PREDICTING FINANCIAL WELL-BEING
USING
BEHAVIORAL TRANSACTIONAL DATA

by

ANADIL MOHAMMAD

Submitted to the
School of Management
in partial fulfillment of the requirements for
the degree of
Master of Science
in
Business Analytics

Sabanci University
July 2018

PREDICTING FINANCIAL WELL-BEING
USING
BEHAVIORAL TRANSACTIONAL DATA

Sabanci University
School of Management

This is to certify that I have examined this copy of a master’s thesis by

ANADIL MOHAMMAD

And have found that it is complete and satisfactory in all respects, and that any and all revisions required by the final examining committee have been made.

Committee Members:

Prof. Burcin Bozkaya ………………………………

Assoc. Prof Abdullah Dasci ………………………………

Asst. Prof. Özay Özaydın ………………………………

Date: ……………………………………..

ABSTRACT

PREDICTING FINANCIAL WELL-BEING
USING
BEHAVIORAL TRANSACTIONAL-DATA

ANADIL MOHAMMAD

M.Sc. Thesis, July 2018

Supervisor: Prof. Burcin Bozkaya

Keywords: spatio-temporal mobility, overspending, trouble, late payment, shopping, channel, bagging, entropy

The recent introduction of using customers' spatio-temporal mobility patterns to predict their financial well-being proved to show significant results when examined on an OECD country’s bank data. In this research, we attempt to validate the same concept using another large bank’s transactional data set and see if it can be generalized. We examine a 1-year dataset spanning 2014 and 2015, calculate the relevant features from the literature and run prediction models using the bagging algorithm. The results show that the models built on spatio-temporal mobility features are still significant when predicting a customer's overspending and the status of financial trouble. In the case of late credit card payments as signs of financial trouble, demographics prove to be more significant than the spatio-temporal mobility features. We conduct further analysis to introduce new input variables related to shopping and channel categories, in an effort to improve the prediction accuracies of these models. The results show that among all the new features we experiment with, shopping categories used as an entropy variable and used as a binary indicator variable were the most significant ones in predicting overspending. The results of this study further validate that spatio-temporal mobility and other behavioral features can successfully predict financial well-being across different datasets, and hence can be used by decision makers in the financial industry.

ÖZET

PREDICTING FINANCIAL WELL-BEING

USING

BEHAVIORAL TRANSACTIONAL DATA

ANADIL MOHAMMAD

Müşterilerin finansal refahlarını tahmin etmek için mekânsal-zamansal hareketlilik modellerinin kullanımı yakın tarihli bir çalışmada bir OECD ülkesinin banka verisi üzerinde önemli sonuçlar ortaya çıkardı. Bu araştırmada, aynı konsepti başka bir büyük bankanın işlemsel veri setini kullanarak doğrulamaya ve genelleştirilip genelleştirilemeyeceğini araştırıyoruz. 2014 ve 2015 yıllarını kapsayan 1 yıllık bir veri seti üzerinde literatürdeki ilgili mekânsal-zamansal endeksleri hesaplayarak ve torbalama algoritmasını kullanarak tahmin modellerini çalıştırdık. Sonuçlar, bir müşterinin aşırı harcama ve mali sıkıntı durumunu tahminlemede, mekansal-zamansal hareketlilik özelliklerinin tahmin modellerinde hala etkin olduğunu göstermektedir. Geç kredi kartı ödemelerinde finansal sıkıntı belirtileri olması durumunda, demografik özelliklerin mekansal-zamansal hareketlilik özelliklerinden daha etkin olduğu kanıtlanmıştır. Bu modellerin tahmin doğruluğunu iyileştirmek amacıyla, alışveriş ve kanal kategorileri ile ilgili yeni girdi değişkenleri tanımladık ve analizlerimizi genişlettik. Sonuçlar, denediğimiz tüm yeni değişkenler arasında, entropi değişkeni olarak kullanılan ve ikili gösterge değişkeni olarak kullanılan alışveriş kategorilerinin, aşırı harcamaları tahmin etmede en etkin olanlar olduğunu göstermektedir. Bu çalışmanın sonuçları, mekânsal-zamansal davranışsal özelliklerin farklı veri setlerinde finansal refahı başarılı bir şekilde tahmin edebildiğini ve dolayısıyla finansal sektördeki karar vericiler tarafından kullanılabileceğini doğrulamaktadır.

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Over the years, different methods of assessing a customer’s financial well-being or credit risk by a bank have been developed. The advent of digital technologies and the growing popularity of big data have also contributed towards the advancement in this assessment. With the availability of numerous and such extensive data sources, banks realized the relationship between an understanding of their customer and making effective business decisions. Mobile banking development and digitalization of expenses have made banks one of the largest sources of data. Yet the enormous data related to the demographics, finance, and mobility of the customer is meaningless if not analyzed, (Kung, Greco, Sobolevsky, & Ratti, 2014). Methodologies born from computational social science presented opportunities for performing analysis of human shopping behavior across the domains of time and space, and in turn allowed the understanding of its relationship with financial outcomes, (Lazer, et al., 2009) and (Singh, Bozkaya, & Pentland, 2015). Hence banks began to dig deeper into what influences a customer to overspend, become delinquent, and to miss payments. The insights produced from this were in turn used to allow banks to improve “good lending”, or to lend to customers who were good borrowers, (Corsetti, Pesenti, & Roubini, 1999). During the initial years, banks were more dependent on the use of demographic characteristics of the customer. Now, there has been a shift from using just demographic-related variables (e.g. age, gender, marital status, job status, income) to using ones which are more related to the customer’s transactional habits and mobility. One benefit of this is that while demographic characteristics provided by the customer can fall prey to intended fraud and unintended ambiguity (e.g. customer might fake their job status or marital status), the shopping data coupled with geographical information generated by the merchant’s and bank’s automated systems (e.g. ATMs) is highly accurate and transparent. Also, mobility is harder to manipulate as compared to an individual’s payment history or economic profile. Some studies may even use data related to phone call timestamps, check-ins at cafés and restaurants, and router locations of wi-fi’s an individual is connected to. For example, Noulas et al. (2012) utilized the check-in data generated by Foursquare (a local search-and-discovery mobile application) and Go Walla to predict the next potential location of the user. They discovered that across 11 different cities, around 60% to 80% of users’ visits were in places which they had not visited previously in the last 30 days, (Noulas, Scellato, Lathia, & Mascolo, 2012). Another example is by Isaacman et al. (2010) who used aggregate and anonymous statistics of the estimated locations of thousands of cell phones in New York City and Los Angeles to illustrate the contrasting mobility patterns between both the cities. Some of their findings were that the residents of Los Angeles had median daily travel distances around two times greater than their New York counterparts, and also that the most mobile New Yorkers travelled on average six times farther than Angelenos, (Isaacman, et al., 2010). Another instance is by Sobolevsky et al. (2014) who explored a bank’s transaction data to understand the relation between an individual’s nationality and their mobility patterns. Hence, the future implications of this might be that loan-seeking customers might be able to prove their credit-worthiness on the basis of their mobility footprints, rather than through collaterals and financial statistics.

One study which has managed to prove a strong relationship between a customer’s spatio-temporal mobility and their credit risk is by Singh et al. (2015). This paper was inspired by animal behavior studies which discovered important connections between an animal’s “spatio-temporal foraging behavior and their life outcomes”, (Singh, Bozkaya, & Pentland, 2015). It is also the first publicly reported study to examine detailed data of a customer’s current shopping patterns with the aim of predicting financial outcome. Based on this idea, the authors analyzed thousands of economic transactions belonging to the customers of a bank in an OECD country. They discovered that an individual’s financial outcomes are strongly linked to his or her spatio-temporal traits such as elasticity, exploration, and engagement, (Singh, Bozkaya, & Pentland, 2015). The results also showed that the spatio-temporal features created based on the existing dataset proved to predict a customer’s potential financial difficulties 30%-49% better than rival demographic models, (Singh, Bozkaya, & Pentland, 2015). Chapter 3 of this thesis gives a clear explanation as to what these features and financial difficulty indicators are.

In this thesis, an effort is made to explore whether the model by Singh et al. (2015) can be generalized for other banks’ customers. This would prove whether there is indeed a strong link between a customer’s mobility and their financial outcomes, hence giving more strength to the generalization of the model and its use in potentially other countries. This would also increase the chances of the results obtained to be consistent across different cultures or banking customer profiles. To ensure consistency, the bank whose dataset is used for this paper also belongs to the same OECD country. Whereas the study by Singh et al. (2015) uses data from 2013, this thesis uses data from July 2014 to June

2015. It was also ensured that the filters and manipulations applied on the former dataset was also applied to the dataset used in this paper.

The second part of this thesis investigates whether the model by Singh et al. (2015) can be improved further by adding more features related to the purchasing habits of customers and their interaction with the bank. Several experiments were carried out with information related to the channel(s) used for interacting with the bank, and shopping categories (e.g. insurance, food, gasoline) appearing in customer transactions. Chapter 2 offers the relevant literature and background on credit risk assessment methodologies, human mobility and its relation to financial behavior, spending categories and their relation to financial behavior, and various classification algorithms which have been used. The performance evaluation criteria, the feature selection/extraction methods, their preparation and manipulation, and the classification algorithms used are presented in Chapter 3. The computational experiments and deductions are examined in Chapter 4. Chapter 5 concludes the thesis and summarizes the main contributors and inferences we obtain during the study.

We believe that findings from this research can be used to provide feedback to banks about how to identify risky customers. For example, if it is found that spending more on alcohol and cigarettes has a strong correlation with paying credit card dues late, then those customers can be highlighted by the bank and suitable remedial action can be taken. In turn, the results can also be of significance to individuals who can be alerted if their credit worthiness is at risk of deteriorating. Our study is expected to benefit analysts and decision makers in the finance and related industries for taking actions towards reducing or pre-empting risky customers, hence increasing financial well-being of organizations as well as individuals.

Chapter 2

LITERATURE REVIEW

2.1. Credit Risk Assessment Methodologies

The market for consumer credit is constantly experiencing unpredictable changes, along with other new challenges and increased competition. To keep up with these developments, highly advanced mathematical and statistical tools are being adopted. Not only are these tools used to differentiate between good and bad risks, but also to characterize different customer behavior patterns, and monitor customer performance, at both the portfolio and individual level, (Hand, 2001). Consumer credit is generally granted by various lending institutions such as banks, retailers, building societies, and other such organizations. Credit risk assessment was traditionally conducted using human judgement and experience of past decisions to determine if the credit applicant should be granted credit or not, (Henley & Hand, 1996). It was common for the assessment process adopted by the institutions to involve the usage of methodologies such as linear or logistic regression, discriminant analysis, decision trees, and linear programming, (Hand & Henley, 1997). The output of the credit assessment process is usually a credit score, which is a term used to signify the use of statistical methods for evaluating applications for credit into majorly two risk classes: good and bad, (Hand & Henley, 1997). The process determines the potential of an application to default on their payments. Due to the rapid growth in consumer credit over recent years, these methods have become increasingly important and researchers are constantly developing new methods of risk measurements.

In 1996, Henley and Hand assessed the creditworthiness of applicants for consumer loans using the k-nearest neighbor method (k-NN). They developed a different version of the Euclidean distance metric which incorporated knowledge about class separation inherent in the data, (Hand & Henley, 1997). They also explored how to select optimal values of the parameters included in the method such as k and D. They eventually discovered that the k-NN classification was quite insensitive to the parameters chosen, and the bad risk curves against the k parameter showed flat valleys, (Henley & Hand, 1996). When this was compared to the performance of other techniques such as linear and logistic regression and decision trees, the k-NN method proved to perform better and achieved the lowest expected bad risk rate, (Henley & Hand, 1996). They were also able to further improve the assessment results by using the adjusted Euclidean metric, as compared to the standard Euclidean metric, (Henley & Hand, 1996). Practically it was possible to use the k-NN method to justify the reason for refusing credit to an application, which also satisfied legal requirements. In addition, the k-NN method was superior to traditional score-card techniques because of its ability to be updated given any changes in the population, (Henley & Hand, 1996).

Within the next decade, several academicians explored various other ways of modeling credit risk. In 2001, Hand tried to model consumer credit risk by utilizing statistical tools such as logistic regression, naïve Bayes, recursive partitioning models, and neural networks, (Hand, 2001). In the same year, Christiansen et al (2001) invented a credit risk assessment method which used a variety of segments to group and classify credit applicants according to their credit risk, (Washington, DC Patent No. US 6,202,053, 2001). The segments were based on reported trades, bank card utilization, reported delinquency, and credit history length. For each segment, a unique scorecard was designed and based on this, a score was generated for each applicant. This allowed for more accurate credit risk assessment since each applicant was evaluated against each segment's likelihood of being a bad credit risk, (Washington, DC Patent No. US 6,202,053, 2001).

In 2006, Crook et al. undertook a research into the recent developments in consumer credit risk assessment. They discovered that the most popular method for classifying applicants into those likely to repay or not repay was logistic regression, and then comparing the logit value to a cut off or threshold, (Crook, Edelman, & Thomas, 2007). Despite the logistic regression's popularity, they found that the most accurate method was the support vector machines. Furthermore, they stated that due to the substantial growth of risk assessment and credit scoring techniques, groups such as institutions, consumers and the economy were able to experience various benefits: easy and quick assessment made possible for the consumers to obtain credit and loans on time, improvement of the lifestyles of scores of people around the world, competition increase in credit markers, and the consequent reduction in the cost of borrowing, (Crook, Edelman, & Thomas, 2007).

In 2010, further advancements in credit risk assessment appeared. Khandani et al (2010) applied machine-learning techniques for analyzing consumer credit risk. They developed nonlinear nonparametric forecasting models for a certain commercial bank and based on a sample of customers, combined credit bureau data and customer transactions from January 2005 to April 2009, (Khandani, Kim, & Lo, 2010). Consequently, they developed forecasts for a test sample of customers which greatly improved the classification rates of credit-card holder defaults and delinquencies, achieving an R^2 of 85% for forecasted vs. realized delinquencies, (Khandani, Kim, & Lo, 2010). The model produced accurate forecasts for credit events in advance of 3 to 12 months. Based on the model's results, the authors predicted that by cutting certain credit links the bank would be able to experience cost savings of between 6% to 25% of current total losses, (Khandani, Kim, & Lo, 2010).

In the same year, Khashman (2010) developed a credit risk evaluation system which used supervised neural network models based on the back-propagation learning algorithm. To decide whether to approve or reject a credit application, he trained and implemented three neural network models, (Khashman, 2010). He investigated 9 learning schemes with different training-to-validation ratios and compared their implementation results. The finding was that the neural network performed best under LS4 in which 400 cases of training and 600 cases for validation were used. While the overall accuracy rate was 83.6%, accuracy rates using the training and validation set were 99.25% and 73.17% respectively, (Khashman, 2010).

Similarly, Constangiora (2011) discovered that complex non-linear estimations were more superior in accuracy when he used statistical modeling to forecast the default probabilities of a dataset of consumer loan applicants. He also found that the bagging model produced better results than the neural network model and traditional tree and logit estimations, (Constangiora, 2011). Furthermore, he proposed a statistical scorecard which offered a 60% improvement compared to the baseline model. His recommendations included that the bank's management should set up a decisional probability threshold in line with its propensity to risk.

Kruppa et al. (2013) also studied default probabilities and found that they offer detailed information regarding consumer creditworthiness. The authors stated that machine learning techniques could be used for the consistent estimation of individual consumer credit risks, (Kruppa, Schwarz, Arminger, & Ziegler, 2013). They also demonstrated probability estimation in Random Jungle, a fast-random forest implementation. The findings proved that random forests outperformed a tuned logistic regression on large credit scoring dataset. They also suggested that machine learning methods should be considered serious competitors of classical models since their implementations are fast, reliable, and simple-to-use, (Kruppa, Schwarz, Arminger, & Ziegler, 2013).

2.2. Human Mobility and its relation to Financial Behavior

Parallel to studies on credit risk assessment, academicians also began to take notice of human mobility and its relation to various life outcomes. In 2008, Gonzalez et al studied the trajectory of 100,000 anonymous mobile phones users whose movements were tracked for a period of six months, (Gonzalez, Hidalgo, & Barabasi, 2008). They discovered that human trajectory shows a high level of spatial and temporal regularity, and a high probability of returning to a few mostly visited locations, (Gonzalez, Hidalgo, & Barabasi, 2008). They also found that the individual travel patterns collapsed into a single spatial probability distribution signifying that despite the diversity of their travel history, humans generally follow simple reproducible patterns, (Gonzalez, Hidalgo, & Barabasi, 2008). In the end, the study implied that similarity in movement patterns could affect all phenomena driven by human mobility such as emergency response to epidemic intervention, agent-based modelling, and urban planning, (Gonzalez, Hidalgo, & Barabasi, 2008).

Another very significant study was conducted by Lazer et al (2009) who discussed the emergence of computational social science, which is the use of computers to model, simulate, and analyze social phenomena. They stated that people's everyday transactions leave numerous 'digital breadcrumbs' which are harbored by data sources such as Google, Yahoo, phone companies and social networking sites. Whereas previous research on human interactions relied mainly on one-time self-reported data on relationships, Lazer et al. proposed that it was now possible to use peoples' digital footprints (based on their movements and physical proximities) to understand cognitive relationships and even the potential spread of disease in a certain community, (Lazer, et al., 2009).

In 2014, Sobolevsky et al. proposed a new and consistent way of developing mobility networks using transactional data. They demonstrated and studied the potential of a new type of extensive data, i.e. bank card transactions executed by both domestic and foreign customers of a Spanish bank. They performed a quantitative study of the impact of tourists' nationality on their mobility behavior and discovered a consistent and positive relationship between the distance from a given country to Spain, and the mobility characteristics of the visitors coming from the said country, (Sobolevsky, et al., 2014).

In one of the latest studies on human mobility, Singh et al. (2015) developed a new model based on several individuals' spatio-temporal mobility for assessing the financial well-being of a certain bank's customers. While traditional assessment systems relied more on a customer's demographic characteristics, e.g. gender, age, marital status, and job type, the authors instead proposed a new system which incorporated information regarding human consumption patterns across space and time, (Singh, Bozkaya, & Pentland, 2015). The study was based on three months of credit cards transactions which took place in 2013. The model's main idea was founded on studies of animal behavior, in which significant relationships between animal foraging behavior and their life outcomes were found. The authors developed a set of 12 features signifying shopping behavior: *spatial radial diversity, spatial radial loyalty, spatial radial regularity, spatial grid diversity, spatial grid loyalty, spatial grid regularity*, *temporal weekly diversity, temporal weekly loyalty, temporal weekly regularity, temporal hourly diversity, temporal hourly loyalty, and temporal hourly regularity*, (Singh, Bozkaya, & Pentland, 2015). They also developed a set of output variables signifying credit risk: *overspending, financial trouble, and late payment,* (Singh, Bozkaya, & Pentland, 2015). The study produced several findings: (1) an individual's financial outcome is intricately linked with his or her spatio-temporal traits like exploration, engagements, and elasticity, (2) models that use these features are 30%-49% better at predicting future financial difficulties than comparable demographic models, (3) the results obtained may have a higher likelihood of being consistent across cultures than those based on culture specific norms and customers, (4) mobility data is harder to manipulate as compared to social or economic profile which makes the model highly strong and efficient in predictions, (5) spatial diversity, loyalty, and regularity had high median scores on their curves which indicated a strong affinity for all three traits in human shopping behavior, (6) ROC area was greatest for all models with behavioral variables predicting financial trouble, overspending, and late payment, and (7) late payments and financial trouble variables had a positive correlation with low education and low age, and a negative correlation with male, and married customers, (Singh, Bozkaya, & Pentland, 2015).

Dong et al. (2016) further studied human purchase behavior at a community level and argued that people who live in different communities but work at close-by locations could potentially act as "social bridges" which link their respective communities and cause the purchase behavior in the community to be similar, (Dong, et al., 2018). To prove this, they studied millions of credit card transactions for thousands of individuals living in the city for over a period of three months. The findings showed that the number of social bridges between communities is a greater indicator of similarity in purchase behavior as compared to traditional factors such as socio-demographic and income variables. Other findings of the study suggested that (1) effect of social bridges can vary across different merchant categories, (2) the presence of female customers in social bridges is a stronger indicator compared to that of their male counterparts, and (3) geographical constraints exist for the effect of social bridges, making it vary across cities, (Dong, et al., 2018). In addition, they found that as the number of bridges between two communities increased, the number of co-visits they shared increased, they had closer temporal distributions of purchases, and had more similar median spending amount per transaction, (Dong, et al., 2018).

