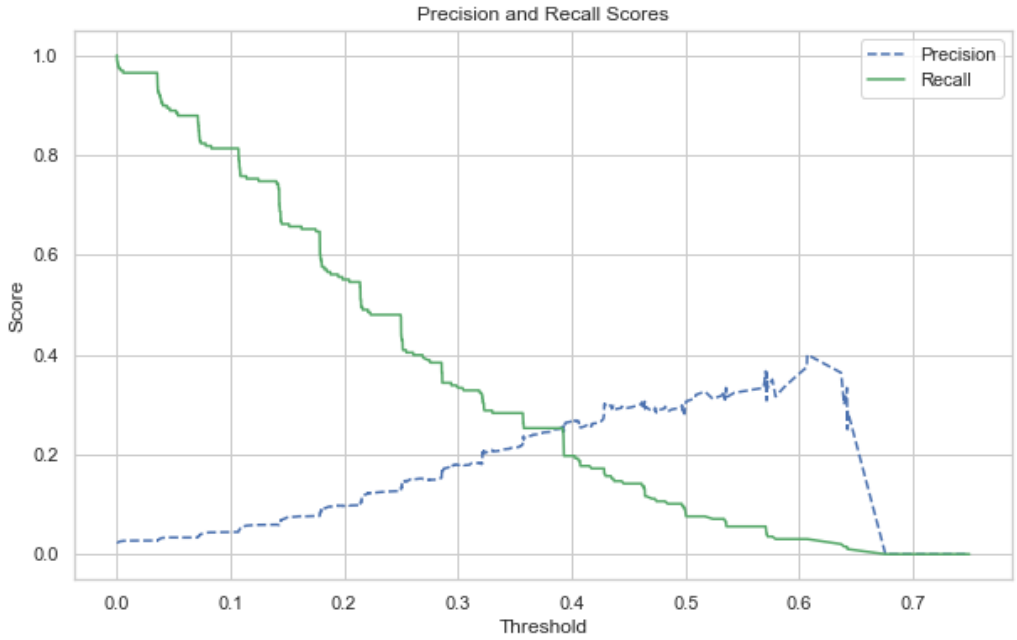


Figure 5.18 Precision-Recall plot for XGBoost

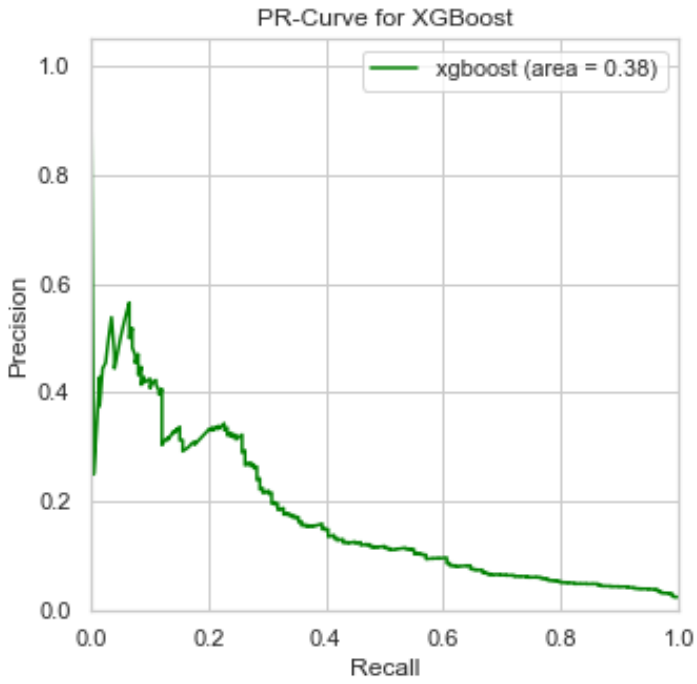


Figure 5.19 Precision-Recall threshold plot for XGBoost

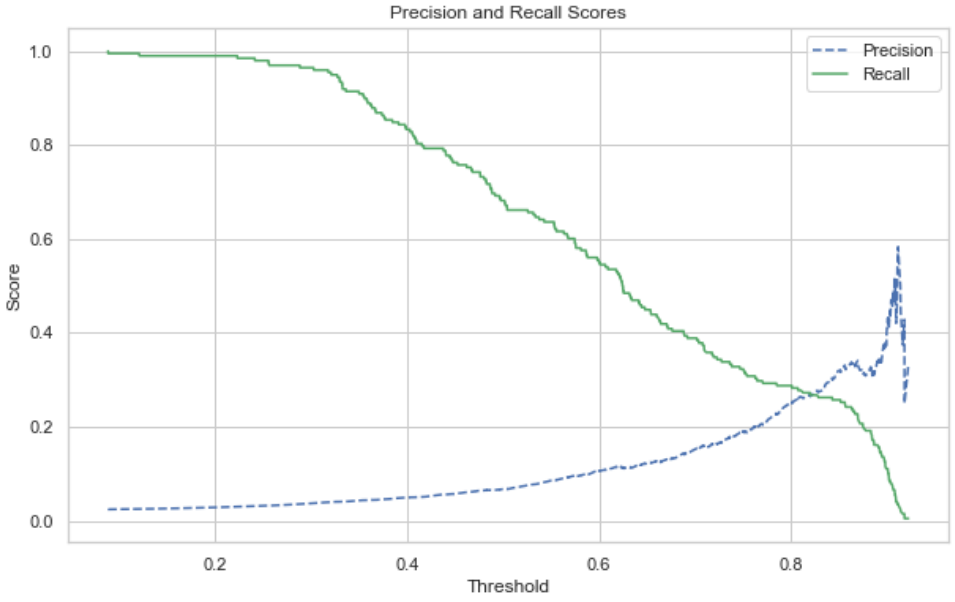


Figure 5.20 shows the precision-recall plots for the Recurrent Neural Networks prediction model. The area under curve value is 0.34 which is better than the Random Forest model

but slightly less than the XGBoost model. However, the decision threshold for RNN (is 0.75) which is less than the XGBoost and F1-value at this point is 0.22. The precision-recall curve value for churning customer in case of graph network is 0.23 as shown in Figure 5.22 which is less than the Recurrent Neural Network model. The reason may be due to the presence of sparse network connections between the customers.