2.3. Spending Categories and its relation to Financial Behavior

Several researchers have studied the relation between shopping and financial behavior of the customers, (Hui-Yi & Nigel, 2012). The authors conducted experiments to compare the decision-making process of compulsive and non-compulsive shoppers. The experiments found that compulsive shoppers were more likely to overspend and were more encouraged to shop due to credit card availability, (Hui-Yi & Nigel, 2012). The authors also claimed that compulsive shopping is one of the reasons why compulsive buyers end up with such a large debt. Since credit cards allow individuals to borrow money easily, this influences them to overspend and become less conscious of their budgets, (Hui-Yi & Nigel, 2012).

In another study, Achtziger et al. (2015) noted that while compulsive buying was positively correlated with debt, self-control was negatively correlated with it. They also found that women and younger individuals were more likely to buy compulsively as compared to men and older age individuals, (Achtziger, Hubert, Kenning, Raab, & Reisch, 2015).

It would then be interesting to note the shopping categories which are common in compulsive shopping behavior. It is widely understood that compulsive shopping behavior is generally characterized by high expenditure levels on hedonic goods. Dhar and Wertenbroch (2000) state that hedonic goods are those whose consumption is characterized by a sensory and affective experience of sensual or aesthetic pleasure, fun, and fantasy. Furthermore, they found that product categories high in terms of hedonic value are more likely to be classified under "want preferences", while product categories high in terms of utilitarian value are more likely to be classified under "should preferences", (Dhar & Wertenbroch, 2000). Hence, it is assumed that items such as clothing, jewelry, restaurants appear under the hedonic shopping category, while items such as insurance, gasoline, and food are included in the utilitarian shopping category. Thus, it is expected that people who spend a greater percentage of their income on the former category are likely to overspend.

Chapter 3

DATA AND MODELING FEATURES

In this chapter, we first discuss the data preparation made for validating the model by Singh et al. (2015) using a new data set provided by a major bank in the same OECD country. In the second part, we introduce new behavioral features for improving the aforementioned model in an effort to achieve better prediction results.

3.1. Validation of the model by Singh et al. (2015)

The dataset considered in this thesis belongs one of the major banks of the same OECD country of interest. For practical reasons, we will refer to this bank as A-Bank in the sequel. Unique identifiers (e.g. citizenship number) and customer names were removed from the dataset before it was delivered to us in order to create an anonymized dataset for the study. The analysis and results reported in this thesis were performed on the de-identified and anonymized dataset.

The dataset under consideration consists of tens of thousands of personal accounts of individuals taken from A-Bank's data warehouse. A random sample of customers were sampled, along with their demographic information and credit card transactions for purchases made between July 2014 to June 2015. This amounted to around 4.05 million transactions for an estimated 20 thousand customers. For the same sample of customers, various other information related to shopping categories and communication channel usage were also provided.

The sampled data included the following information regarding the customer's demographics and their transactions:

- Customer's gender
- Customer's age
- Customer's education level
- Customer's marital status
- Customer's job status
- Customer's income
- Customer's home and work coordinates
- Transaction timestamp (date, hour, minute)
- Transaction amount
- Merchant's coordinates
- Customer's credit card statement details: Statement Date, Statement Due Date, Payment Date, and Statement Amount
- Information on risk codes assigned to each customer for each month in the analysis period

The above attributes were further processed into other measures and indicators for which the steps are described below. Most of these steps are the same as defined by Singh et al. (2015).

Data cleaning/characteristics/validation: A random sample of 20,000 individuals and information regarding their 4,058,641 transactions were selected from the database. Online transactions were excluded from the dataset (e.g. electronic funds transfer, remittances, etc.). Only individuals with more than 40 transactions for the entire 12-month period were considered. Also, customers who did not have valid coordinates for home and work were excluded from the dataset. To be included in the sample, customers' incomes had to be a non-zero figure. Some customers had job status listed as "Retired", "Housewife", and "Not Working" and yet they had valid work coordinates stated. Since this was illogical, the work coordinates for them were removed and only the coordinates for home were considered. In addition, customer's age was standardized through z-scores.

Features: Each customer's spatio-temporal behavior was measured based on the same three features defined by Singh et al. (2015). These three features are defined as follows:

- <u>Diversity:</u> when a customer's shopping experience varies significantly over space and time and hence the transactions are distributed equitably over several bins. The formula for calculating the entropy is:

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

  where p_{ij} are the fraction of transactions that fall within bin *j* for customer *i*, and *M* is the number of non-empty bins over which the customer's shopping experience is divided, (Singh, Bozkaya, & Pentland, 2015). The output values range between 0 to 1, where large numbers signify high diversity and low numbers signify low diversity.

- Loyalty: is defined as the percentage of a customer's transactions occurring in the top 3 most-frequented bins. The formula for calculating the loyalty is:

$$\boldsymbol{L_i = \frac{f_i}{\sum_{j=1}^{N} p_{ij}}} \tag{2}$$

  where f_i is the aggregate proportion of all transactions of customer i occurring in the top 3 bins, (Singh, Bozkaya, & Pentland, 2015). The output values range between 0 to 1, where large numbers signify high loyalty and low numbers signify low loyalty.

- Regularity: is an indicator of an individual's similarity in behavior over shorter (4 months) and longer (1 year) time periods. The formula for calculating the regularity is:

$$\boldsymbol{R_i = 1 - \frac{\sqrt{(D_i^1 - D_i^T)^2 + (L_i^1 - L_i^T)^2}}{\sqrt{2}}} \tag{3}$$

  where $\boldsymbol{D_i^1}$ and $\boldsymbol{D_i^T}$ are the diversity values for the individual in the first four months (July to October) and the entire year respectively. Similarly, $\boldsymbol{L_i^1}$ and $\boldsymbol{L_i^T}$ are the loyalty values of the individual for the same times periods, (Singh, Bozkaya, & Pentland, 2015). The output values range between 0 to 1, where large numbers signify high regularity and low numbers signify low regularity.

Bins: Each of the three features above was computed based on the same four different bins defined by Singh et al. (2015). These bins were *Spatial Radial, Spatial Grid, Temporal Hourly,* and *Temporal Weekly,* (Singh, Bozkaya, & Pentland, 2015).

Dependent variables signifying financial well-being: The following variables were considered as the dependent features:

- Overspending: this involves comparing an individuals' total credit card transactions for the year against their total income for the year. Hence if cc_i is a person's total credit card spending for the year, and I_i is the annual income, then overspending is defined as (Singh, Bozkaya, & Pentland, 2015):

  $$O_i = \frac{cc_i}{I_i} \tag{4}$$

  The output values range between 0 to below infinity, where low numbers signify less to none overspending, and high numbers signify major overspending. Individuals who had an overspending value less than or equal to 1 were assigned a 0, whereas those with values above 1 were assigned a 1.

- Trouble: A-Bank keeps track of their customer's payment history by assigning each customer one of six risk codes, or description of payment performance, to each of their 12 months. These risk codes in order of least severe to most severe are as follows: *Without Risk, Delayed 1-15 days, Delayed 16-30 days, Delayed 30-59 days*, *Delayed 60+ days*, and *Follow*. If a customer shows any of the last four risk codes (*Delayed 16-30 days, Delayed 31-59 days, Delayed 60+ days, Follow*) in any of the 12 months, then that customer is considered as being in "trouble" and is assigned a 1. If the customer shows only any of the other two risk codes (*Without Risk, Delayed 1-15 days*) in any of the 12 months, then they are considered as being "not in trouble" and are assigned a 0.

- Late Payment: For this feature, credit card statements only in local currency were considered. This variable signifies whether the customer paid late against their credit card statement. The combined number of total late days were considered for each customer and if any customer had total late days greater than 0, then they were assigned a 1 (late payer), otherwise 0 (pays on time). A grace period of 3 days was given to each individual to compensate for reasons of paying late not associated with financial trouble such as forgetting to pay, missing the deadline.

Classification Method: After the dataset cleaning, the final sample on which modeling was performed consisted of 16,291 customers. To prepare the dataset for model-building, one-hot encoding was performed on all categorical features. These included Gender, Marital Status, Education, and Job Status. The Bagging algorithm was used for classification. To divide the dataset into training and testing, a ratio of 70:30 was used. The models' results are based on 30-fold classification with re-sampling and replacement, and a unique random seed was used for each round. The complete modeling process was run in R.

SMOTE (synthetic minority over-sampling technique) was used to lessen the effects of imbalances in the dataset. The imbalances were such that while 94% of individuals were over-spenders only 6% were not, and while 97% of individuals were in trouble only 3% were not. Late Payment was relatively better distributed with the late-payer to not-late-payer ratio as 75:25.

3.2. New features for predicting financial well-being

In this part of our study, we introduce new features with the motivation to potentially improve the model by Singh et al. (2015). We provide the definitions of the new features, and also discuss the relevant preparations and manipulations performed on them.

For the same set of customers in Section 3.1, we obtain further information regarding the customers' shopping transactions, and banking channel usage. This new information includes:

- Customer's spending category or merchant type. (e.g. insurance, food, accommodation)
- Customer's channel usage while interacting with the bank. These included five categories: ATM, Branch Visit, Internet, Mobile Application, and Call Center.

Data cleaning/characteristics/validation: Purchase transactions in which the merchants' coordinates were not geo-coded were also excluded. The merchant codes had to be processed as well since there were a total of 1078 codes for each specific merchant. For example, Saudi Airlines, Qatar Airways, Accent Rent-a-Car, Dollar Rent-a-Car, Four Seasons Hotel, and Shangri-La Hotel all had a unique merchant code. All similar transactions were grouped into one category, which eventually narrowed down to just 22 categories such as Airlines, Car Hire, and Accommodation.

Features:

- Customer's spending category information was used to create two different types of features which were used separately. The purpose was to see which of the two representations had more influence on the model's predictive power.

i. **Category Diversity, Loyalty, and Regularity**: Based on the same entropy formulas defined in Section 3.1, further measures related to categories were defined.

- Category Diversity: when a customer's total shopping transactions are spread over different shopping categories. The dataset in total had 22 shopping categories such as insurance, gasoline, food, jewelers, and accommodation. Each shopping category is treated as a different bin which amounts to 22 different bins. The formula for calculating the entropy is the same as Equation (1), but where p_{ij} are the fraction of transactions that fall within shopping category bin j for customer i, and M is the number of non-empty shopping category bins over which the customer's shopping transactions are divided. The output value ranges between 0 to 1, where large numbers signify high category diversity and low number signify low category diversity.

- Category Loyalty: this is the percentage of a customer's transactions which occur in the top 3 most-frequently bought shopping categories. The formula for calculating the loyalty is the same as Equation (2), but where f_i is the aggregate proportion of all transactions of customer i which occur in the top 3 most frequently bought shopping categories. The output values range between 0 to 1, where large numbers signify high category loyalty and low numbers signify low category loyalty.

- Category Regularity: is an indicator of an individual's similarity in shopping purchases over shorter (4 months) and longer (1 year) time periods. The formula for calculating the regularity is the same as Equation (3), but where $\boldsymbol{D_i^1}$ and $\boldsymbol{D_i^T}$ are the category diversity values for the individual in the first four months (July to October) and the entire year respectively. Similarly, $\boldsymbol{L_i^1}$ and $\boldsymbol{L_i^T}$ are the category loyalty values of the individual for the same time periods. The output values range between 0 to 1, where large numbers signify high category regularity and low numbers signify low category regularity.

ii. **Shopping Categories used as a Binary Variable:** Each customer's top-most frequented shopping category was identified. Simultaneously, all 22 shopping categories were treated as 22 different dummy variables. A customer was assigned a 1 for the top-most frequented shopping category among the 22 and was assigned a 0 for all the remaining 21 categories. For example, if a customer spent the most on Gasoline, he/she was assigned a 1 for the dummy variable 'Gasoline', and a 0 for the dummy variables 'Insurance', 'Food', 'Clothing', and so on.

- Similarly, the customer's channel usage behavior was also modeled into two different types of features which were used separately. The purpose was to see which of the two representations had more influence on the model's predictive power.

i. **Channel Usage Diversity, Loyalty, and Regularity**: Based on the same entropy formulas defined in Section 3.1, further measures related to channel usage were defined.

- Channel Usage Diversity: when a customer's channel usage behavior is spread over different channels. The dataset in total had 5 channel types: *ATM, Call Center, Branch Visit, Internet,* and *Mobile Applications*. Each channel is treated as a different bin which accumulates to 5 different bins. The formula for calculating the entropy is the same as Equation (1), but where p_{ij} are the fraction of total contacts made with bank that fall within channel category bin *j* for customer *i*, and *M* is the number of non-empty channel category bins over which the customer's channel usage behavior is divided. The output value ranges between 0 to 1, where large numbers signify high channel usage diversity and low numbers signify low channel usage diversity.

- Channel Usage Loyalty: this is the percentage of a customer's total contacts made with bank which occur in the top 2 most-frequently used channel categories. The formula for calculating the entropy is the same as Equation (2), but where f_i is the aggregate proportion of all total contacts made with the bank of customer *i* through the top 2 most frequently used channels. The output values range between 0 to 1, where large numbers signify high channel usage loyalty and low numbers signify low channel usage loyalty.

- Channel Usage Regularity: is an indicator of an individual's similarity in channel usage over shorter (4 months) and longer (1 year) time periods. The formula for calculating the entropy is the same as Equation (3), but where $\boldsymbol{D_i^1}$ and $\boldsymbol{D_i^T}$ are the channel usage diversity values for the individual in the first four months (July to October) and the entire year respectively. Similarly, $\boldsymbol{L_i^1}$ and $\boldsymbol{L_i^T}$ are the channel usage loyalty values of the individual for the same time periods. The output values range between 0 to 1, where large numbers signify high channel usage regularity and low numbers signify low channel usage regularity.

ii. **Channel Categories used as a Binary Variable:** Each customer's top-most used channel was identified. Simultaneously, all 5 channel categories were treated as 5 different dummy variables. A customer was assigned a 1 for the top-most used channel category among the 5 and was assigned a 0 for all the other 4 channels. For example, if a customer used the call center the most to contact with the bank, he/she was assigned a 1 for the dummy variable 'Call Center', and a 0 for the dummy variables 'Branch', 'ATM, 'Mobile', and 'Internet'.

Bins: The category diversity, loyalty, and regularity entropy measures, along with the 'shopping categories used as a binary variable' values, all considered the 22 different shopping categories as 22 different bins.

Similarly, the channel diversity, loyalty, and regularity entropy values, along with the ‘channel categories used as a binary variable’ values, all considered the 5 different channel categories as 5 different bins.

Dependent variables signifying financial well-being: The same dependent features mentioned in Section 3.1 were predicted using the new features identified in this part of Chapter 3. These dependent features were *Overspending*, *Trouble*, and *Late Payment*.

Classification Method: The final sample on which modeling was performed consisted of 16,291 customers. In the case of using shopping categories and channel categories as binary variables, one-hot encoding was performed to prepare the dataset for model-building. Hence, 22 and 5 dummy variables were created separately in each case, respectively. The information related to algorithm used, dividing the dataset into train and test, classification rounds, and balancing the dataset are all the same as what were described in Section 3.1.

Chapter 4

RESULTS AND DISCUSSION

In this Chapter, we present the results of our experiments for predicting the financial well-being of A-Bank customers using the features described in the previous chapter. First, we present the findings related to validating the generalization of the model by Singh et al. (2015) on the A-Bank dataset, while in the second part, we present the findings related to using the new features (introduced Section 3.2) for potentially improving the aforementioned model.

4.1. Results and Discussion for the model by Singh et al. (2015) using A-Bank dataset

The analysis of the spatio-temporal features calculated for the 16,291 randomly selected customers based on their transactions illustrated that in most cases, spatio-temporal mobility of the customers significantly influenced their financial outcomes. The three main financial outcome variables were overspending, late payment, and financial trouble, (Singh, Bozkaya, & Pentland, 2015). We start by providing descriptive statistics and distributions on the features we calculated.

A) The following table shows the summary statistics of the A-Bank dataset (total of 16,291 customers) for all the demographic and spatio-temporal features, along with the three ‘financial outcomes’ dependent variables.

Gender	**Marital Status**
FEMALE: 4983	DIVORCED: 688
MALE: 11308	MARRIED: 12005
	SINGLE: 3074
	UNKNOWN: 455
	WIDOW: 69

Education	**Age**	**Job Status**
COLLEGE: 7195	Min.: 19.00	WAGE(PRIVATE): 11381
UNDERGRADUATE: 4860	1st Qu.: 32.00	SELF-EMPLOYED: 2177
HIGH SCHOOL: 1460	Median: 38.00	WAGE(PUBLIC): 1178
MIDDLE SCHOOL: 1156	Mean: 38.77	RETIRED: 765
PRIMARY SCHOOL: 922	3rd Qu.: 45.00	RETIRED EMPLOYEE(WAGE): 327
GRADUATE: 526	Max.: 83.00	HOUSEWIFE: 157
(Other): 172		(Other): 306

Spatial Radial Diversity	**Spatial Radial Loyalty**	**Spatial Radial Regularity**
Min.: 0.0000	Min.: 0.5570	Min.: 0.2657
1st Qu.: 0.6026	1st Qu.: 0.8444	1st Qu.: 0.8714
Median: 0.7403	Median: 0.9111	Median: 0.9221
Mean: 0.6944	Mean: 0.8971	Mean: 0.8983
3rd Qu.: 0.8322	3rd Qu.: 0.9659	3rd Qu.: 0.9571
Max.: 1.0000	Max.: 1.0000	Max.: 1.0000

Spatial Grid Diversity	**Spatial Grid Loyalty**	**Spatial Grid Regularity**
Min.: 0.0000	Min.: 0.3380	Min.: 0.2478
1st Qu.: 0.4747	1st Qu.: 0.7800	1st Qu.: 0.8482
Median: 0.6378	Median: 0.8841	Median: 0.9036
Mean: 0.5988	Mean: 0.8552	Mean: 0.8828
3rd Qu.: 0.7557	3rd Qu.: 0.9588	3rd Qu.: 0.9435
Max.: 1.0000	Max.: 1.0000	Max.: 1.0000

Temporal Weekly Diversity	**Temporal Weekly Loyalty**	**Temporal Weekly Regularity**
Min.: 0.2943	Min.: 0.4417	Min.: 0.2045
1st Qu.: 0.9239	1st Qu.: 0.5490	1st Qu.: 0.8784
Median: 0.9552	Median: 0.5946	Median: 0.9324
Mean: 0.9417	Mean: 0.6066	Mean: 0.9057
3rd Qu.: 0.9756	3rd Qu.: 0.6512	3rd Qu.: 0.9669
Max.: 1.0000	Max.: 1.0000	Max.: 1.0000

Temporal Hourly Diversity	Temporal Hourly Loyalty	Temporal Hourly Regularity
Min.: 0.1311	Min.: 0.2159	Min.: 0.1526
1st Qu.: 0.8805	1st Qu.: 0.3939	1st Qu.: 0.8737
Median: 0.9106	Median: 0.4513	Median: 0.9263
Mean: 0.8993	Mean: 0.4664	Mean: 0.8962
3rd Qu.: 0.9330	3rd Qu.: 0.5238	3rd Qu.: 0.9570
Max.: 0.9920	Max.: 1.0000	Max.: 1.0000

Overspending	Trouble	Late Payment
Min.: 0.00000	Min.: 0.00000	Min.: 0.0000
1st Qu.: 0.00000	1st Qu.: 0.00000	1st Qu.: 0.0000
Median: 0.00000	Median: 0.00000	Median: 1.0000
Mean: 0.05991	Mean: 0.02928	Mean: 0.7486
3rd Qu.: 0.00000	3rd Qu.: 0.00000	3rd Qu.: 1.0000
Max.: 1.00000	Max.: 1.00000	Max.: 1.0000

Table 1

Comparison with the Singh et al. (2015) dataset: Several demographic differences exist between the A-Bank customer dataset and the Singh et al. dataset. The male-to-female ratios in the A-Bank and Singh et al. dataset are 69:31 and 73:27 respectively. The A-Bank customers appear to be more married (74%) and less single (19%) as compared to the Singh et al. dataset (66% and 29% respectively). In terms of education, around 77% of the A-Bank customers either have a college, undergraduate, or master's degree, whereas only 40.5% of the Singh et al. customers have a college, masters, or Ph.D. degree. Hence, the A-Bank customers appear to be more educated than the customers in the Singh et al. dataset. Age-wise both datasets are quite similar. In terms of job status, more A-Bank customers are self-employed (13%) and less are in public sector (7%) as compared to the Singh et al. dataset (6% and 18% respectively). Almost the same proportion of customers are in the private sector in the A-Bank dataset (70%) and Singh et al. dataset (68%). However, more of the customers are either retired or unemployed in the A-Bank dataset (9%) as compared to the Singh et al. dataset (4%). These can be cited as few of the reasons why the A-Bank customers are more regular than diverse, as compared to the

Singh et al. customers, who are more diverse than regular. This is also proven by the mean regularity values for all of the four spatio-temporal bins in the A-Bank dataset, which are higher compared to the Singh et al. dataset.