Figure 5.20 Precision-recall plot for RNN-LSTM

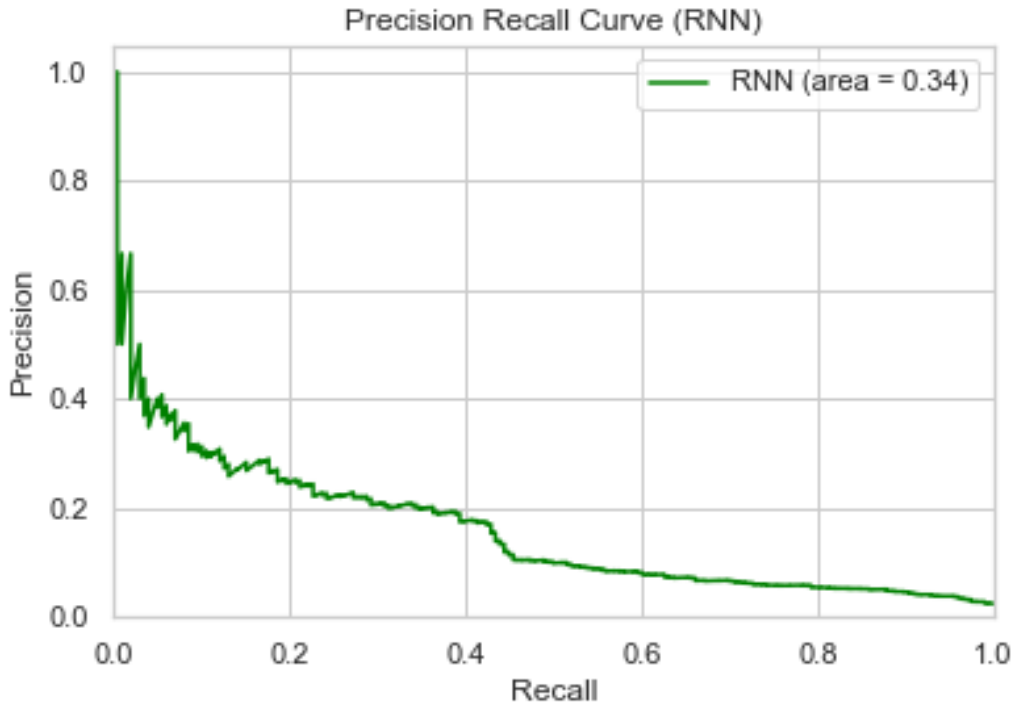


Figure 5.21 Precision-Recall threshold plot for RNN

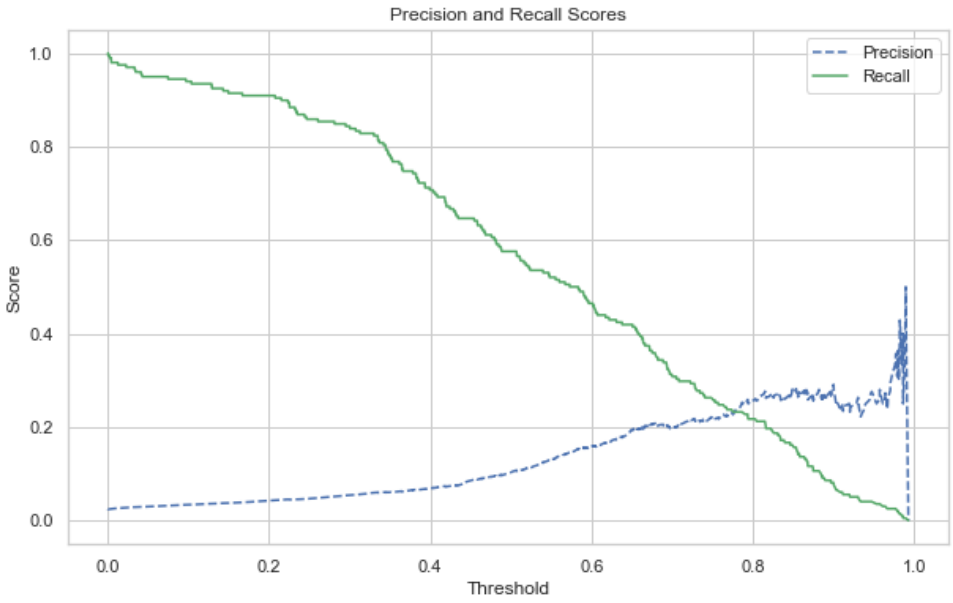


Figure 5.22 Precision-Recall plot for GCN

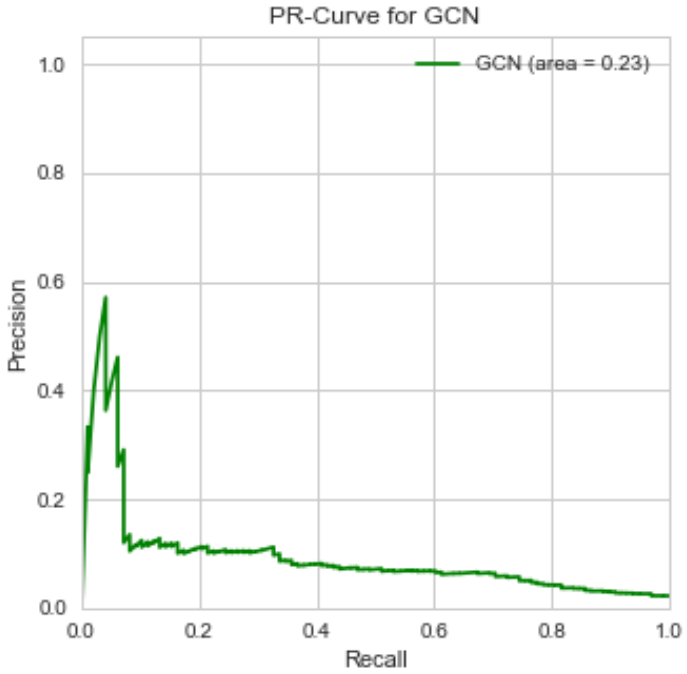
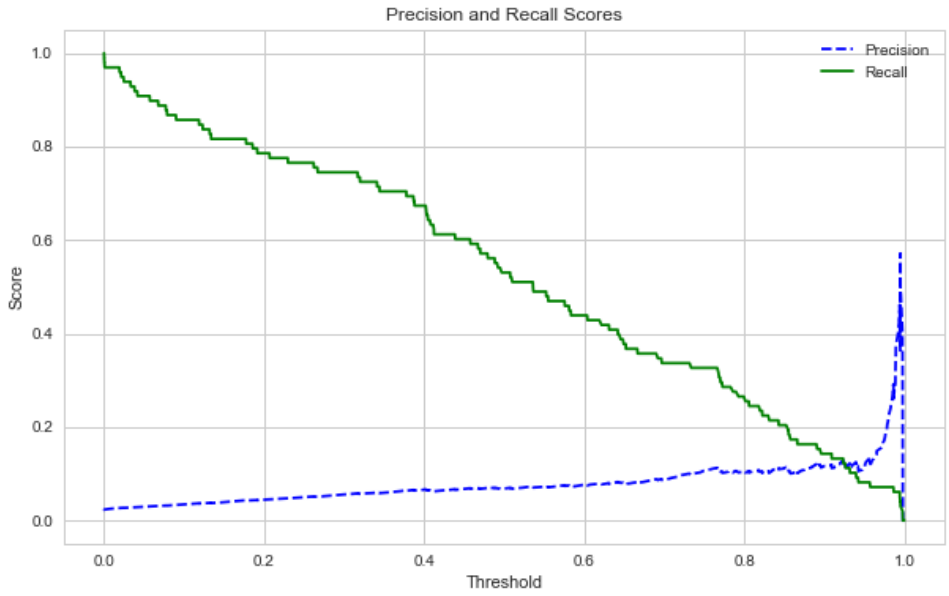


Figure 5.23 Precision-Recall threshold plot for GCN



5.3 Feature Importance and Dimensionality Reduction

In large data sets, where the number of attributes or features is quite large, selection of significant features which contribute the most in the performance of the prediction algorithm is an important step. Data with a large number of features are called high-dimensional and modeling with large number of attributes may lead to what is known as "curse of dimensionality" (Guyon and Elisseeff [2003]). Reducing the number of features for the learning model shortens the model training time, reduces the model complexity and reduces the chances of overfitting. In our experiments along with the given data attributes such as the demographic features, we have derived features from the transactions data. There are 51 attributes in total used in the gradient boosted trees (XGBoost) model. It is obvious that all the attributes do not have the same importance and contribution in the prediction model. We have employed a dimensionality reduction technique known as the Principal Component Analysis (PCA) and from the variance of the components it is revealed that most of the variation in the data can be explained by 20 components using a threshold value for component selection. Feature dimensions are reduced using PCA in

such a way that the new features are independent and not correlated.

In gradient boosted trees, the feature scores are retrieved after the tree creation, which shows how valuable each feature is in building the tree for the learning model. The feature importance rank is the average value of all the trees created during the training phase. We have used the built-in function of XGBoost to get the feature importance scores and their contribution value in the model. Similarly, important features of the random forest model plotted in Figure 5.22 shows regularity (weekly, hourly, grid wise), diversity (grid wise, weekly, radial weekly), age, bank age, fund transfer options, loyalty as the contributing features in the prediction.