B) The following graphs show the cumulative density functions for diversity, loyalty, and regularity of the customers in each of the four different bins: spatial grid, spatial radial, temporal weekly, and temporal hourly.

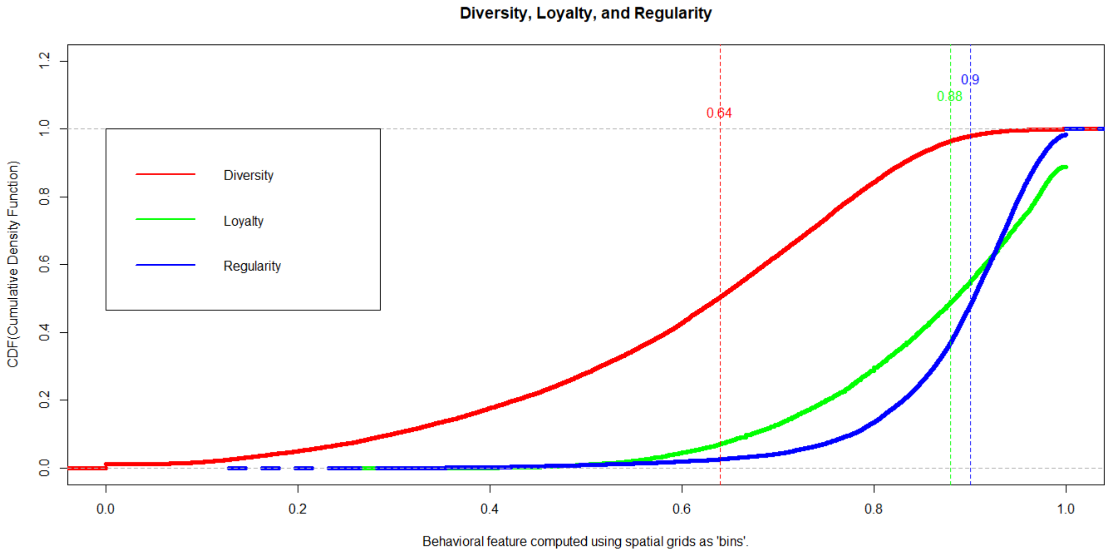

Figure 1A: The cdf graph shows the median scores for diversity, loyalty, and regularity in the 'spatial grids' bin. All three curves have high median scores, which indicates their strong tendency to occur in human shopping behavior. In general, consumers are more regular and loyal in terms of the 'grids' they shop in, as compared to how diverse they are.

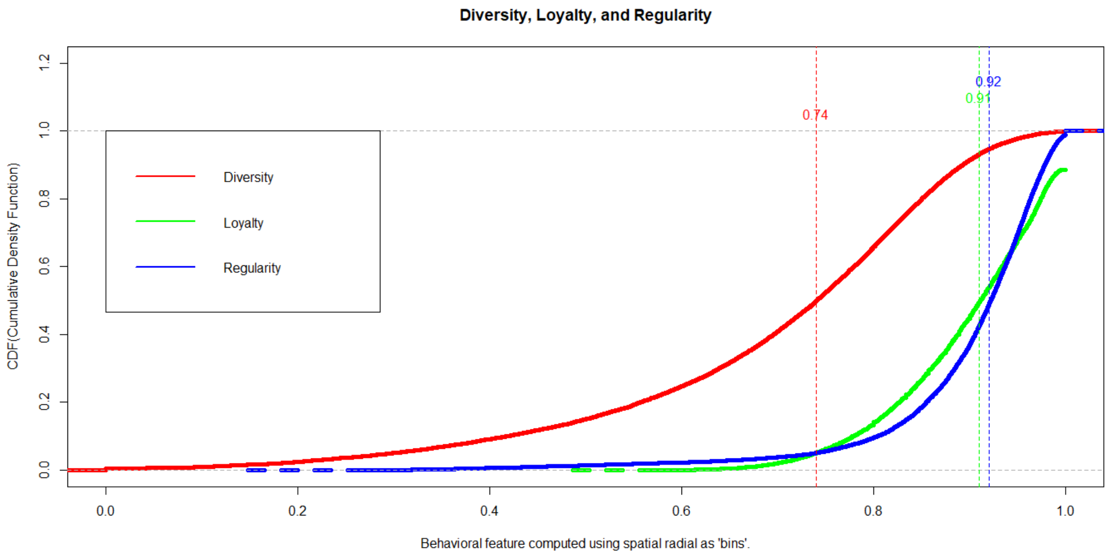

Figure 1B: The cdf graph shows the median scores for diversity, loyalty, and regularity in the 'spatial radials' bin. All three curves have high median scores, which indicates their strong tendency to occur in human shopping behavior. In general, consumers are more regular and loyal in terms of the 'radials' within which they shop, as compared to how diverse they are.

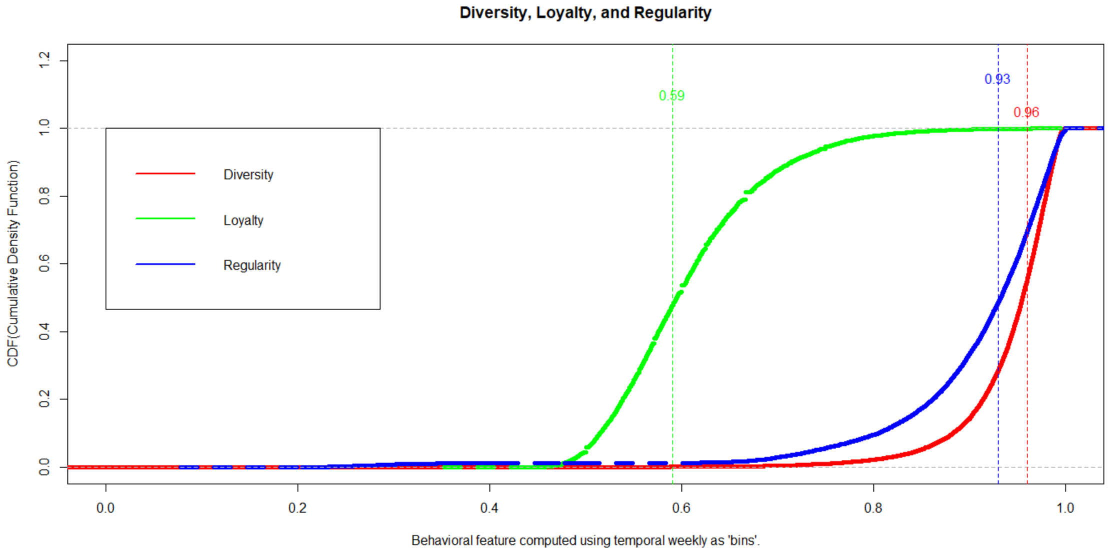

Figure 1C: The cdf graph shows the median scores for diversity, loyalty, and regularity in the 'temporal weekly' bin. Diversity and Regularity have high median scores compared to Loyalty. This indicates their strong tendency to occur in human shopping behavior. In general, consumers are more diverse and regular in terms of the 'day of the week' they shop at. The relatively low median score of 0.59 for loyalty shows that people are not as loyal in terms of the day they shop at, as compared to how diverse and regular they are.

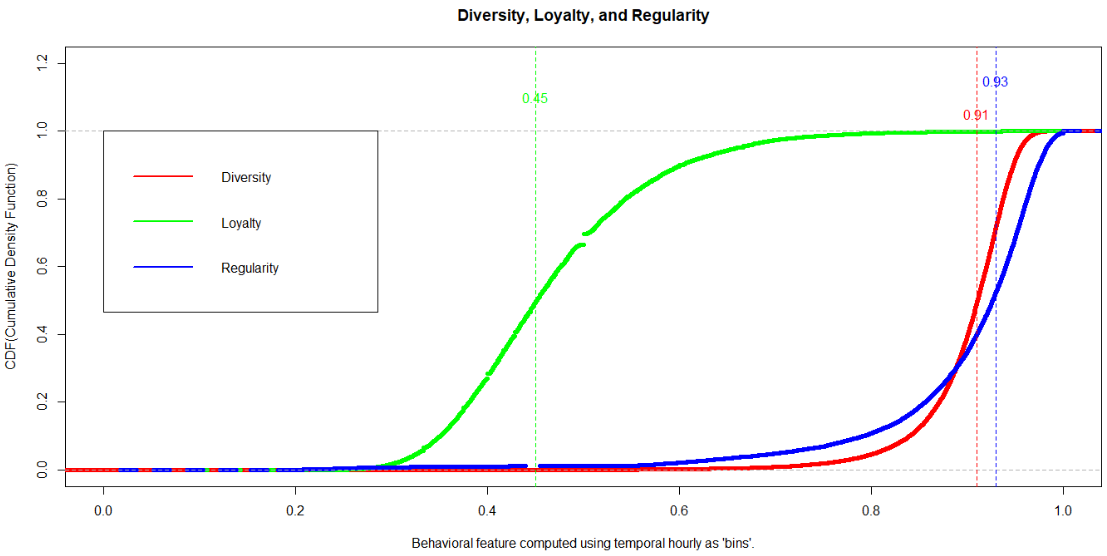

Figure 1D: The cdf graph shows the median scores for diversity, loyalty, and regularity in the 'temporal hourly' bin. Regularity and Diversity have high median scores compared to Loyalty. This indicates their strong tendency to occur in human shopping behavior. In general, consumers are more regular and diverse in terms of the 'hour of the day' they shop at. The relatively low median score of 0.45 for loyalty shows that people are not as loyal in terms of the hour of the day they shop at, as compared to how diverse and regular they are.

Comparison with the Singh et al. (2015) dataset: In the Singh et al. dataset, the customers exhibited higher loyalty (0.9) and diversity (0.79) in the spatial grids bin, as compared to the A-Bank customers which had higher regularity (0.9). In the spatial radial bins, both the datasets exhibited almost the same level of loyalty and regularity, but the A-Bank customers were less diverse (0.74) as compared to the Singh et al. dataset. In the temporal weekly bins, the level of diversity, loyalty and regularity were very similar. In the temporal hourly bins, the A-Bank customers were slightly more regular than diverse compared to the Singh et al. dataset, while the loyalty level was low for both.

C) The following graphs show the combined entropy for each of the three features (diversity, loyalty, and regularity) in each of the four bins (spatial grid, spatial radial, temporal hourly, temporal weekly):

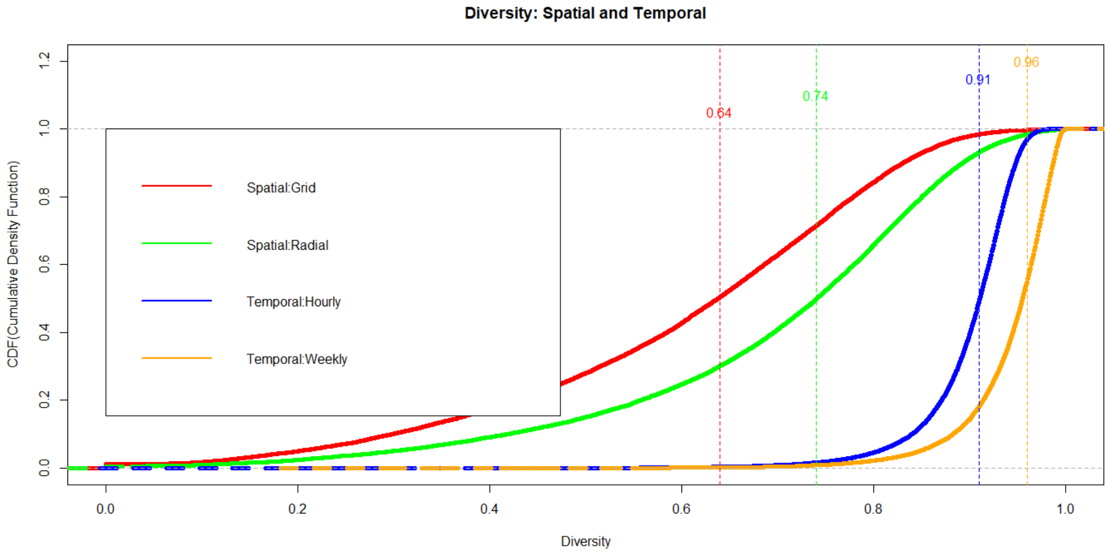

Figure 2A: The cdf graph shows the median scores for diversity in each of the four different bins: spatial grid, spatial radial, temporal hourly, and temporal weekly. It shows that the customers were more diverse in terms of the day of the week and hour of the day they shopped at, as compared to the locations they visited and the distances they travelled.

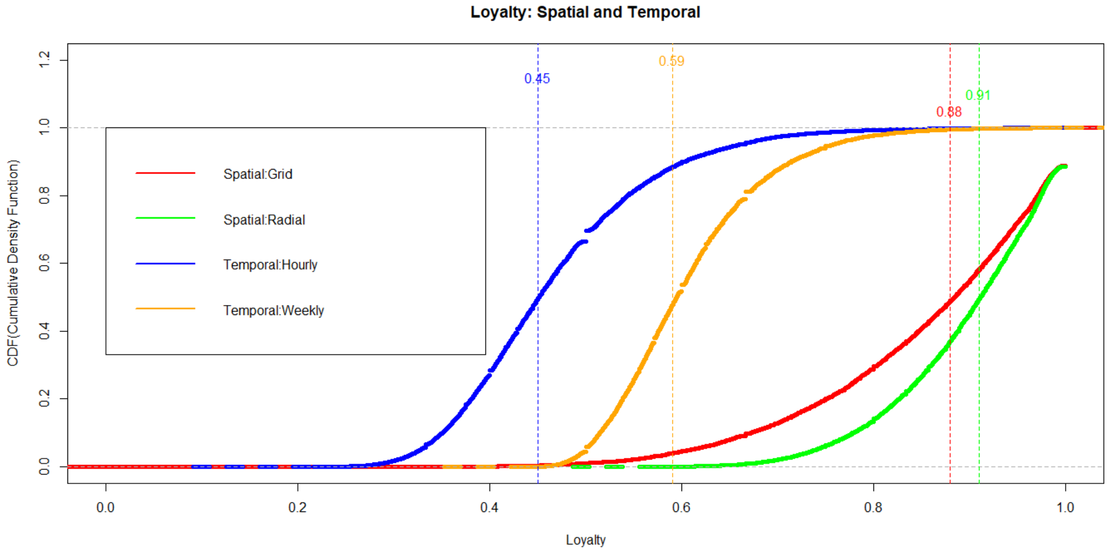

Figure 2B: The cdf graph shows the median scores for loyalty in each of the four different bins: spatial grid, spatial radial, temporal hourly, and temporal weekly. It shows that the customers were more loyal in terms of the distances they travelled and the locations they visited, as compared to the day of the week and hour of the day they shopped at. In fact, customers had more of a tendency to not shop at the same time of the day every time, and less than 60% of their purchases were made during their favored day of the week and hour of the day. Overall, the customer's three most frequented locations accounted for a very large percentage (0.88) of all their shopping.

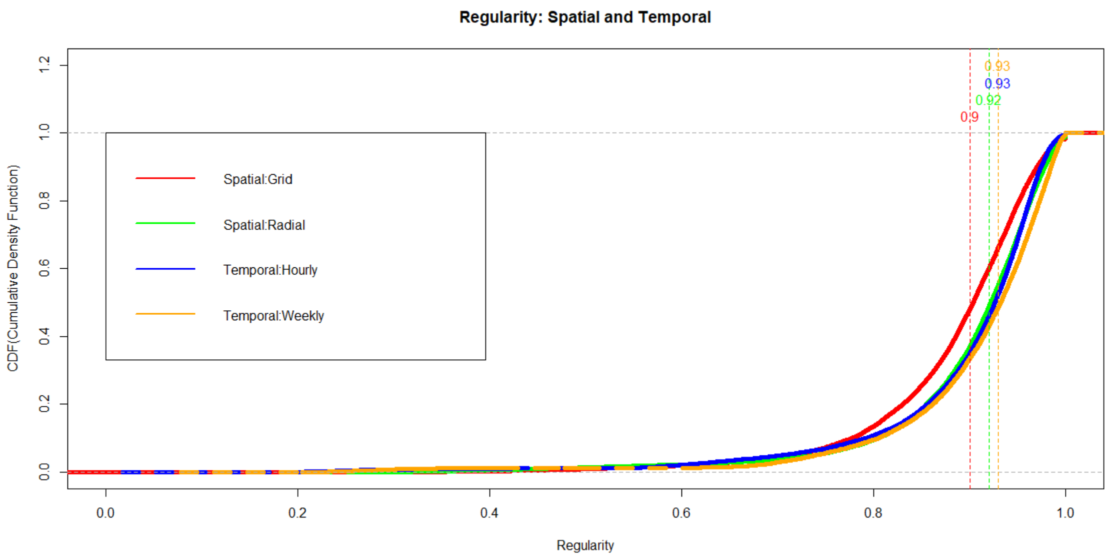

Figure 2C: The cdf graph shows the median scores for regularity in each of the four different bins: spatial grid, spatial radial, temporal hourly, and temporal weekly. It shows that customers were regular in all the bins, i.e. they were regular in terms of the day of the week and hour of the day they shopped at, and regular with regards to the distances they travelled and they locations they visited while shopping. Overall, they exhibited very similar behavioral patterns over time.

Comparison with Singh et al. (2015) dataset: The individuals in the Singh et al. dataset and A-Bank dataset both exhibited high diversity (greater than 0.90) in both the temporal hourly and temporal weekly bins. However, the A-Bank dataset individuals showed less diversity in the spatial grid and spatial radial bins, compared to the Singh et al. dataset.

Regarding loyalty in spatial and temporal bins, both the Singh et al. and A-Bank datasets exhibited almost the same level of loyalty in all bins. The A-Bank individuals were slightly less loyal (0.59) in terms of the day of they shopped at, as compared to the Singh et al. dataset (0.64).

Regarding regularity in spatial and temporal bins, both the Singh et al. and A-Bank dataset illustrated high levels of regularity in the temporal weekly, temporal hourly, and spatial radial bins. However, with regards to the spatial grid bin, the A-Bank dataset exhibited higher regularity (0.90) as compared to the Singh et al. dataset (0.75). This means that the A-Bank dataset individuals were more regular in terms of the locations they visited while shopping.

D) Regression analysis of the dependent variables (overspending, trouble, late payment) was conducted, which revealed that several demographic and spatio-temporal variables had a significant association with them. The basis of this significance was the *p*-value of the coefficients produced during the performance of generalized linear modeling for logistic regression. The following graphs show the odds ratio for each of the three financial outcome variables, i.e. overspending, trouble, late payment, in terms of each of the four spatio-temporal bins and their entropies.

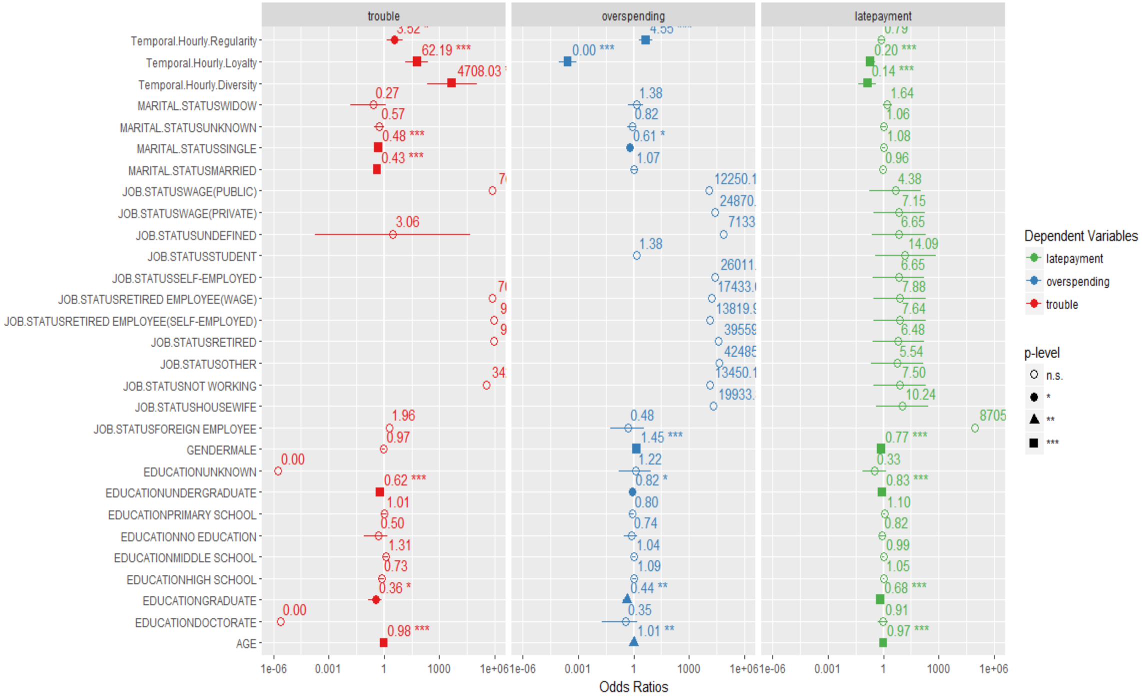

Figure 3A: The figure shows significant associations observed during logistic regression performed between the demographic and spatio-temporal features and the financial outcomes. Customers who were more likely to be in trouble were those who were either more regular, loyal, or diverse in terms of the time of the day they shopped at. Single and married customers were less likely to in trouble, just like those who had an undergraduate or graduate degree. Older customers were also less likely to be in trouble, as compared to younger customers. Customers who were more likely to overspend were those were more regular in terms of the hour of the day they shopped at, whereas those who were more loyal were less likely to overspend. Single individuals, males, and those with a graduate degree were also likely to overspend, whereas older customers were marginally less likely to overspend. Customers who were less likely to pay their bills late were those who were more loyal and diverse in terms of the time of the day they shopped at, were male, had either an undergraduate or master's degree, and were of an older age.

Comparison of Figure 3A with the Singh et al. (2015) dataset: In both datasets, customers who were more likely to be in trouble were those who were more loyal or diverse in terms of the hour of the day they shopped at. Older age customers were also less likely to be in trouble in both datasets. However, while more regular customers were more likely to be in trouble in the A-Bank dataset, in the Singh et al. dataset these customers had no significance.

In both datasets, regular and male customers were more likely to overspend. Also, loyal customers were less likely to overspend. In both datasets, older age and male customers were the ones less likely to pay their bills late. However, while more loyal and diverse A-Bank customers were less likely to miss their bill payments, in the Singh et al. dataset these same customers were more likely to miss them.

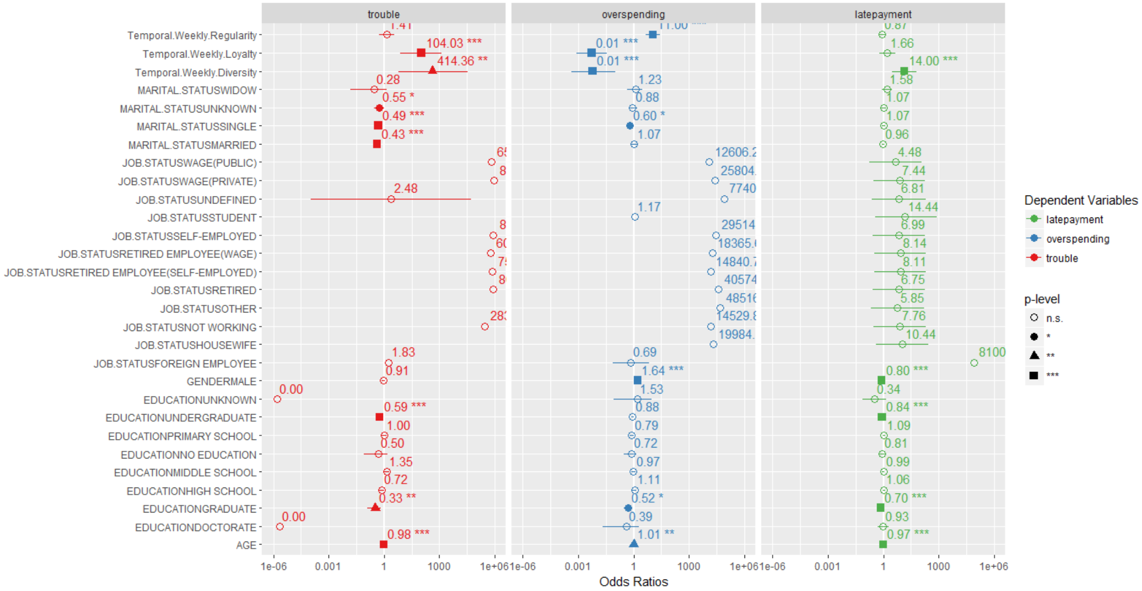

Figure 3B: The figure shows significant associations observed during logistic regression performed between the demographic and spatio-temporal features and the financial outcomes. Customers who were more likely to be in trouble were those who were either more loyal or diverse in terms of the day of the week they shopped at. Customers who were less likely to be in trouble were those whose marital status was single, married, or unknown. Older customers, and those with an undergraduate or graduate, were also less likely to be in trouble. Male customers and those who were regular in terms of the day of the week they shopped at were more likely to overspend, whereas those who were more loyal and diverse were less likely to overspend. Single customers and those with a graduate degree were less likely to overspend, and older customers were marginally less likely to overspend. Customers who were more likely to miss their payments were those who were more diverse in terms of the day of the week they shopped at, while older and male customers, along with those who had an undergraduate or graduate degree, were less likely to miss their payments.

Comparison of Figure 3B with the Singh et al. (2015) dataset: In both datasets, customers who were more likely to be in trouble were those who were more loyal or diverse in terms of the day of the week they shopped at. Also, older age customers were less likely to be in trouble.

Male and regular customers in both datasets were more likely to overspend. On the other hand, customers who were less likely to overspend in both datasets were those who were more loyal, diverse, or were of an older age.

Customers who were less likely to miss their bills in both datasets were those who were either male or were of an older age. However, while more diverse A-Bank customers were more likely to miss their payments, these same class of customers had no significance in the Singh et al. dataset.

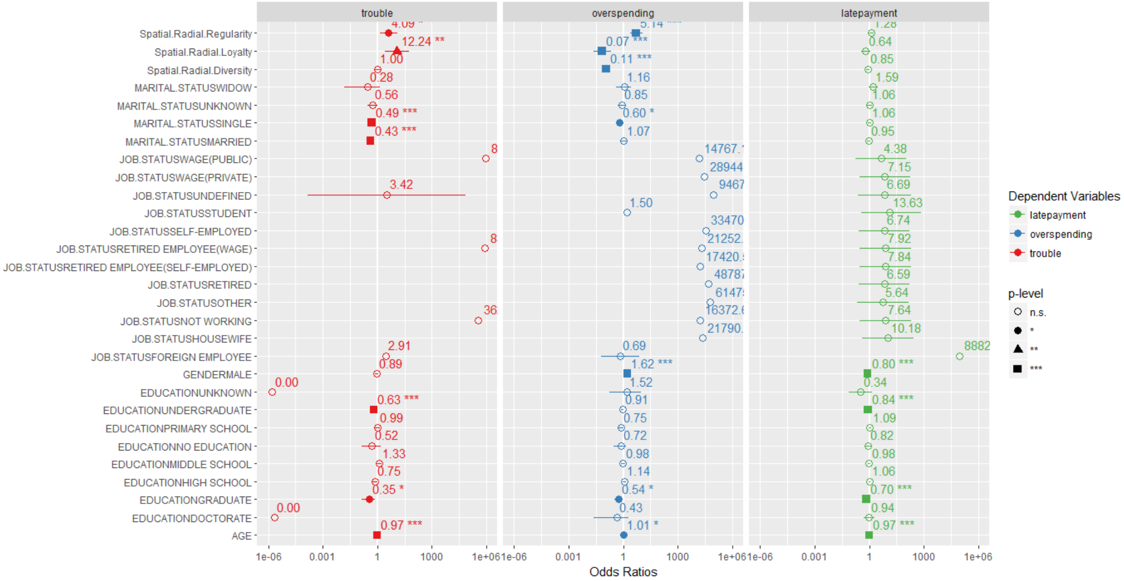

Figure 3C: The figure shows significant associations observed during logistic regression performed between the demographic and spatio-temporal features and the financial outcomes. Customers who were more likely to be in trouble were those who were more regular or loyal in terms of the distances they travelled while shopping. Customers who were less likely to be in trouble were those who were single, married, had an undergraduate or graduate degree, and were of an older age. Male customers, and those who were more regular in terms of the distances they travelled while shopping, were more likely to overspend. Customers who were less likely to overspend were those who were more loyal and diverse in terms of the distances they travelled while shopping, were single, or had a graduate degree. Older customers were marginally less likely to overspend. Customers who were less likely to miss their payments were those who were male, were of an older age, or had an undergraduate or graduate degree.

Comparison of Figure 3C with the Singh et al. (2015) dataset: In both datasets, customers who were more likely to be in trouble were those who were more loyal in terms of the distances they travelled while shopping. Older age customers were less likely to be in trouble in both datasets. However, while more regular customers were more likely to be in trouble in the A-Bank dataset, in the Singh et al. dataset these kinds of customers were less likely to be in trouble.

Customers who were more likely to overspend in both datasets were those who were either male or were more regular. Also, in both datasets, loyal and diverse customers were less likely to be in trouble, while older age customers were marginally less likely to be in trouble. Similarly, older age and male customers were also less likely to miss their payments. However, while more loyal customers were more likely to miss their payments in the Singh et al. dataset, these same customers had no significance in the A-Bank dataset.

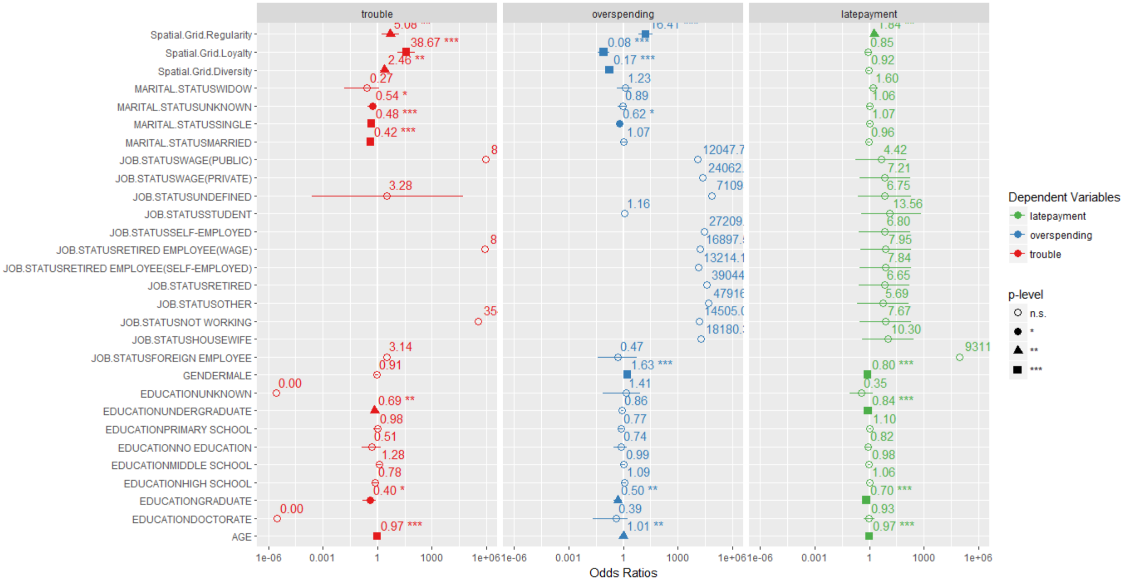

Figure 3D: The figure shows significant associations observed during logistic regression performed between the demographic and spatio-temporal features and the financial outcomes. Customers who were more likely to be in trouble were those were more regular, loyal, and diverse in terms of the locations they shopped at. Customers with their marital status as single, married, or unknown, who had either an undergraduate degree or graduate degree, and those who were older in age were less likely to be in trouble. Male customers and those who were more regular in terms of the locations they shopped at were more likely to overspend. Customers who were less likely to overspend were those who were more diverse and loyal in terms of the locations they shopped at, were single, or had a graduate degree. Older age customers were also less likely to overspend. Customers who were more likely to be late for their payments were those who were more regular in terms of the locations they shopped at. Those who were less likely to miss their bill payments were either male, had an undergraduate or graduate degree, or were of an older age.

Comparison of Figure 3D with the Singh et al. (2015) dataset: In both datasets, customers who were more likely to be in trouble were those who were more loyal or diverse in terms of the locations they shopped at. Older age customers were less likely to be in trouble in both the datasets. However, while regular customers were more likely to be in trouble in the A-Bank dataset, these same customers were less likely to be in the trouble in the Singh et al. dataset.

Customers who were more likely to overspend in both datasets were those who were either regular in their shopping locations or were male. Loyal and diverse customers were less likely to overspend in both datasets.

Older age and male customers were less likely to miss their payments in both datasets. However, while more regular A-Bank customers were more likely to miss their payments, the same class of Singh et al. customers were less likely to miss their payments.

E) To predict each of the three financial outcome variables, the bagging algorithm was used. The following graphs show the ‘area under the curve / receiver operating characteristic curve’ (AUC) values for each of the three dependent variables: overspending, trouble, and late payment. The AUC values are based on 30 different rounds (30 different random seeds) of classification. Each graph compares the AUC values produced when models based on only baseline, demographic, or spatio-temporal features are used to predict each of the three dependent variables.

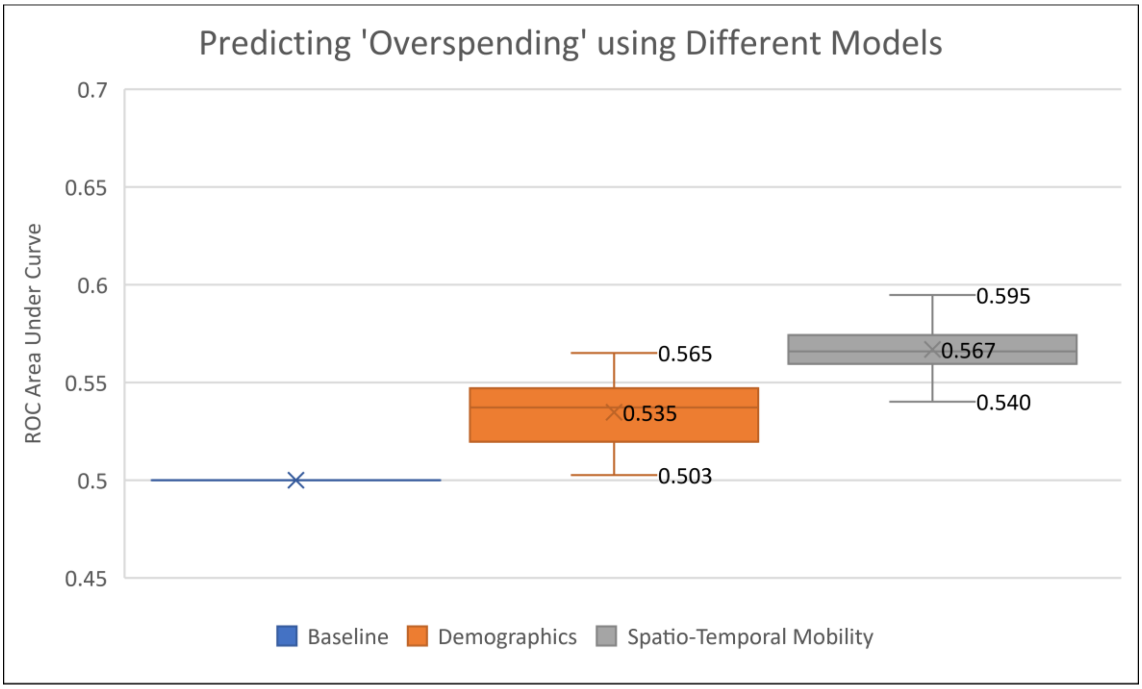

Figure 4A: The box plots above show the prediction performance for overspending using a baseline, demography-based, and spatio-temporal mobility model. The spatio-temporal mobility model performs 6% better than the demography model for predicting overspending.

	Overspending	
	Demography	Spatio-Temporal Mobility
Minimum	0.503	0.540
Mean	0.535	0.567
Median	0.537	0.566
Maximum	0.565	0.595

Table 2A: The above table summarizes basic summary statistics for the AUCs of the models built using demography and spatio-temporal mobility features for predicting overspending.

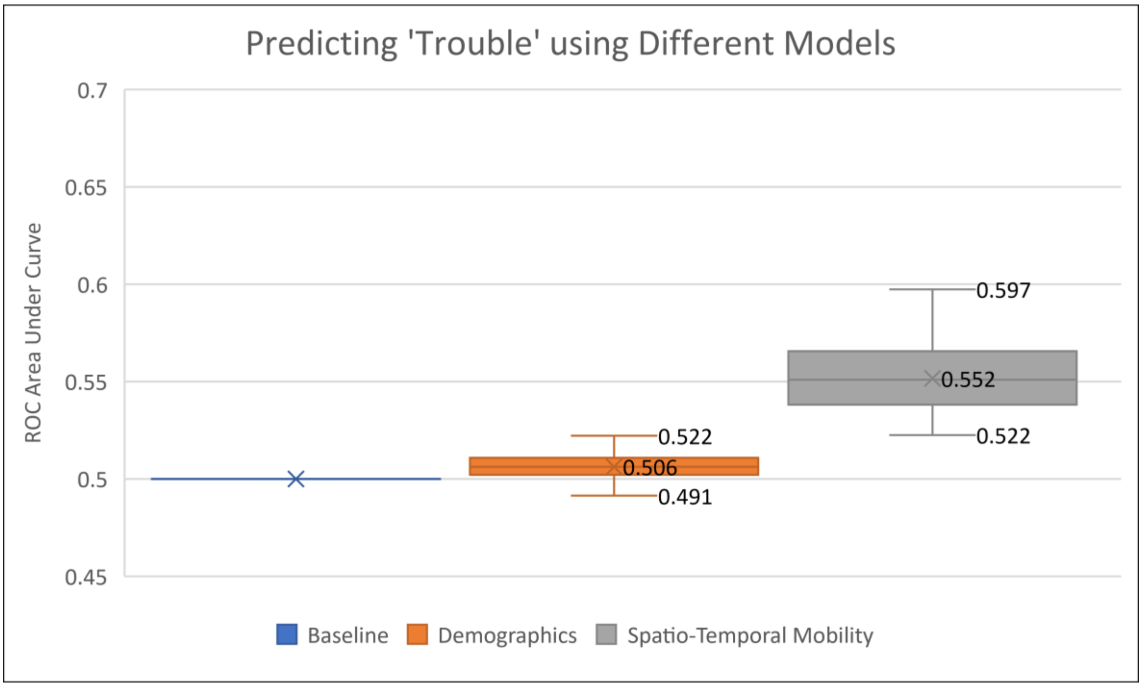

Figure 4B: The box plots above show the prediction performance for trouble using a baseline, demography-based, and spatio-temporal mobility model. The spatio-temporal mobility model performs 9% better than the demography model for predicting trouble.