Figure 5.24 Feature Importance - XGboost

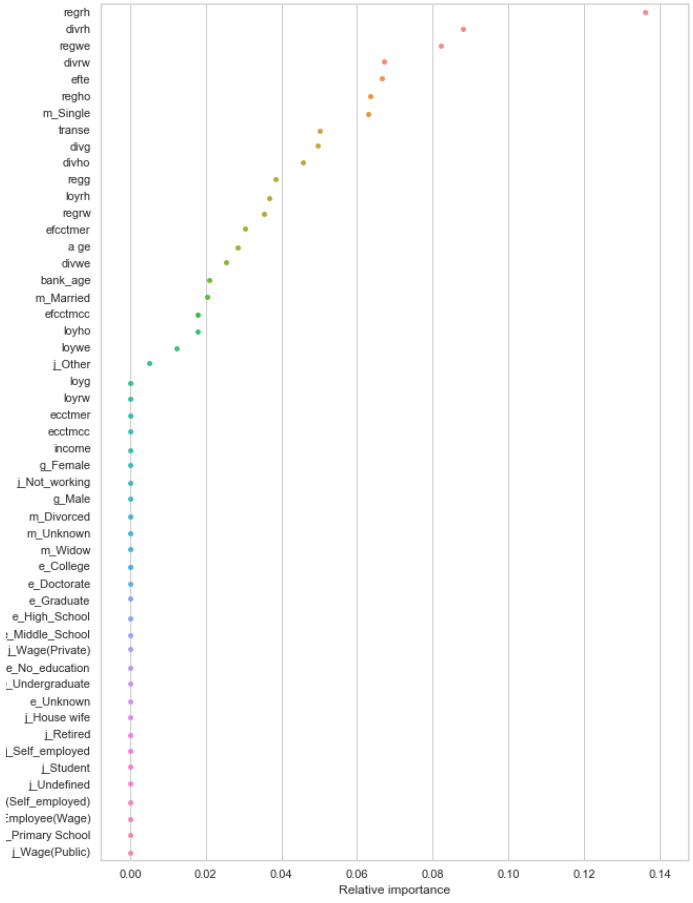
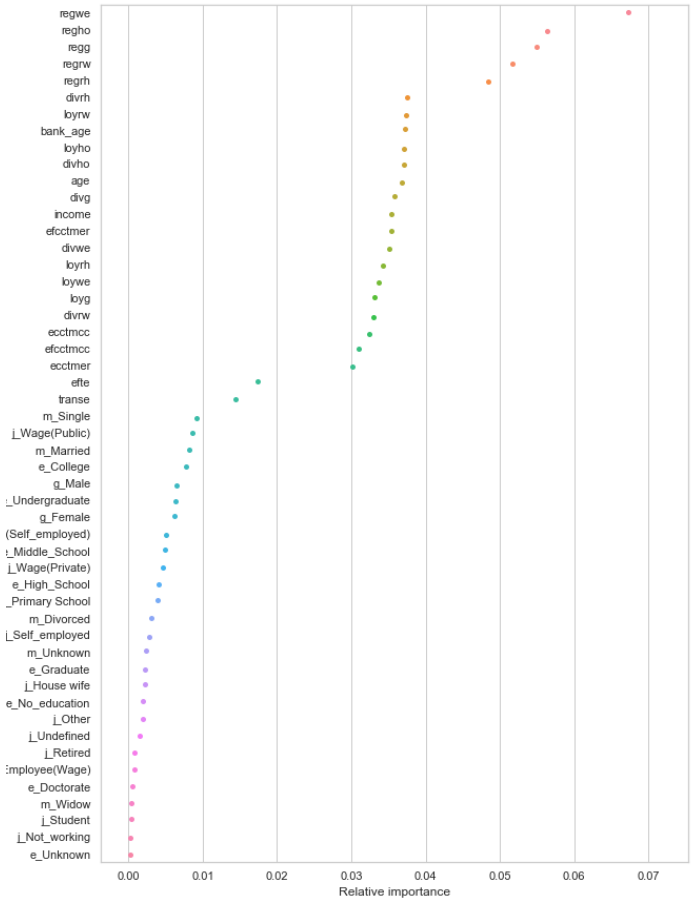


Figure 5.23 illustrates the feature importance plot for the gradient boosted tree model.

In the gradient boosted tree, regularity (radial with home location) is a more important feature than diversity (radial with home location), which seems logical as the regularity feature encapsulates both the diversity and loyalty values. Among the other important features, marital status (divorced, widow), customer age, age with bank, diversity and regularity (weekly), electronic amount transfer and loyalty features are included. Some of the features are marked as important in both models, however marital status, job type and education-related features do not seem to impact the prediction model performances as much.

Figure 5.25 Feature Importance - Random Forest Model



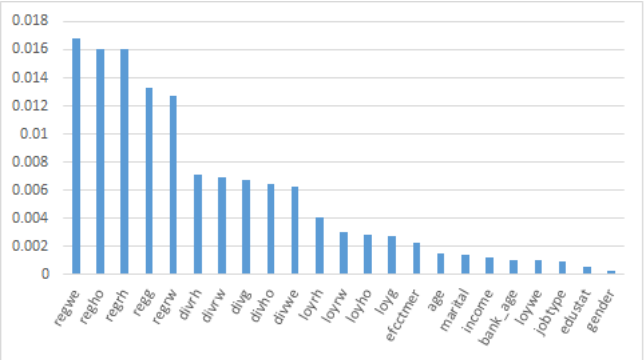
To find the contributing and uncorrelated attributes in our data set, we have employed two methods. First, by calculating the information gain or entropy value of each feature and

second by applying the correlation-based feature selection. Information gain or entropy of each feature is calculated against the target class and it gives the worth-score to each attribute between 0 (no information) and 1 (maximum information) using:

$$InfoGain(Class, Attribute) = Entropy(Class) - Entropy(Class|Attribute)$$

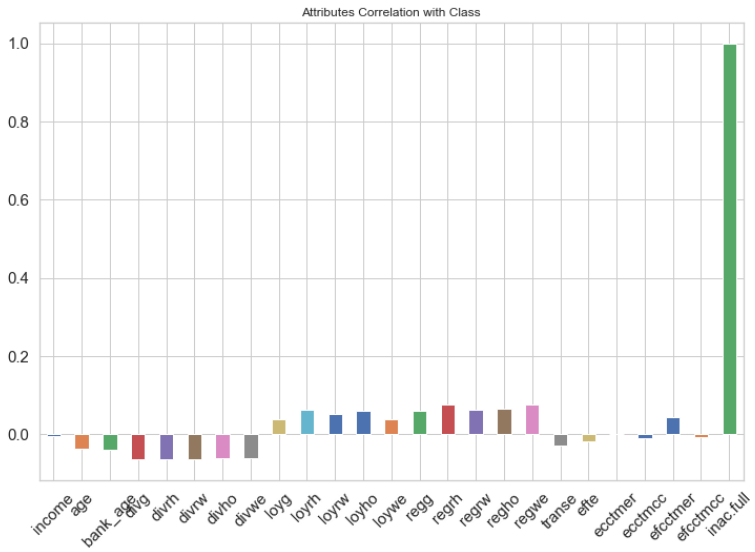
The attributes with zero or very low score have no contribution in the learning model. Figure 5.26 illustrates the ranking of the behavioral attributes with respect to their calculated information gain score.