	Trouble	
	Demography	**Spatio-Temporal Mobility**
Minimum	0.491	0.522
Mean	0.506	0.552
Median	0.506	0.551
Maximum	0.522	0.597

Table 2B: The above table summarizes basic summary statistics for the AUCs of the models built using demography and spatio-temporal mobility features for predicting trouble.

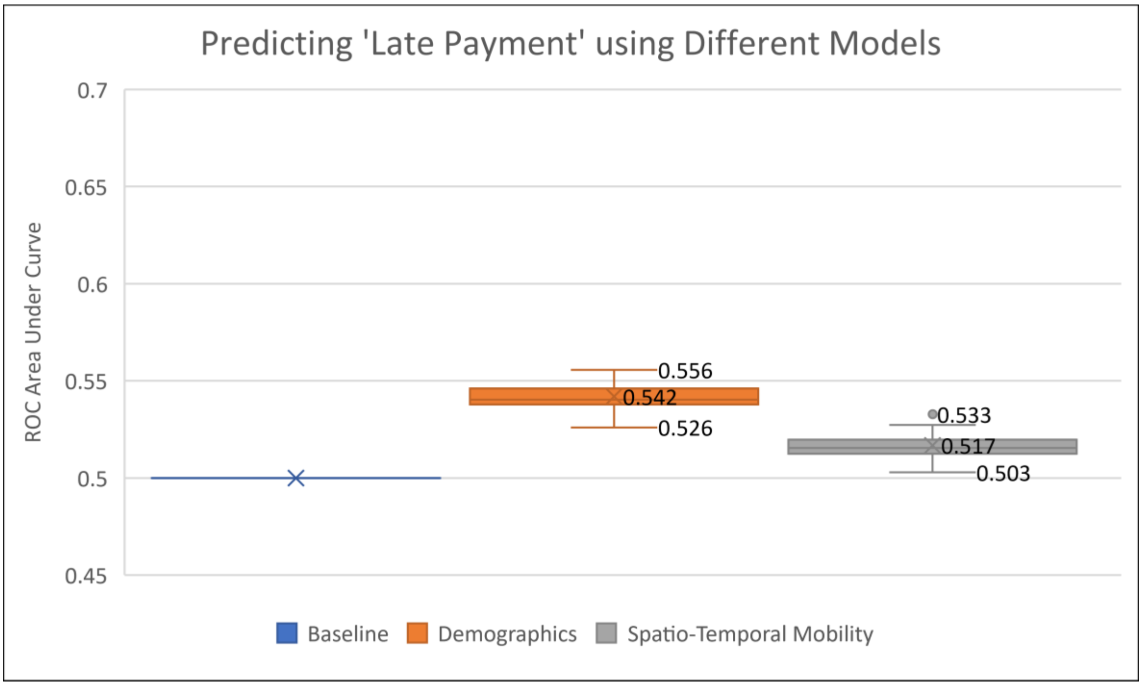

Figure 4C: The box plots above show the prediction performance for late payment using a baseline, demography-based, and spatio-temporal mobility model. The demography model performs 5% better than the spatio-temporal mobility model for predicting late payment.

	Late Payment	
	Demography	**Spatio-Temporal Mobility**
Minimum	0.526	0.503
Mean	0.542	0.517
Median	0.540	0.515
Maximum	0.556	0.533

Table 2C: The above table summarizes basic summary statistics for the AUCs of the models built using demography and spatio-temporal mobility features for predicting late payment.

Since late payment proved to not have a strong association with behavioral variables when using a dataset spanning one year, a further analysis was conducted whereby the year was broken down into four groups of three months. This was done to investigate whether any sort of relationship between late payment and behavioral variables could be found if we used the same time span as in the Singh et al. (2015) paper. However, no such significant association could be found. Below are the AUC findings for each quarter of the year from July 2014 to June 2015:

Features	**July-Aug-Sept**	**Oct-Nov-Dec**	**Jan-Feb-Mar**	**April-May-June**
Demographics	0.508	0.503	0.504	0.514
Spatio-Temporal Mobility	0.505	0.506	0.500	0.499
All	0.516	0.515	0.514	0.516

Table 2D: Table showing the AUC values for predicting Late Payment using both demographic and spatio-temporal features, for each quarter separately.

Overall, the spatio-temporal features allowed us to predict ‘overspending’ and ‘trouble’ better than the demographic features by nearly 6% and 9%. Although significant, these results are weaker compared to the findings of Singh et al. (2015). The authors were able to better predict the same financial outcomes by 49% and 31% respectively. They were also able to prove that ‘late payment’ can be predicted 30% better using spatio-temporal features as compared to demographics. However, several timeline-related and demographic differences exist between the dataset used in this research and the one used by Singh et al. (2015). Where the Singh et al. (2015) dataset spanned just three months, the A-Bank dataset spans 12 months. Furthermore, there is a greater percentage of females in the A-Bank dataset as compared to the Singh et al. (2015) dataset. Also, the A-Bank dataset customers are more married, less single, more educated, more self-employed, less publicly-employed, and more retired or unemployed than the Singh et al. (2015) dataset. All these differences may be a cause for the high regularity and less diversity seen in the A-Bank dataset (Figure 1A), and which might eventually have influenced the strength of the relationship between their behavior and their financial outcomes.

4.2. Results and Discussion with the proposed new behavioral features

The analysis of the features related to shopping categories and channel usage calculated for the randomly selected customers (16,291 and 15,388 respectively) based on their transactions illustrate that shopping categories significantly influenced ‘overspending’ of the customers. The predictability of the other two financial outcomes (trouble and late payment) is not found to be significantly affected by any of the new features.

A) The following tables shows the summary statistics of the A-Bank dataset for the new features related to shopping categories and banking channels, along with the three 'financial outcomes' dependent variables.

Table 3A: Summary statistics related to using shopping categories as an entropy variable.

Category Diversity	**Category Loyalty**	**Category Regularity**
Min: 0.0000	Min.: 0.3421	Min.: 0.2233
1st Qu.: 0.5761	1st Qu.:0.7238	1st Qu.: 0.8585
Median: 0.7027	Median: 0.8176	Median: 0.9099
Mean: 0.6642	Mean: 0.8075	Mean: 0.8880
3rd Qu.: 0.7902	3rd Qu.:0.9038	3rd Qu.: 0.9451
Max.: 0.9992	Max.: 1.0000	Max.: 1.0000

Table 3B: Summary statistics (frequencies) related to using shopping categories as a binary variable.

Top Shopping Categories among Customers
Market and Shopping Centers: 5731
Clothing and Accessories: 4934
Gasoline and Fuel Stations: 3092
Electric/Electronic Goods/Computers: 621
Insurance: 414
Various Food: 323
(Other): 1176

Table 3C: Summary statistics related to using channel categories as an entropy variable.

Channel Diversity	**Channel Loyalty**	**Channel Regularity**
Min.: 0.0000	Min.: 0.5000	Min.: 0.2468
1st Qu.: 0.4753	1st Qu.: 0.8605	1st Qu.: 0.8263
Median: 0.6323	Median: 0.9344	Median: 0.9070
Mean: 0.6031	Mean: 0.9092	Mean: 0.8607
3rd Qu.: 0.7725	3rd Qu.: 0.9851	3rd Qu.: 0.9568
Max.: 1.0000	Max.: 1.0000	Max.: 1.0000

Table 3D: Summary statistics (frequencies) related to using channel categories as a binary variable.

Top Channel Categories among Customers
Atm: 4251
Bank Branch: 2454
Call Center: 1719
Internet: 2739
Mobile: 4926

B) The following graphs show the cumulative density functions of the shopping category and banking channel entropies (diversity, loyalty, and regularity) shown by the customers in each of the respective bins for shopping and channels.

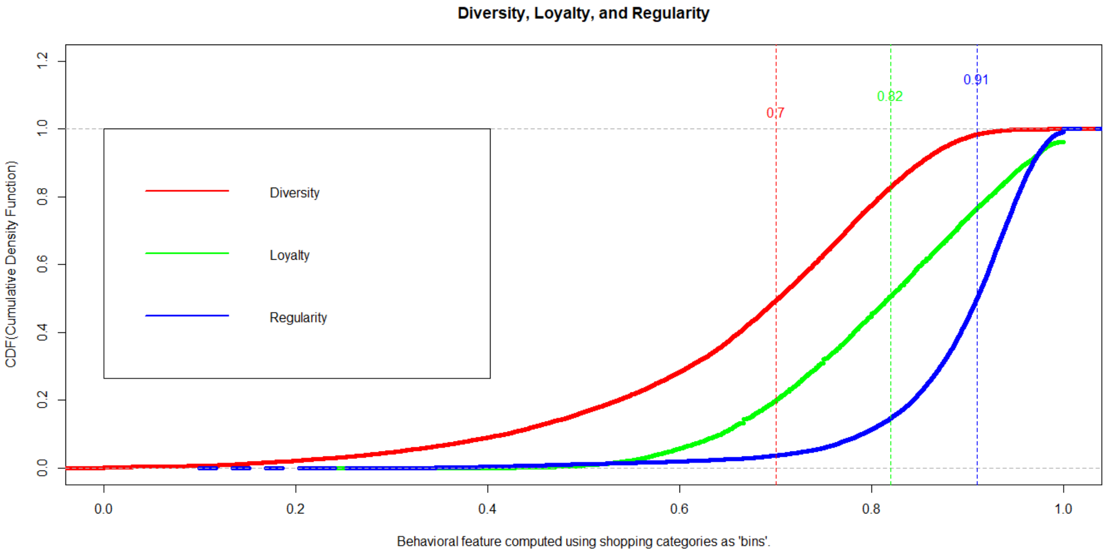

Figure 5A: The cdf graph shows the median scores for diversity, loyalty, and regularity when using shopping categories as bins. All three curves have high median scores, indicating their strong affinity with human shopping behavior. In general, consumers are more regular and loyal in terms of the 'shopping categories' they buy, as compared to how diverse they are.

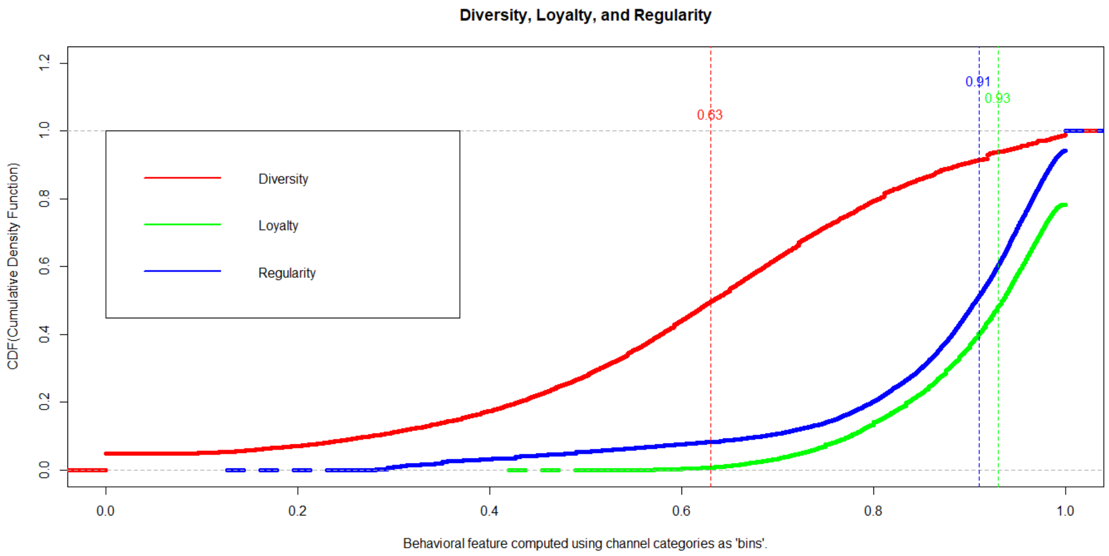

Figure 5B: The cdf graph shows the median scores for diversity, loyalty, and regularity when using channel categories as bins. All three curves have high median scores, indicating their strong affinity with customer's channel usage behavior. In general, consumers are more loyal and regular in terms of the channels they use to communicate with the bank, as compared to how diverse they are.

C) The following graphs show the combined entropy for each of the three features (diversity, loyalty, and regularity) related to the 'shopping category' bins and 'banking channel' bins, along with for the 'spatial radial', 'spatial grid', 'temporal hourly', and 'temporal weekly' bins.

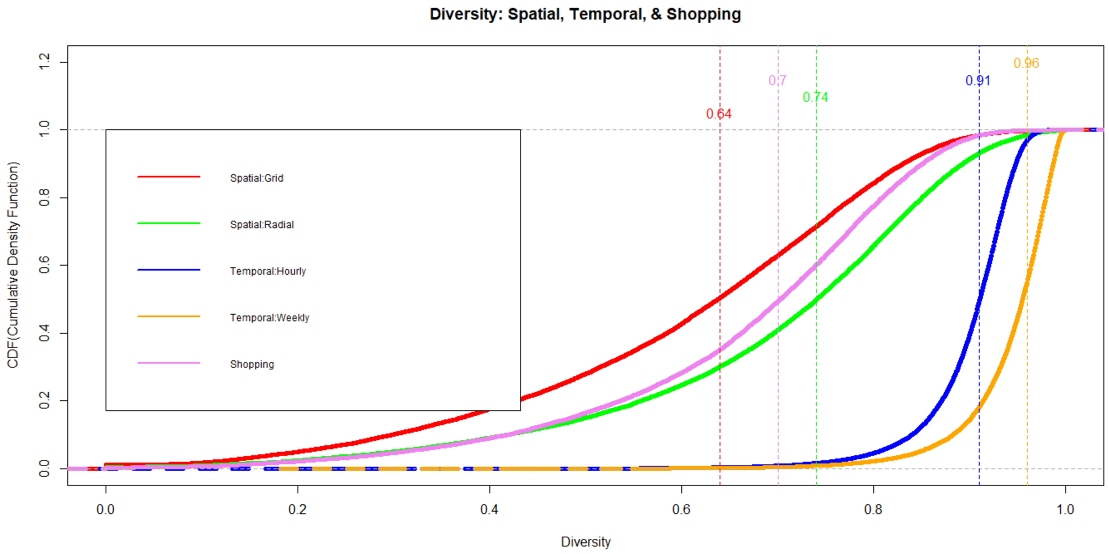

Figure 6A: The cdf graph shows the median scores for diversity in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and shopping categories. It shows that the customers had a high level of diversity (0.70) in the shopping categories which they preferred. Customers were more diverse in terms of the shopping categories they preferred, as compared to the locations which they visited.

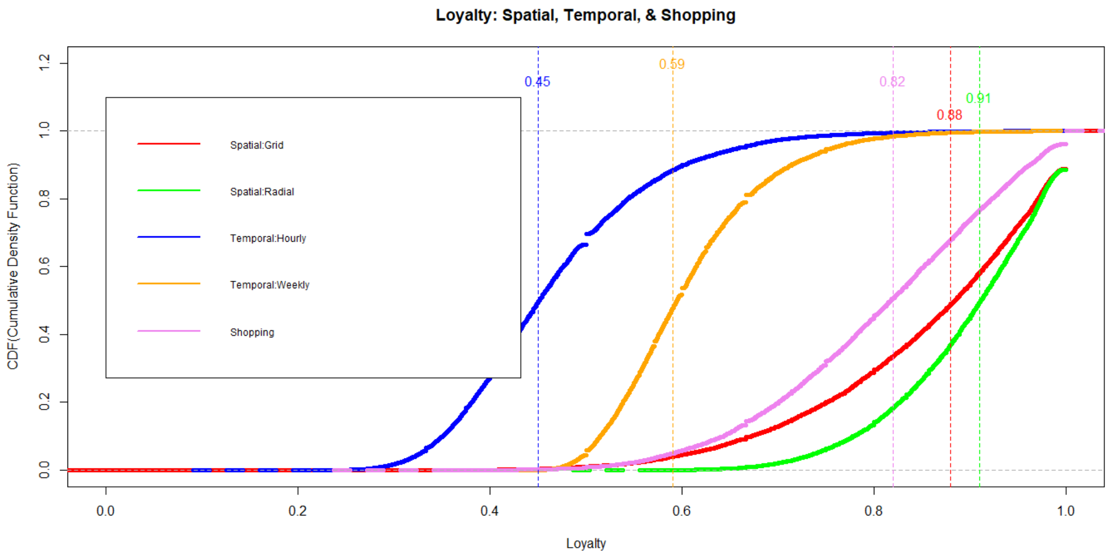

Figure 6B: The cdf graph shows the median scores for loyalty in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and shopping categories. It shows that the customers had a high level of loyalty (0.82) in the shopping categories which they preferred. Customers were more loyal in terms of the shopping categories they preferred, as compared to the day of the week they made their purchases and hour of the day at which they bought them. However, customers were more loyal with regards to the locations they visited and the distances they travelled to make their purchases.

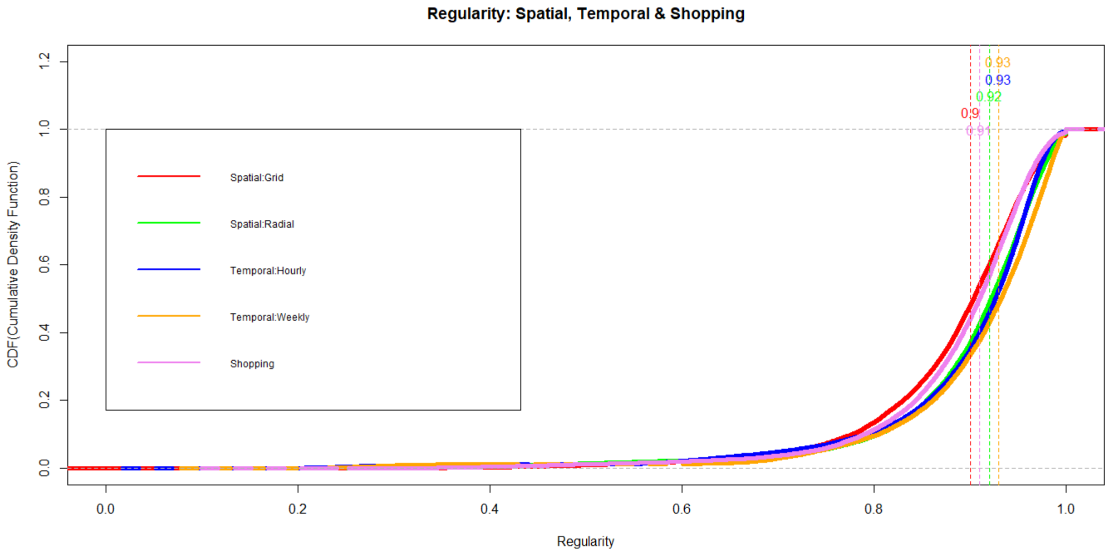

Figure 6C: The cdf graph shows the median scores for regularity in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and shopping categories. It shows that customers were regular in all the bins, including the shopping categories which they preferred to purchase (0.91). Overall, the customers exhibited very similar behavioral patterns over time.

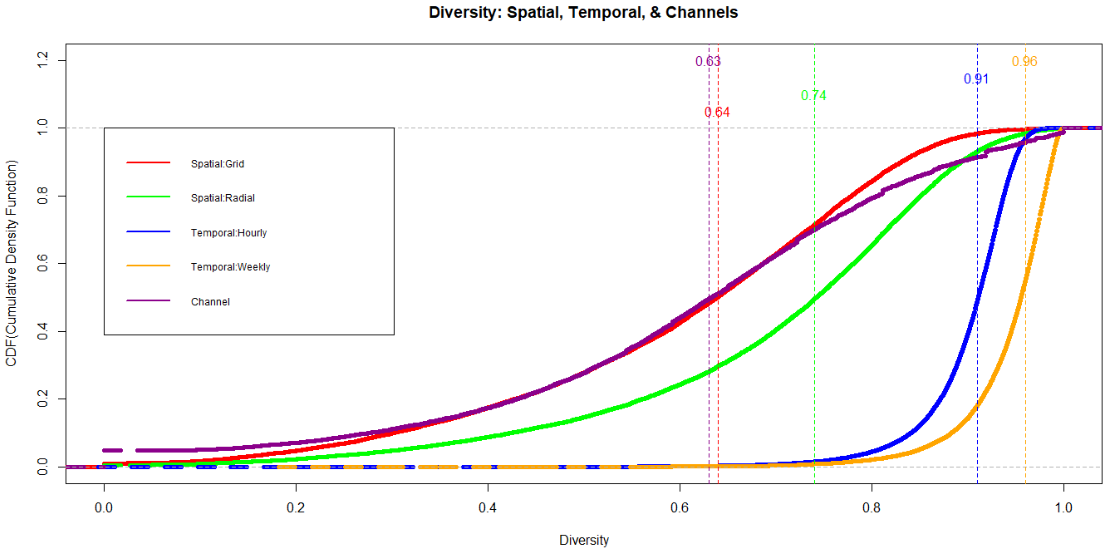

Figure 6D: The cdf graph shows the median scores for diversity in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and channel categories. It shows that the customers were more diverse (0.63) than not in using the different channels when interacting with the bank. However, it had the lowest level of diversity amongst all bins.

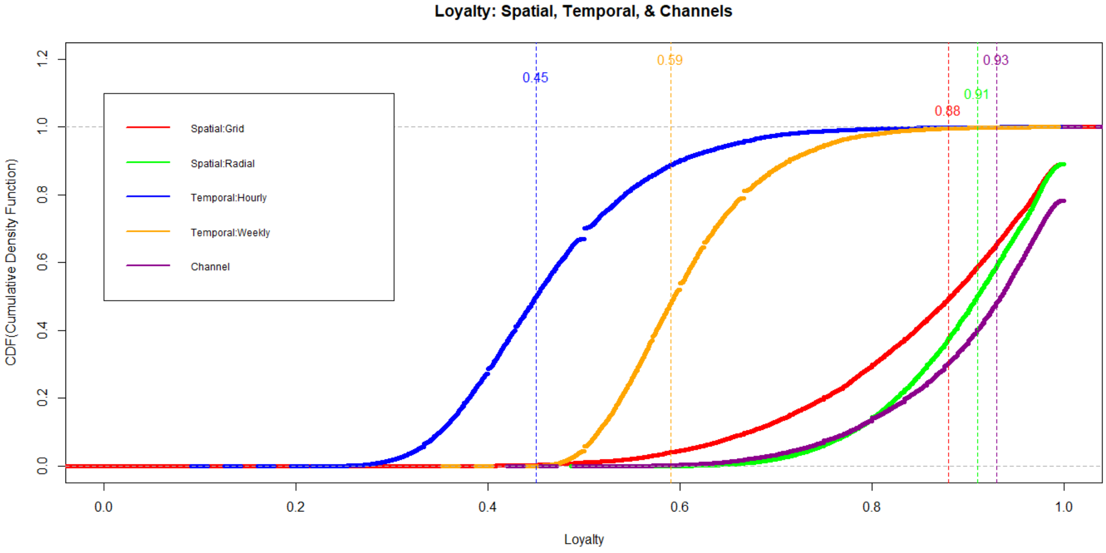

Figure 6E: The cdf graph shows the median scores for loyalty in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and channel categories. It shows that out of all the bins, the customers were the most loyal (0.93) when deciding which channel to use when interacting with the bank.

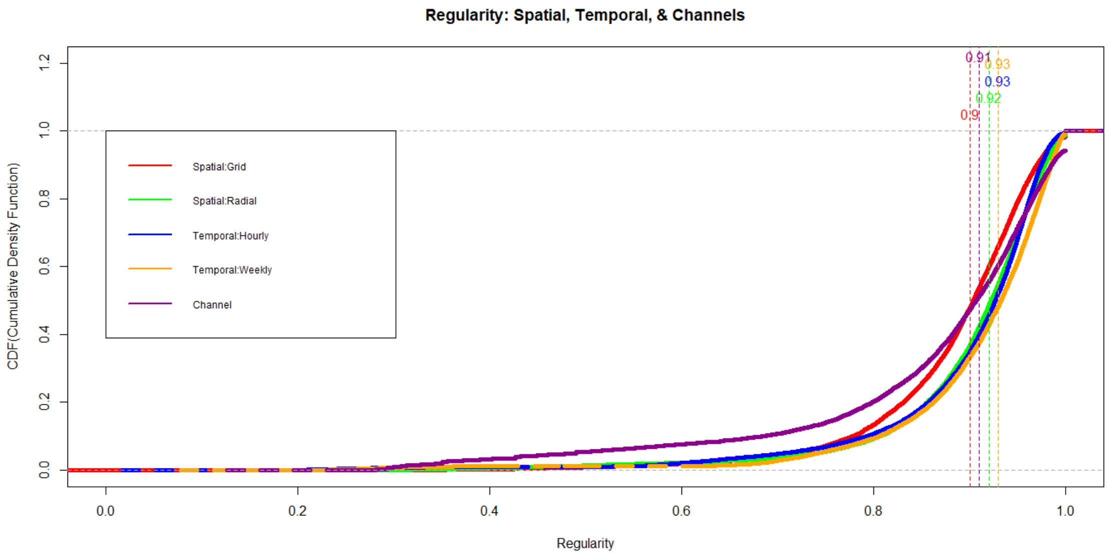

Figure 6F: The cdf graph shows the median scores for regularity in each of the five different bins: spatial grid, spatial radial, temporal hourly, temporal weekly, and channel categories. It shows that customers were regular in all the bins, including the channels they preferred to use when interacting with the bank. Overall, the customers exhibited very similar behavioral patterns over time.

D) Regression analysis of the dependent variables (overspending, trouble, late payment) was conducted and which revealed that the features related to both shopping categories and channel categories had a significant association with them. The following graphs show the odds ratios for each of the three financial outcome variables, i.e. overspending, trouble, late payment, in terms of each of the shopping entropies and categories, and channel entropies and categories.

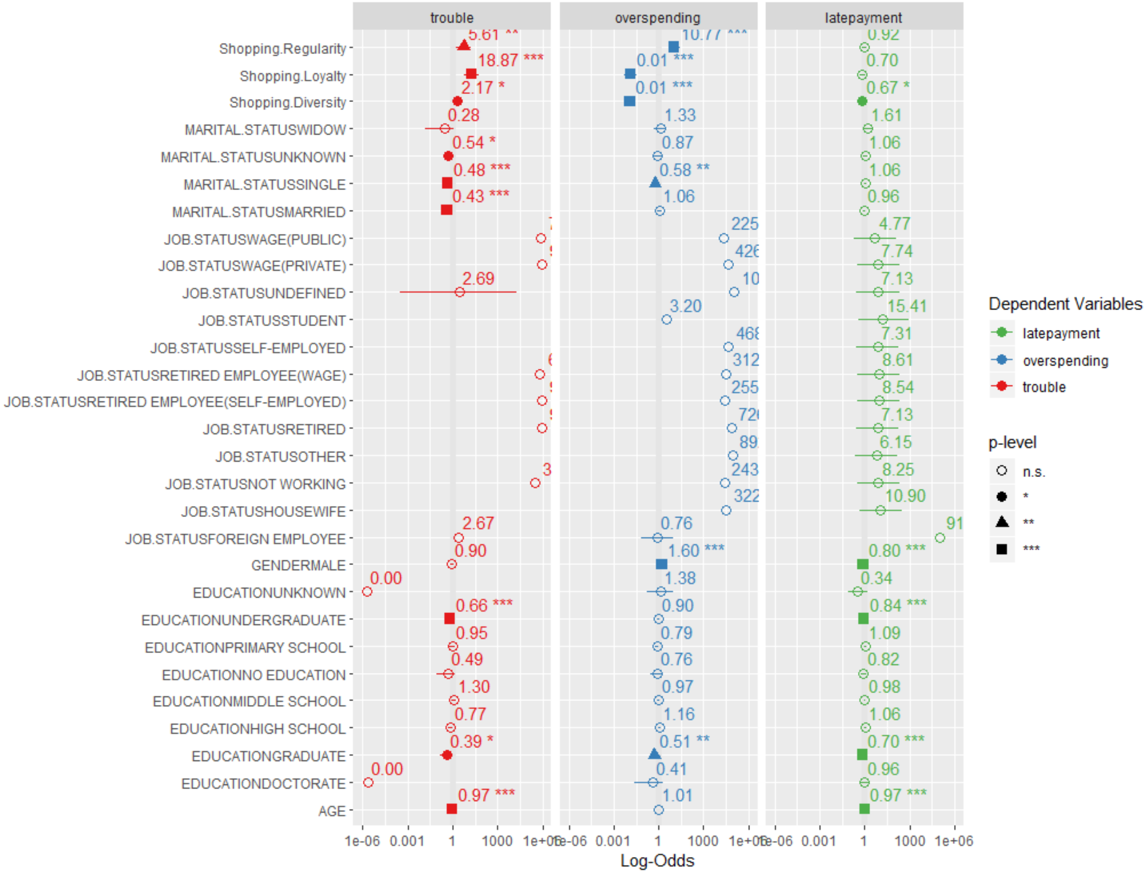

Figure 7A: The figure shows significant associations observed during logistic regression performed between the demographic and shopping entropy features and the financial outcomes. Customers who were more regular in the shopping categories which they preferred to buy were more likely to be in trouble and overspend. Customers who were more loyal and for whom a great percentage of their shopping was accounted for by the top three categories which they preferred to buy were also more likely to be in trouble but were less likely to overspend. Diverse shoppers were also more likely to be in trouble, but less likely to overspend or miss their payments. A new demographic category, individuals whose marital status was unknown, also showed a new significant association and they were less likely to be in trouble.

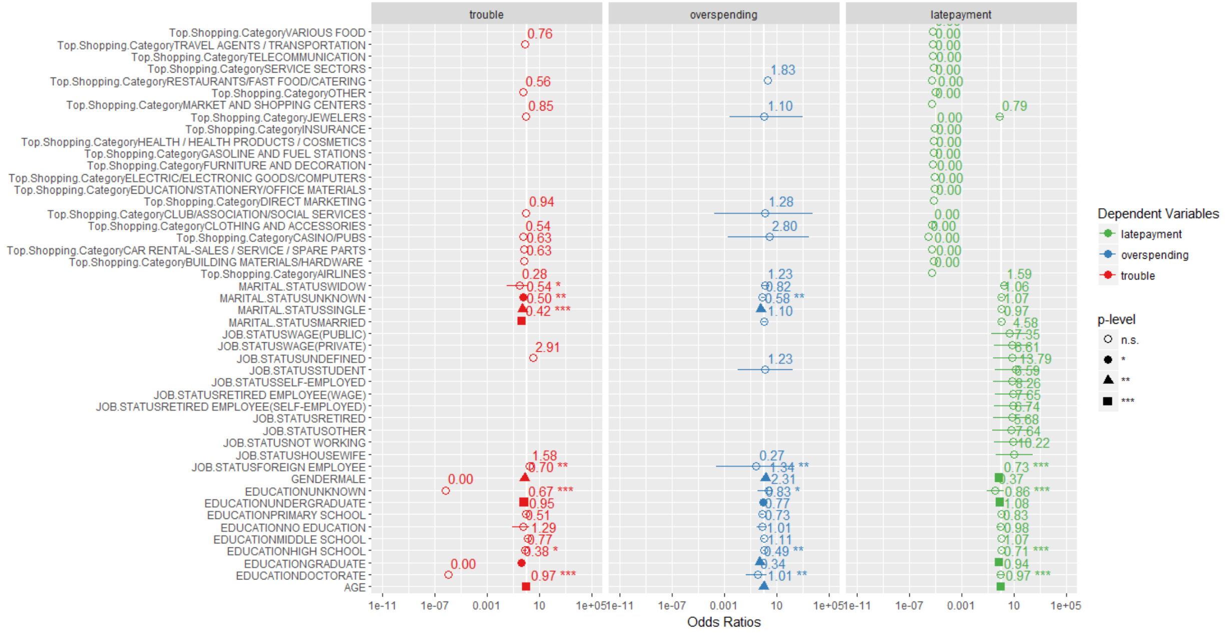

Figure 7B: The figure shows significant associations observed during logistic regression performed between the demographic and the financial outcomes. However, no significant associations between the shopping categories and the financial outcomes could be found in the above graph.

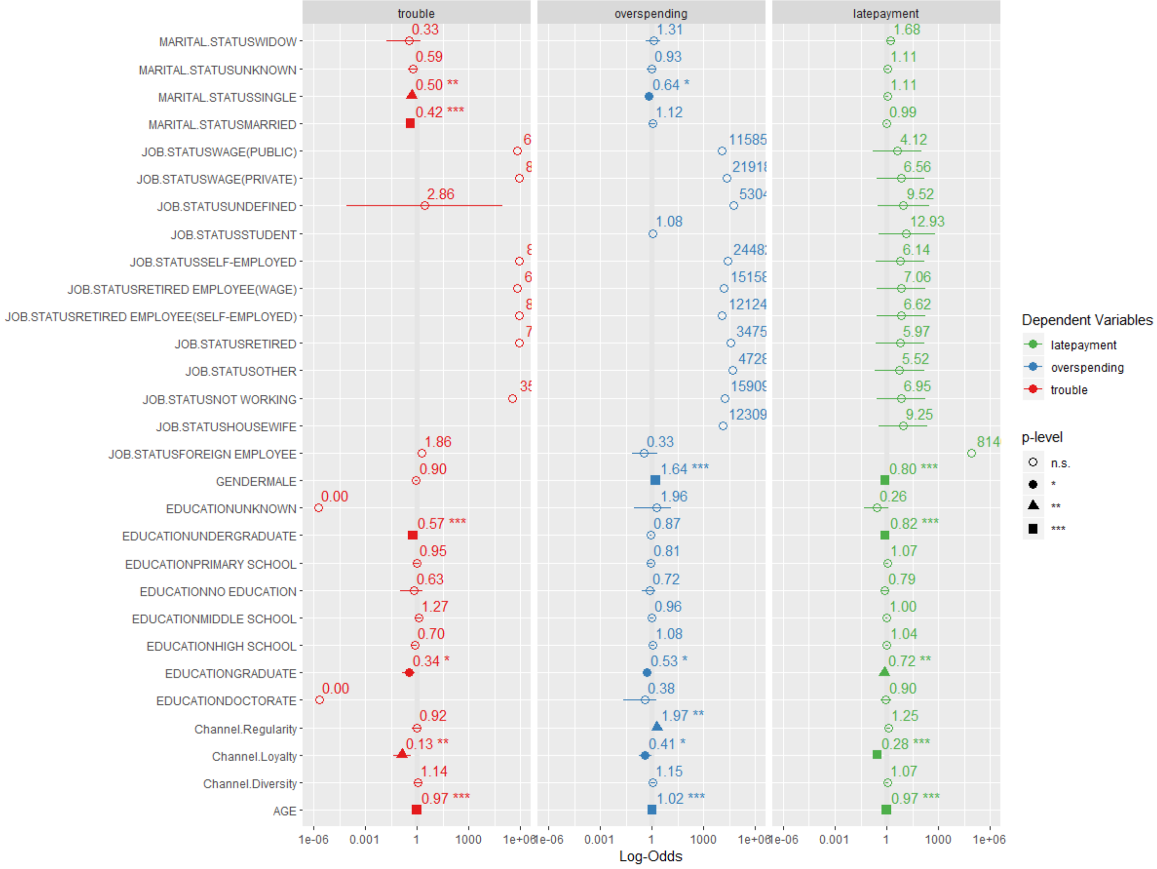

Figure 7C: The figure shows significant associations observed during logistic regression performed between the demographic and 'channel usage' entropy features and the financial outcomes. Customers who were more regular in channel usage were more likely to overspend. Customers who were more loyal in terms of the top three channels which they preferred to use were less likely to be in trouble, overspend, or miss their payments. No significant association between channel usage diversity and any of the financial outcomes could be found in the above graph.

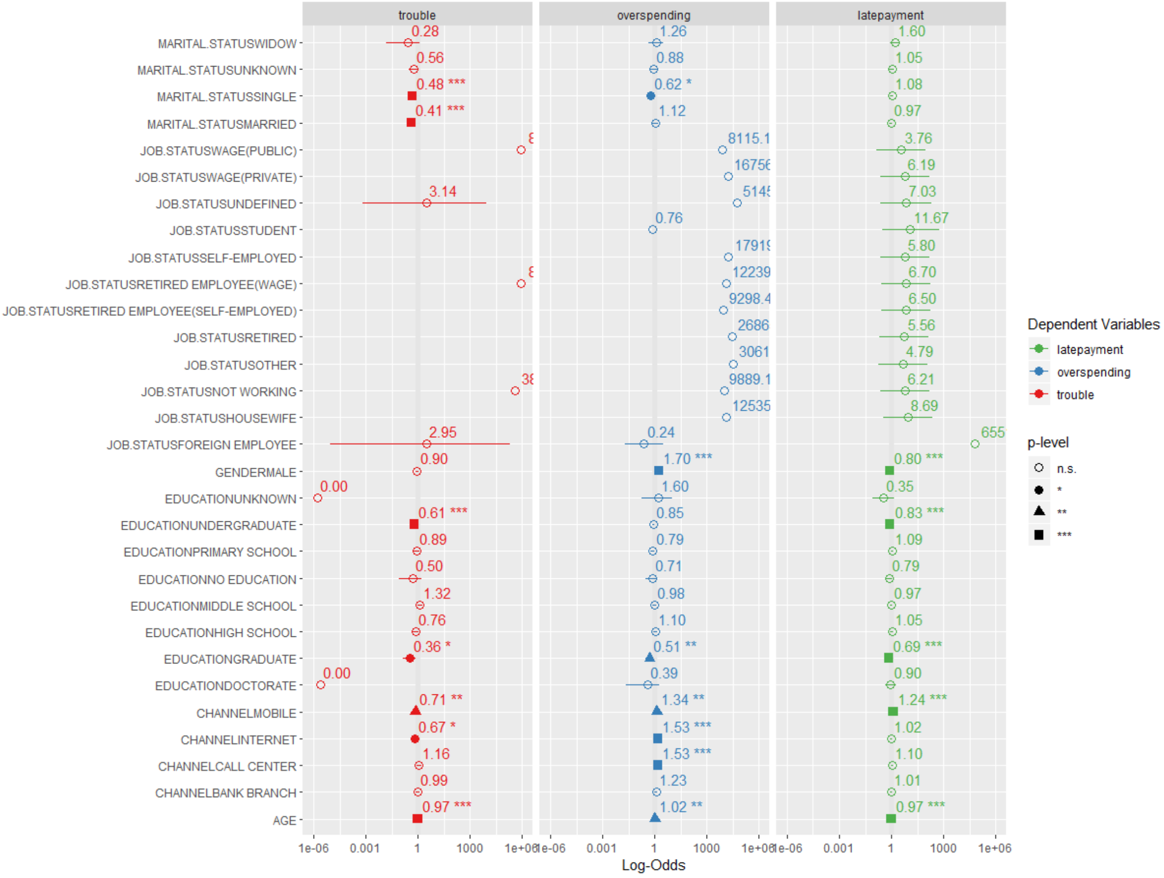

Figure 7D: The figure shows significant associations observed during logistic regression performed between the demographic features and channel categories and the financial outcomes. Customers who preferred to use the mobile application for their banking needs were less likely to be in trouble, but more likely to overspend and miss their payments. Customers who preferred to use the internet were less likely to be in trouble but more likely to overspend. Similarly, customers who preferred to use the call center were also more likely to overspend.