Figure 5.26 Ranking of Behavioral Features based on Information Gain Value



In the correlation-based feature selection method, the correlation between each attribute and the output variable is calculated. The correlation score can be a positive or negative value depicting positive or negative correlation, and the attributes may have high positive or negative correlation values (values close to +1 or -1) or negligible correlation values (values equal to zero or close to zero). Figure 5.27 shows the Pearson's correlation coefficient values of the significant attributes.

Figure 5.27 Attributes Correlation Chart



In our study, we first attempted to replicate the churn prediction model of Kaya et al. [2018] while using one of the data set and the same behavioral features that they have used. Further, we analyzed and compared the prediction performance of a boosted tree algorithm and two deep-learning techniques with the base model. From the evaluation results, we see improvement in the values of recall, area under the ROC curve (AUROC) and the area under precision-recall curve. The new features and techniques explored for the churn prediction shows that the implicit behavioral patterns extracted from transaction data give better classification results than the baseline model. We have experimented with two deep learning techniques while using the sequenced and the graph network features that illustrates the customers' behavior. The results from our experiments prove that the recurrent neural network model using Long Short Term Memory (LSTM) methodology can be used for the classification tasks where data have temporal sequences. And the graph network model can learn and predict well while using the sparse connectivity data of customers in the multi-layered neural network model. We have observed that GCN model can perform comparable classification task while using the relatively fewer number of graph features.

6. Conclusion

Our study was aimed to predict the churning customers - who are going to leave the bank. The main contribution of our study is to develop a churn prediction model which assists the bank analysts' and management to predict the customers who are most likely subject to churn. Random forest classification technique is used as a baseline in our work for the comparison with gradient boosted tree (XGBoost) which is a decision tree based algorithm and two deep learning approaches i.e. Recurrent Neural Network (LSTM) and the Graph Convolutional Network (GCN).

We extracted and transformed the behavioral attributes from the transaction data of the customers to build the prediction models. Location and time based behavioral attributes (i.e. diversity, loyalty, regularity) which were used in earlier studies were transformed into sequence based features so that the dependency information can be revealed by using these attributes in the Recurrent Neural Network model. As Recurrent Neural Network based models depends on the sequential data to make a prediction, we have calculated customer's diversity, loyalty and regularity values over a constant time window along with the transaction frequency and spending amount patterns for our model. The use of sequence-based behavioral attributes enhanced the performance of Recurrent Neural Network model as illustrated in the results section. In our study, we have used customers' credit card transactions data for the calculation of quarter wise nine temporal features which reveals the behavioral information of the user while he/she is deciding to leave the bank. The results show that the Recurrent Neural Network model (LSTM) show comparable rather better performance than the base model. Based on the prediction results, it can be concluded, that the new features carry the dependency information from the sequential input data. The results of RNN (LSTM) model can be further improved by extracting more features which carries the temporal information of customers' behavior.

Our second deep learning based model uses the graph features which are calculated using the customers' transaction and the demographics data. We tried to build a prediction model, which utilizes these network based attributes where customers are defined as

the nodes and the attributes which explain the connections between customers are the connectivity points and edges of the network. Multi-layer graph network utilizes this network information to predict the churning customers. From the experiments, we obtain comparable prediction results from the graph network model. Use of the graph network attributes define the virtual connection between customers' and the results we have obtained in our experiments has confirmed that the customers' connections information contribute in the churn prediction. From the analysis of the graph features, customer co-churn terminology can be defined, such that if one of the customer who is a part of the network churn, than there is a possibility that the connected customer may churn in the near future. The virtual graph network of customers give insights of the community behavior and also reveal the community trend towards a service provider.

Both deep learning techniques have shown better prediction results as compared to the base model. The prediction performance of the gradient boosted decision tree (XGBoost) model and the deep learning models is almost the same, with minor advantages in the precision, recall and F1 values in either of the prediction model.

We have used three network connectivity dimensions to predict the churning customers, like the customers' who are either living in the same proximity or have similar patterns in the context of making purchases from same merchants and/or doing payments to each other. And the prediction result prove that the connectivity features contribute well to get better prediction results where data set is large and the connections among the customers are not strong enough. While we have defined customer connectivity using only three data dimensions, the prediction results can be more promising, if more features are extracted from other data sets like social media.

Based on the prediction results obtained from the graph network model, incorporation of social network analysis is suggested for future work. This will improve the prediction results as now people tends to share their experiences and opinions regarding service offerings on the social media, which creates a cumulative community perspective against a service provider. Finally, the decision of a customer to leave a company effects the overall community behavior and attitude towards a service provider. The deep learning based model can support the organizations to devise the retention plans for upcoming churning customers while observing the deviations in their transaction or service usage patterns. Our proposed prediction model is a contribution in the literature and towards the prediction models to define a framework of connected users, where the behavior of customers can be analyzed and the prediction of connected customers can be made earlier.

BIBLIOGRAPHY

Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588, 1997.

Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial intelligence in medicine*, 33(2):139–155, 2005.

Jonathan Burez and Dirk Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, 2009.

Longbing Cao. In-depth behavior understanding and use: the behavior informatics approach. *Information Sciences*, 180(17):3067–3085, 2010.

Mu-Chen Chen, Ai-Lun Chiu, and Hsu-Hwa Chang. Mining changes in customer behavior in retail marketing. *Expert Systems with Applications*, 28(4):773–781, 2005.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

Ding-An Chiang, Yi-Fan Wang, Shao-Lun Lee, and Cheng-Jung Lin. Goal-oriented sequential pattern for network banking churn analysis. *Expert Systems with Applications*, 25(3):293–302, 2003.

Mark Craven and Joseph Bockhorst. Markov networks for detecting overalpping elements in sequence data. In *Advances in Neural Information Processing Systems*, pages 193–200, 2005.

Piew Datta, Brij Masand, DR Mani, and Bin Li. Automated cellular modeling and prediction on a large scale. *Artificial Intelligence Review*, 14(6):485–502, 2000.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.

Leslie De Chernatony. *Creating powerful brands*. Routledge, 2010.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

Timm Euler. Churn prediction in telecommunications using miningmart. In *Proceedings of the workshop on data mining and business (DMBiz)*, 2005.

Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37, 1996.

Nicolas Glady, Bart Baesens, and Christophe Croux. Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197(1):402–411, 2009.

Clara-Cecilie Günther, Ingunn Fride Tvete, Kjersti Aas, Geir Inge Sandnes, and Ørnulf Borgan. Modelling and predicting customer churn from an insurance company. *Scandinavian Actuarial Journal*, 2014(1):58–71, 2014.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

John Hadden, Ashutosh Tiwari, Rajkumar Roy, and Dymitr Ruta. Computer assisted customer churn management: State-of-the-art and future trends. *Computers & Operations Research*, 34(10):2902–2917, 2007.

JF Hair Jr, MF Wolfinbarger, DJ Ortinau, and RP Bush. Essential of marketing research. mcgraw hill, 2010.

Benlan He, Yong Shi, Qian Wan, and Xi Zhao. Prediction of customer attrition of commercial banks based on svm model. *Procedia Computer Science*, 31:423–430, 2014.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge & Data Engineering*, (9):1263–1284, 2008.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Nan-Chen Hsieh. An integrated data mining and behavioral scoring model for analyzing bank customers. *Expert systems with applications*, 27(4):623–633, 2004.

Wagner Kamakura, Carl F Mela, Asim Ansari, Anand Bodapati, Pete Fader, Raghuram Iyengar, Prasad Naik, Scott Neslin, Baohong Sun, Peter C Verhoef, et al. Choice models and customer relationship management. *Marketing letters*, 16(3-4):279–291, 2005.

Jaya Kawale, Aditya Pal, and Jaideep Srivastava. Churn prediction in mmorpgs: A social influence based approach. In *2009 International Conference on Computational Science and Engineering*, volume 4, pages 423–428. IEEE, 2009.

Erdem Kaya, Xiaowen Dong, Yoshihiko Suhara, Selim Balcisoy, Burcin Bozkaya, et al. Behavioral attributes and financial churn prediction. *EPJ Data Science*, 7(1):41, 2018.

Abbas Keramati and Seyed MS Ardabili. Churn analysis for an iranian mobile operator. *Telecommunications Policy*, 35(4):344–356, 2011.

Muhammad Raza Khan, Joshua Manoj, Anikate Singh, and Joshua Blumenstock. Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty. In *2015 IEEE International Congress on Big Data*, pages 677–680. IEEE, 2015.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Clement Kirui, Li Hong, and Edgar Kirui. Handling class imbalance in mobile telecoms customer churn prediction. *International Journal of Computer Applications*, 72(23): 7–13, 2013.

VISWANATHAN Kumar and Denish Shah. Building and sustaining profitable customer loyalty for the 21st century. *Journal of retailing*, 80(4):317–329, 2004.

Jonathan Lee, Janghyuk Lee, and Lawrence Feick. The impact of switching costs on the customer satisfaction-loyalty link: mobile phone service in france. *Journal of services marketing*, 15(1):35–48, 2001.

Miguel APM Lejeune. Measuring the impact of data mining on churn management. *Internet Research*, 11(5):375–387, 2001.

Aurélie Lemmens and Christophe Croux. Bagging and boosting classification trees to predict churn. *Journal of Marketing Research*, 43(2):276–286, 2006.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*, 2015.

Sunil Mallya, Marc Overhage, Navneet Srivastava, Tatsuya Arai, and Cole Erdman. Effectiveness of lstms in predicting congestive heart failure onset. *arXiv preprint arXiv:1902.02443*, 2019.

David Martens, Foster Provost, Jessica Clark, and Enric Junqué de Fortuny. Mining massive fine-grained behavior data to improve predictive analytics. *MIS quarterly*, 40 (4), 2016.

Teemu Mutanen, Jussi Ahola, and Sami Nousiainen. Customer churn prediction-a case study in retail banking. In *Proc. of ECML/PKDD Workshop on Practical Data Mining*, pages 13–19, 2006.

María Óskarsdóttir, Cristián Bravo, Wouter Verbeke, Carlos Sarraute, Bart Baesens, and Jan Vanthienen. Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications*, 85: 204–220, 2017.

María Óskarsdóttir, Tine Van Calster, Bart Baesens, Wilfried Lemahieu, and Jan Vanthienen. Time series for early churn detection: Using similarity based classification for dynamic networks. *Expert Systems with Applications*, 106:55–65, 2018.