E) The bagging algorithm was again used to predict each of the three financial outcome variables using the new features. The following graphs show the AUC values for each of the three dependent variables (overspending, trouble, and late payment), based on 30-fold classification rounds. Each graph compares the AUC values produced when models based on only baseline, demographic, spatio-temporal, category-related or channel-related features are used to predict each of the three dependent variables.

a. Predicting Financial Outcomes using Channel Usage Entropy

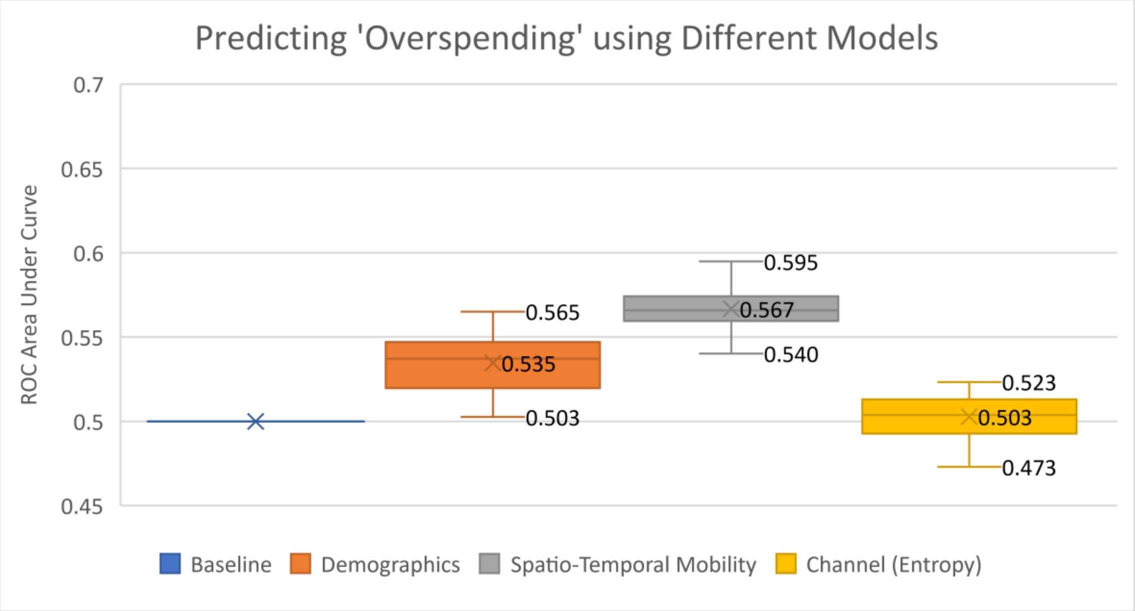

Figure 8A: The box plots above show the prediction performance for overspending using a baseline, demography-based, spatio-temporal, and channel (entropy) model. The channel (entropy) based model fails to predict better than the other models. The spatio-temporal model performs the best out of all the models.

	Overspending		
	Demography	**Spatio-Temporal Mobility**	**Channel-Usage Entropy**
Minimum	0.503	0.540	0.473
Mean	0.535	0.567	0.503
Median	0.537	0.566	0.504
Maximum	0.565	0.595	0.523

Table 4A: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage entropy features for predicting overspending.

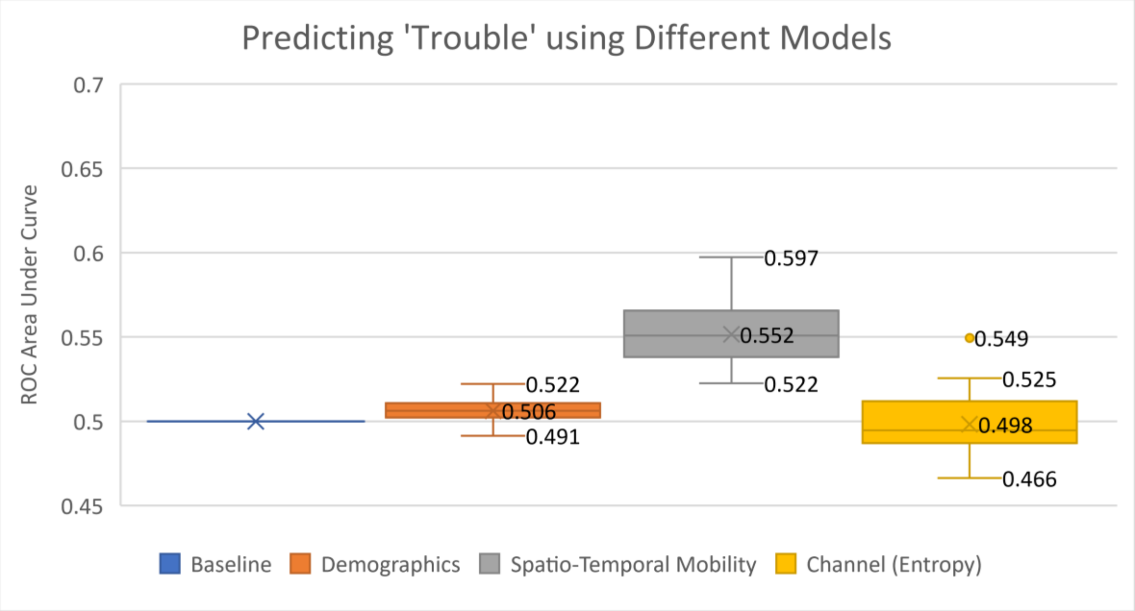

Figure 8B: The box plots above show the prediction performance for trouble using a baseline, demography-based, spatio-temporal, and channel (entropy) model. The channel (entropy) based model fails to predict better than the other models. The spatio-temporal model performs the best out of all the models.

	Trouble		
	Demography	**Spatio-Temporal Mobility**	**Channel-Usage Entropy**
Minimum	0.491	0.522	0.466
Mean	0.506	0.552	0.498
Median	0.506	0.551	0.495
Maximum	0.522	0.597	0.549

Table 4B: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage entropy features for predicting trouble.

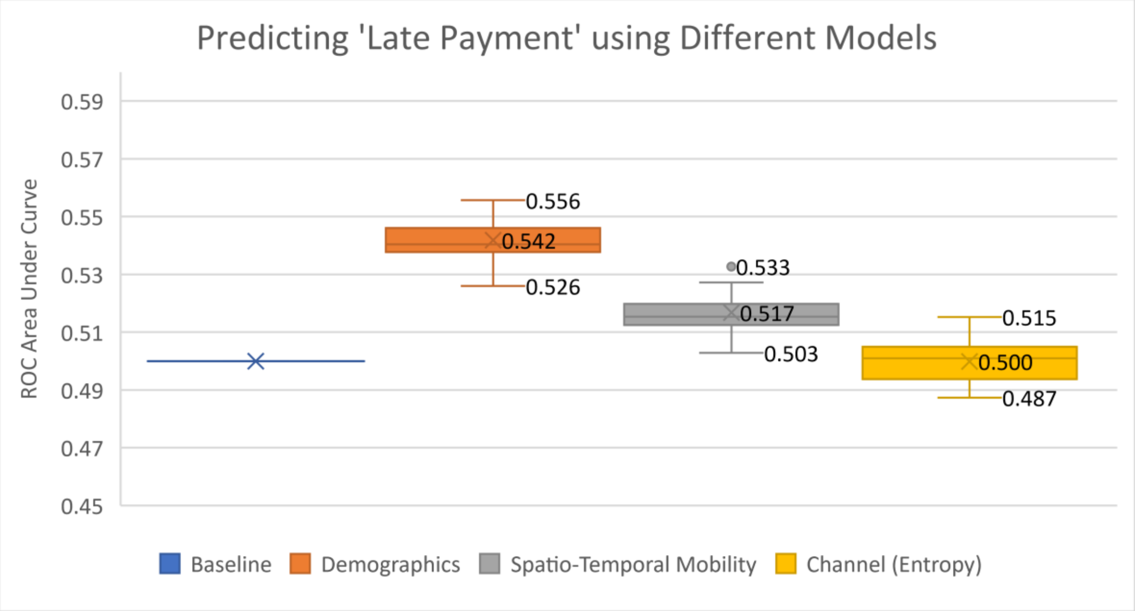

Figure 8C: The box plots above show the prediction performance for late payment using a baseline, demography-based, spatio-temporal, and channel (entropy) model. The channel (entropy) based model fails to predict better than the other models. The demography model performs the best out of all the models.

	Late Payment		
	Demography	Spatio-Temporal Mobility	Channel-Usage Entropy
Minimum	0.526	0.503	0.487
Mean	0.542	0.517	0.500
Median	0.540	0.515	0.501
Maximum	0.556	0.533	0.515

Table 4C: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage entropy features for predicting late payment.

b. Predicting Financial Outcomes using Channel Usage as a Binary Variable

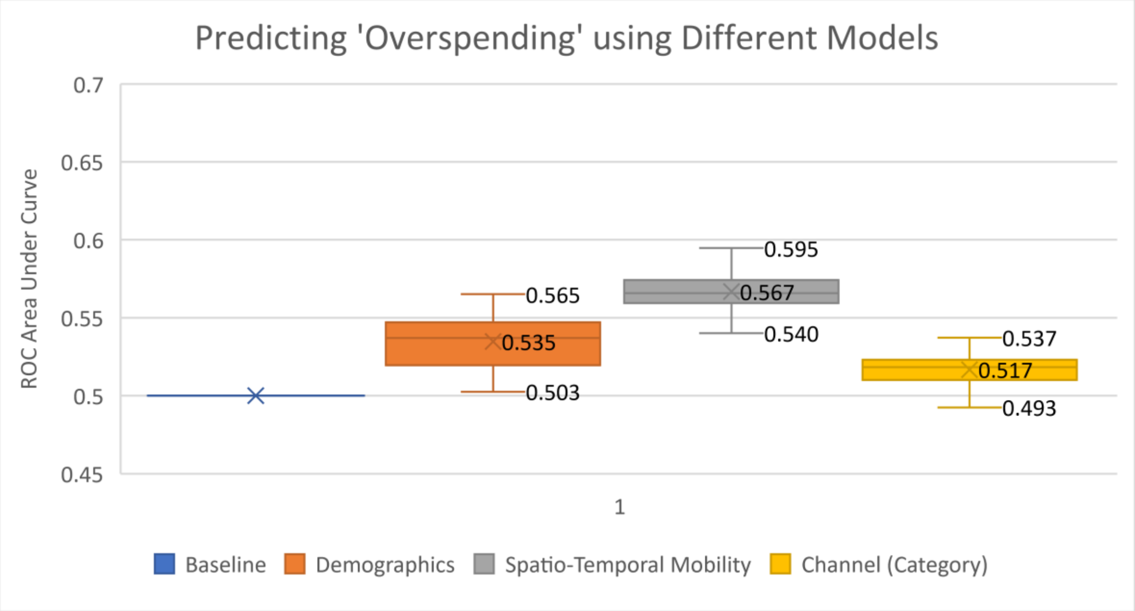

Figure 9A: The box plots above show the prediction performance for overspending using a baseline, demography-based, spatio-temporal, and channel (category) model. The channel (category) based model fails to predict better than the other models. The spatio-temporal model performs the best out of all the models.

	Overspending		
	Demography	**Spatio-Temporal Mobility**	**Channel-Usage (Binary)**
Minimum	0.503	0.540	0.493
Mean	0.535	0.567	0.517
Median	0.537	0.566	0.518
Maximum	0.565	0.595	0.537

Table 5A: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage (binary) features for predicting overspending.

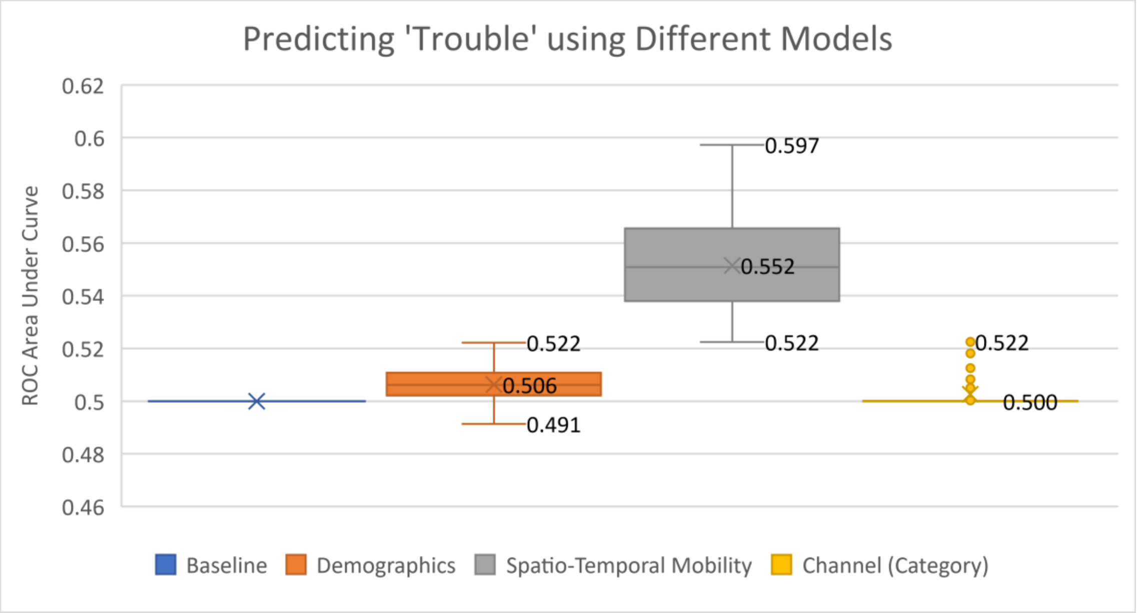

Figure 9B: The box plots above show the prediction performance for trouble using a baseline, demography-based, spatio-temporal, and channel (category) model. The channel (category) based model fails to predict better than the other models and in fact performs almost as the same as the baseline model. The spatio-temporal model performs the best out of all the models.

	Trouble		
	Demography	**Spatio-Temporal Mobility**	**Channel-Usage (Binary)**
Minimum	0.491	0.522	0.500
Mean	0.506	0.552	0.503
Median	0.506	0.551	0.500
Maximum	0.522	0.597	0.522

Table 5B: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage (binary) features for predicting trouble.

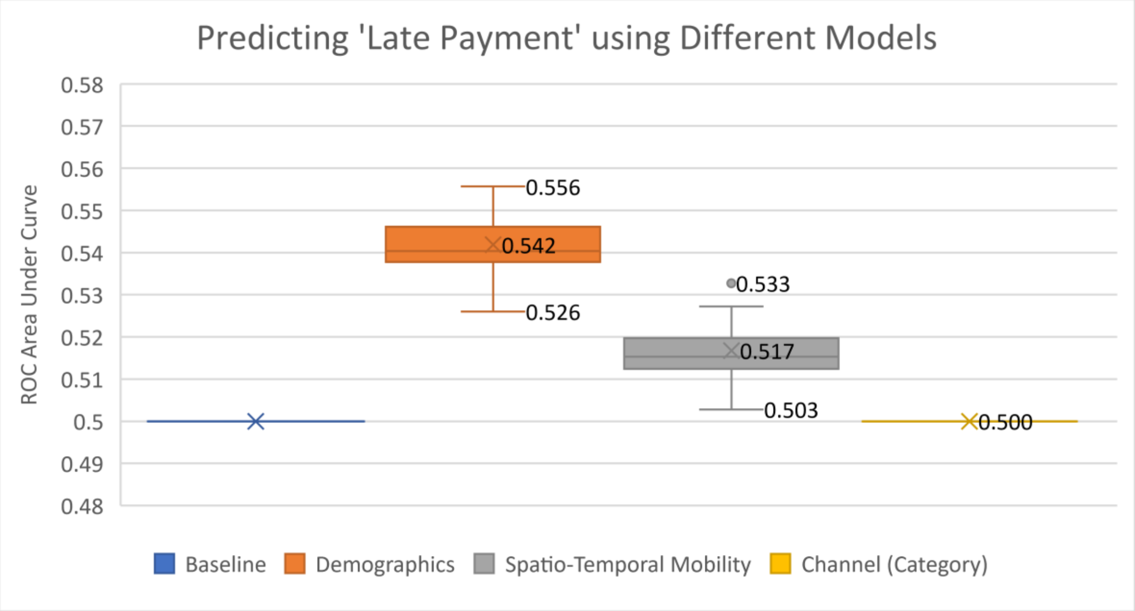

Figure 9C: The box plots above show the prediction performance for late payment using a baseline, demography-based, spatio-temporal, and channel (category) model. The channel (category) based model fails to predict better than the other models and in fact performs the same as the baseline model. The demography model performs the best out of all the models.

	Late Payment		
	Demography	**Spatio-Temporal Mobility**	**Channel-Usage (Binary)**
Minimum	0.526	0.503	0.500
Mean	0.542	0.517	0.500
Median	0.540	0.515	0.500
Maximum	0.556	0.533	0.500

Table 5C: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and channel-usage (binary) features for predicting late payment.

c. Predicting Financial Outcomes using Category (Shopping) Entropy

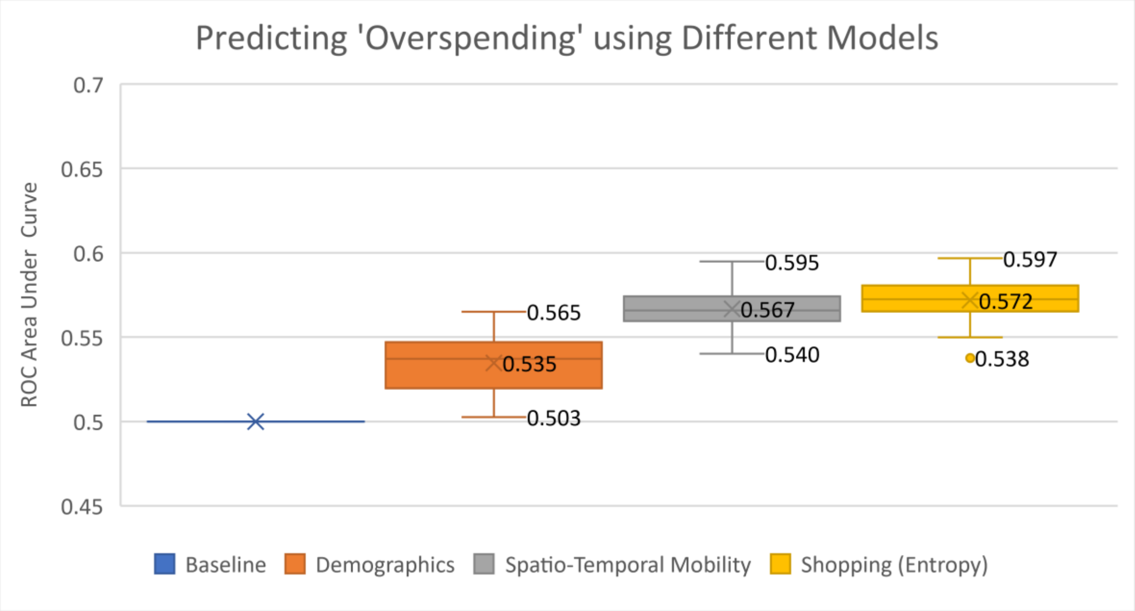

Figure 10A: The box plots above show the prediction performance for overspending using a baseline, demography-based, spatio-temporal, and shopping (entropy) model. The shopping (entropy) based model proves to predict slightly better than the other models and is the best out of all.

	Overspending		
	Demography	**Spatio-Temporal Mobility**	**Shopping (Entropy)**
Minimum	0.503	0.540	0.538
Mean	0.535	0.567	0.572
Median	0.537	0.566	0.572
Maximum	0.565	0.595	0.597

Table 6A: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (entropy) features for predicting overspending.