AO Oyeniyi, AB Adeyemo, AO Oyeniyi, and AB Adeyemo. Customer churn analysis in banking sector using data mining techniques. *Afr J Comput ICT*, 8(3):165–174, 2015.

Parag C Pendharkar. Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services. *Expert Systems with Applications*, 36(3): 6714–6720, 2009.

U Devi Prasad and S Madhavi. Prediction of churn behaviour of bank customers using data mining tools. *Indian Journal of Marketing*, 42(9):25–30, 2012.

R Prashanth, K Deepak, and Amit Kumar Meher. High accuracy predictive modelling for customer churn prediction in telecom industry. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 391–402. Springer, 2017.

Foster J Provost, Tom Fawcett, Ron Kohavi, et al. The case against accuracy estimation for comparing induction algorithms. In *ICML*, volume 98, pages 445–453, 1998.

Roland T Rust and Anthony J Zahorik. Customer satisfaction, customer retention, and market share. *Journal of retailing*, 69(2):193–215, 1993.

Jia-Lang Seng and TC Chen. An analytic approach to select data mining for business decision. *Expert Systems with Applications*, 37(12):8042–8057, 2010.

Vivek Kumar Singh, Burcin Bozkaya, and Alex Pentland. Money walks: implicit mobility behavior and financial well-being. *PloS one*, 10(8):e0136628, 2015.

Philip Spanoudes and Thomson Nguyen. Deep learning in customer churn prediction: unsupervised feature learning on abstract company independent feature vectors. *arXiv preprint arXiv:1703.03869*, 2017.

Chih-Fong Tsai and Yu-Hsin Lu. Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10):12547–12553, 2009.

Wouter Verbeke, David Martens, and Bart Baesens. Social network analysis for customer churn prediction. *Applied Soft Computing*, 14:431–446, 2014.

Chih-Hsuan Wang. Apply robust segmentation to the service industry using kernel induced fuzzy clustering techniques. *Expert Systems with Applications*, 37(12):8395–8400, 2010.

Jo-Ting Wei, Shih-Yen Lin, and Hsin-Hung Wu. A review of the application of rfm model. *African Journal of Business Management*, 4(19):4199–4206, 2010.

Yaya Xie and Xiu Li. Churn prediction with linear discriminant boosting algorithm. In *2008 International Conference on Machine Learning and Cybernetics*, volume 1, pages 228–233. IEEE, 2008.

Yaya Xie, Xiu Li, EWT Ngai, and Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

APPENDIX A

Data Visualization

Figure A.1 Gender distribution - Churn/ Non-churn Customers

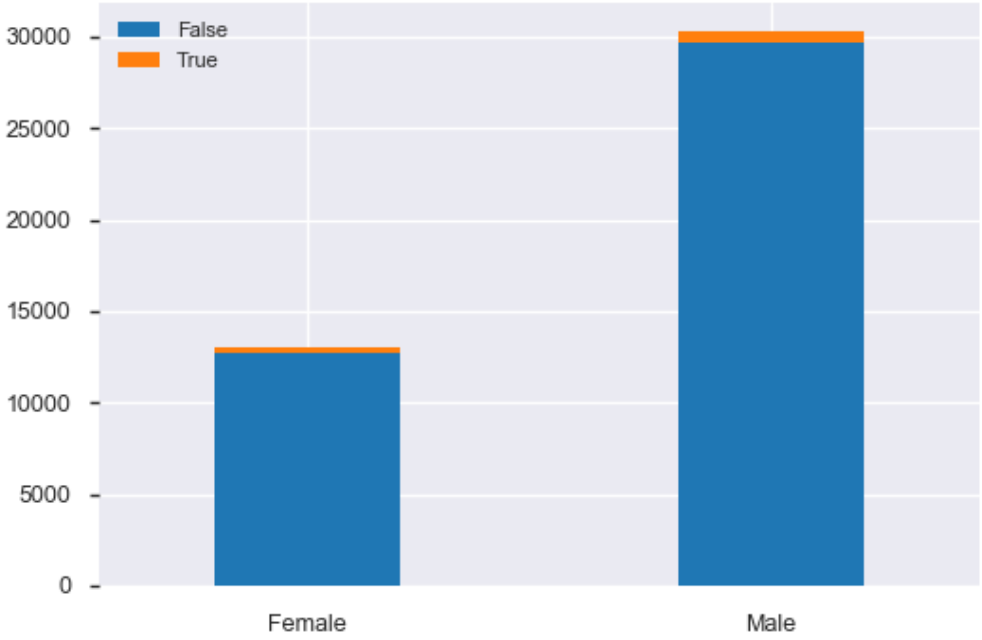


Figure A.2 Education level - Churn/ Non-churn Customers

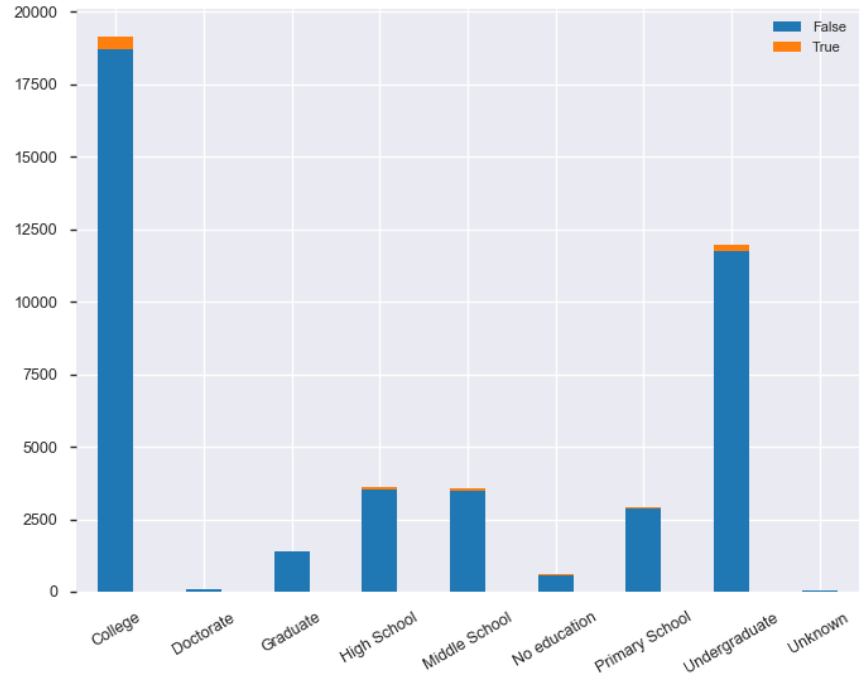


Figure A.3 Marital Status - Churn/ Non-churn Customers

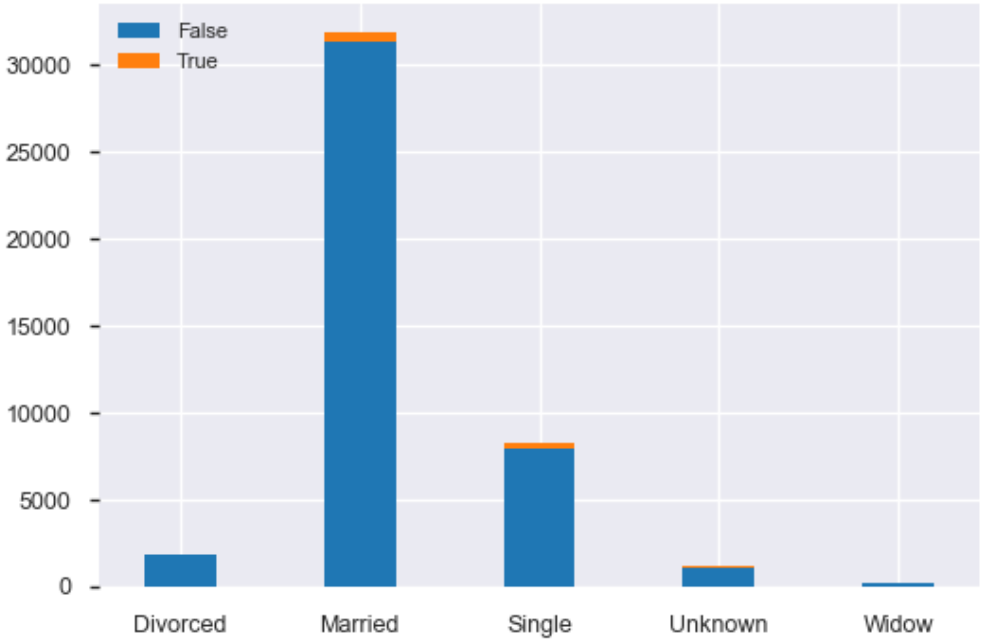


Figure A.4 Customer age with Bank

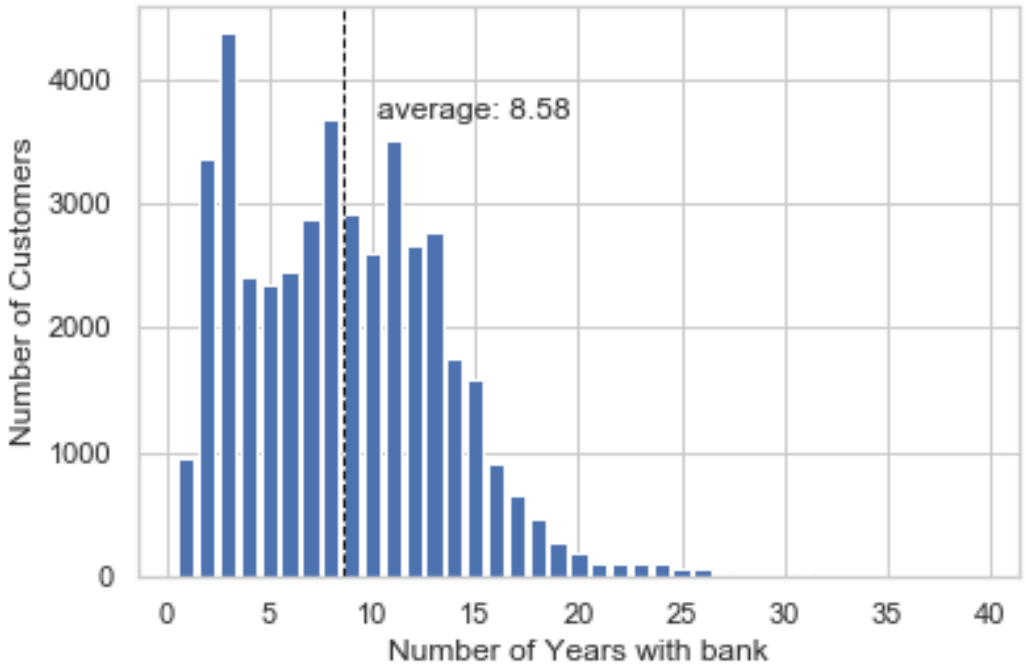


Figure A.5 Job Status distribution - Churn/ Non-churn Customers