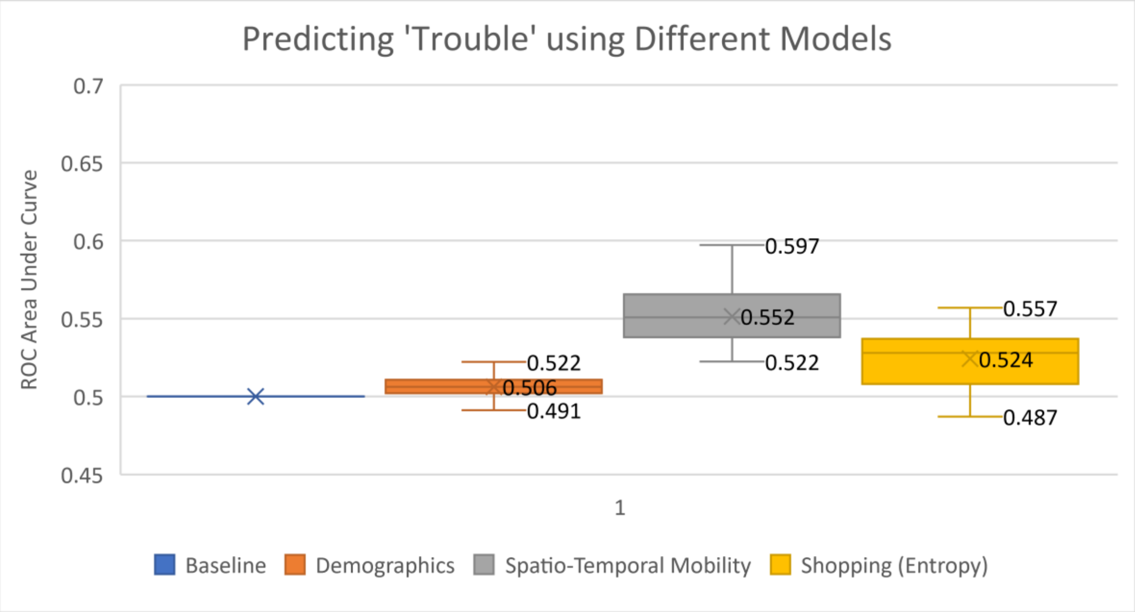

Figure 10B: The box plots above show the prediction performance for trouble using a baseline, demography-based, spatio-temporal, and shopping (entropy) model. The shopping (entropy) based model fails to predict better than the other models. The spatio-temporal model performs the best out of all the models.

	Trouble		
	Demography	**Spatio-Temporal Mobility**	**Shopping (Entropy)**
Minimum	0.491	0.522	0.487
Mean	0.506	0.552	0.524
Median	0.506	0.551	0.528
Maximum	0.522	0.597	0.557

Table 6B: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (entropy) features for predicting trouble.

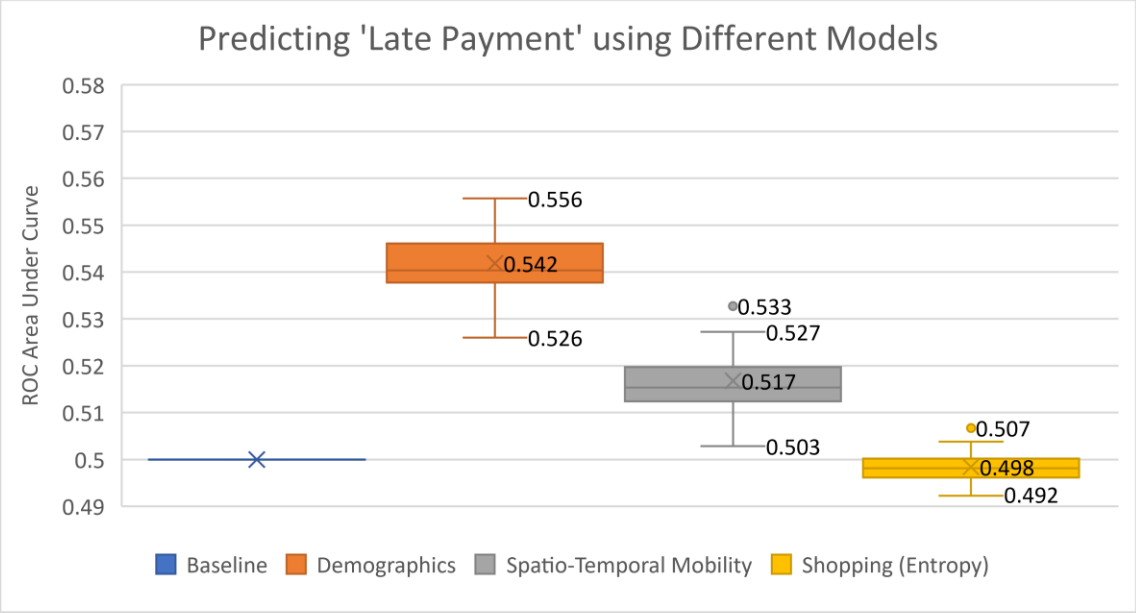

Figure 10C: The box plots above show the prediction performance for late payment using a baseline, demography-based, spatio-temporal, and shopping (entropy) model. The shopping (entropy) based model fails to predict better than the other models and in fact performs slightly worse than the baseline model. The demography model performs the best out of all the models.

	Late Payment		
	Demography	**Spatio-Temporal Mobility**	**Shopping (Entropy)**
Minimum	0.526	0.503	0.492
Mean	0.542	0.517	0.498
Median	0.540	0.515	0.498
Maximum	0.556	0.533	0.507

Table 6C: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (entropy) features for predicting late payment.

d. Predicting Financial Outcomes using Shopping Categories as a Binary Variable

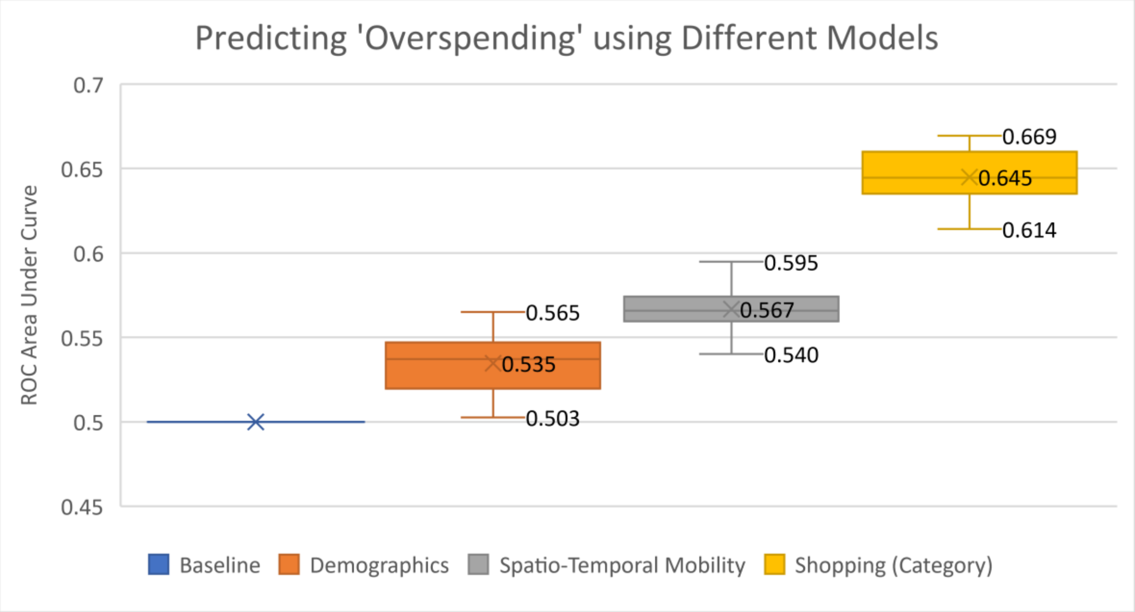

Figure 11A: The box plots above show the prediction performance for overspending using a baseline, demography-based, spatio-temporal, and shopping (category) model. The shopping (category) based model proves to predict overspending better than the other models and is the best out of all. It predicts overspending nearly 14% better than the spatio-temporal mobility model.

	Overspending		
	Demography	Spatio-Temporal Mobility	Shopping (Category)
Minimum	0.503	0.540	0.614
Mean	0.535	0.567	0.645
Median	0.537	0.566	0.645
Maximum	0.565	0.595	0.669

Table 7A: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (category) features for predicting overspending.

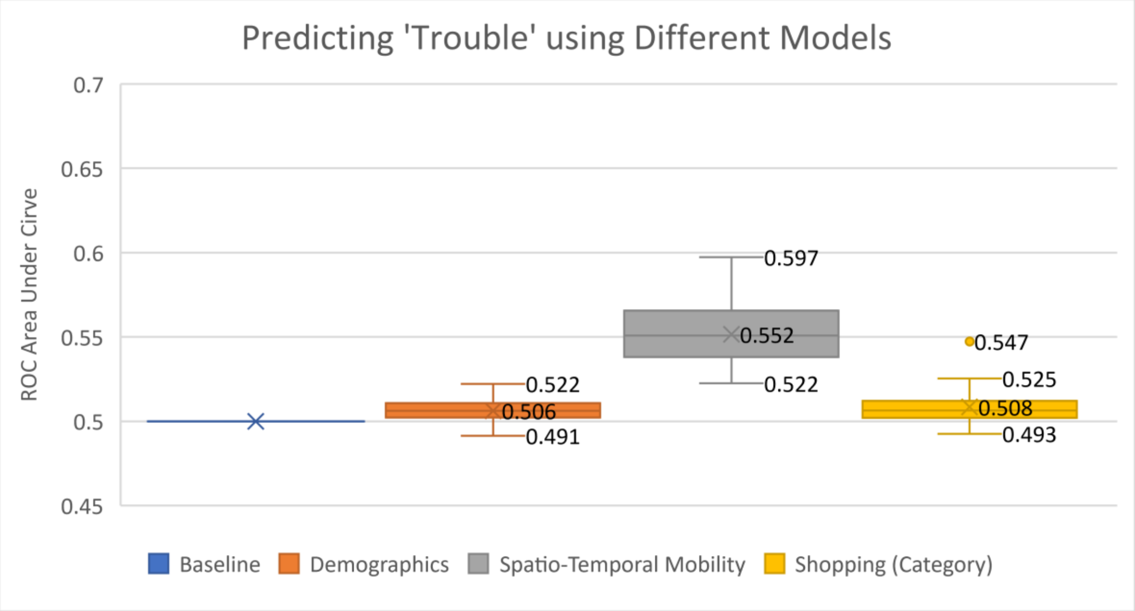

Figure 11B: The box plots above show the prediction performance for trouble using a baseline, demography-based, spatio-temporal, and shopping (category) model. The shopping (category) based model fails to predict better than the other models. The spatio-temporal model performs the best out of all the models.

	Trouble		
	Demography	**Spatio-Temporal Mobility**	**Shopping (Category)**
Minimum	0.491	0.522	0.493
Mean	0.506	0.552	0.508
Median	0.506	0.551	0.506
Maximum	0.522	0.597	0.547

Table 7B: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (category) features for predicting trouble.

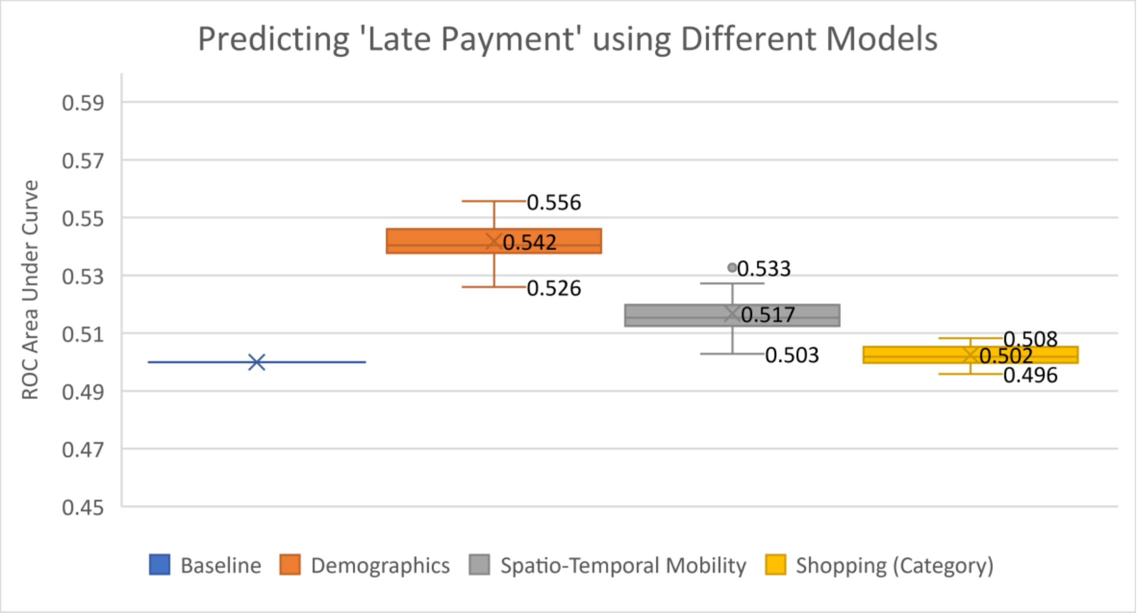

Figure 11C: The box plots above show the prediction performance for late payment using a baseline, demography-based, spatio-temporal, and shopping (category) model. The shopping (category) based model fails to predict better than the other models and in fact performs almost the same as the baseline model. The demography model performs the best out of all the models.

	Late Payment		
	Demography	**Spatio-Temporal Mobility**	**Shopping (Category)**
Minimum	0.526	0.503	0.496
Mean	0.542	0.517	0.502
Median	0.540	0.515	0.502
Maximum	0.556	0.533	0.508

Table 7C: The above table summarizes basic summary statistics for the AUCs of the models built using demography, spatio-temporal mobility, and shopping (category) features for predicting late payment.

Chapter 5

CONCLUSION

In this thesis, we first attempted to validate the findings by Singh et al. (2015) on a transactional database provided by a different major bank in the same OECD country. We implemented the same features defined and introduced by Singh et al. (2015) and applied the same bagging model for predicting three output indicators of financial well-being: overspending, trouble due to bank's administrative action, and late payment.

The results of our research validate the significance of spatio-temporal mobility features over demographic features in predicting an individual's propensity to 'overspend' or be in financial 'trouble'. In the case of predicting whether a customer is bound to miss their payments or not, the demographic features of the individual helped us predict to a better accuracy. Specifically, spatio-temporal features predicted 'overspending' and 'trouble' better than demographic features by around 6% and 9% respectively. Whereas the demographic features predicted 'late payment' by at least 5% better than the spatio-temporal features.

In the second part of the thesis, we introduced new input variables designed as behavioral features using the same dataset provided by A-Bank. Here we considered shopping categories and banking channels as new measures depicting customer behavior. The results with this new set of features suggest that considering the most frequented shopping category helped in predicting 'overspending' better than all the other features. Precisely, it predicted 'overspending' better than the spatio-temporal features by at least 14%. This proves that an individual's behavior, which includes both his/her spatio-temporal mobility and purchasing behavior, can significantly predict whether they will overspend or not. These findings can be considered to be of high importance since they have the ability to affect the markets of credit repayment and credit card limit decisions worth over a trillion dollars, (Singh, Bozkaya, & Pentland, 2015).

In addition to the entropy measures and binary indicator variables, there are other types of variables or features which could be considered as inputs to the prediction model. For example, 'age' can be treated as an interval variable and several age interval groups (e.g. age 19-25, 26-35) can be created out of the data. In addition, data regarding customer's banking age, product ownership of equity- and debt-related products, and even the number of credit cards owned could be experimented with to see if any such association can be found with the financial outcomes.

An important point to remember is that this research was carried out on a dataset belonging to a single major metropolitan city of the OECD country. Big-city residents have the propensity to travel long distances both within the city and outside of it. Hence, for other cities the mobility and financial behavior dynamics can be quite different. For example, residents of small cities may not have great variation in mobility overall and may prefer to travel very less distances within the city due to the proximity of places such as stores and bank branches. Another issue is the possible lack of point-of-sale (POS) and other mobility-tracking devices, which might not be available say at the local butcher or clothing shop. Hence, this can affect the range of data available to generate behavioral features out of them. Furthermore, due to a lower level of education and income, small-city residents may not even be aware or prefer to own a credit card due to their aversion to risk (Tavor & Garyn-Tal, 2016), which eventually affects the level of credit card transaction information available for data modeling.

In addition, there are several factors which can affect the validity of the model presented in this research. Firstly, dataset differences in terms of customer profiles can affect the findings. For example, a younger-aged dataset may have a higher percentage of over-spenders and late payers, as compared to an older-aged dataset. Also, a dataset where the average customer income is quite low may not show a lot of variation in mobility and may also be very regular and loyal in behavior. In terms of the types of banks, where a private bank encourages lending and charges higher interest rates, a public bank will discourage lending but charge low interest rates. This can have implications for the spending behavior of customers: the dataset of a private bank may have high diversity in both shopping behavior and even the shopping categories which are preferred, due to the easy availability of credit. This may even lead to a high percentage of people who are in financial trouble, over-spending, and paying their dues late. On the other hand, the dataset of a public bank may have high regularity in terms of shopping behavior and have low number of people who are in financial trouble, overspending, or paying their dues late. And finally, as transactions are moving online, the validity of mobility-based features to predict financial outcomes can be threatened, and hence there can be a shift to using other types of variables. These might then be related more to temporal features, or even to demographics such as age and marital status.

Overall, the availability of an individual's transactional profile now allows us to create personalized models which can be used to predict their behavior. The modeling techniques and analysis described in this thesis can even be done in advance of the individual's credit card payment deadlines, hence allowing banks to take preemptive measures before the customer gets into 'trouble' or 'overspends'. As financial data arising from POS machines and ATMs increasingly become geo-coded, it is expected that the information related to a person's spatio-temporal mobility and spending habits will become increasingly available, allowing us to create more robust and accurate models.

Individuals may even use the observations and information presented in this research to rectify their own behavior before they display any of the negative financial outcomes. For example, if a user is becoming increasingly regular in the shopping locations which they visit or in the day of the week they shop at, he/she can be notified that they have a high risk of overspending. This will allow the user to modify and become conscious of their behavior. However, the user may or may not wish to take heed of this information and may either ignore the message altogether or use it to rectify their behavior.

In the end, a side piece of information albeit an important one is that all these findings are based on an individual's mobility which can have implications on their privacy. It is imperative that users become aware of the level of information which they share, both consciously and unconsciously. In a Bit9 report (2012), it was discovered that more than 40% of mobile applications ask permission to use the location of the user (Sverdlove & Cilley, 2012), and according to Wired.com, some of these applications such as a flashlight do not even need such level of information to function, (McMillan, 2014). Hence, to not become a victim to a breach of privacy, one should be conscious of which applications and systems one has allowed for tracking.

REFERENCES

Achtziger, A., Hubert, M., Kenning, P., Raab, G., & Reisch, L. (2015). Debt out of control: The links between self-control, compulsive buying, and real debts. *Journal of Economic Psychology*, 141-149.

Christiansen, J., Fatnani, S., Kolhatkar, J. S., & Srinivasan, K. (2001). *Washington, DC Patent No. US 6,202,053.*

Constangiora, A. (2011). Consumer Credit Scoring. *Romanian Journal of Economic Forecasting, 3*, 162-177.

Corsetti, G., Pesenti, P., & Roubini, N. (1999). What caused the Asian currency and financial crisis? *Japan and the world economy*, 305-373.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465.

Dhar, R., & Wertenbroch, K. (2000). Consumer Choice Between Hedonic and Utilitarian Goods. *Journal of Marketing Research*, 60-71.

Dong, X., Suhara, Y., Bozkaya, B., Singh, V. K., Lepri, B., & Pentland, A. S. (2018). Social Bridges in Urban Purchase Behavior. *ACM Transactions on Intelligent Systems and Technology (TIST), 9*(3), 33.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779.

Hand, D. J. (2001). Modelling consumer credit risk. *IMA Journal of Management Mathematics, 12*(2), 139-155.

Hand, D. J., & Henley, W. E. (1997). Statistical Classification Methods in Consumer Credit Scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 160*(3), 523-541.

Henley, W. E., & Hand, D. J. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *The statistician*, 77-95.

Hui-Yi, L., & Nigel, H. (2012, September 25). Effects of shopping addiction on consumer decision-making: Web-based studies in real time. *Journal of Behavioral Addictions, 1*(4), 162-170.

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Rowland, J., & Varshavsky, A. (2010). A tale of two cities. *HotMobile '10 Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications* (pp. 19-24). Annapolis, Marland: ACM New York.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767-2787.

Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications, 37*(9), 6233-6239.

Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individial probability estimates using machine learning. *Expert Systems with Applications, 40*(13), 5125-5131.

Kung, K. S., Greco, K., Sobolevsky, S., & Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *Plos One,* 9.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., . . . Jebara, T. (2009). Life in the network: the coming age of computation social science. *Science, 323*(5915), 721.

McMillan, R. (2014, October 20). *THE HIDDEN PRIVACY THREAT OF ... FLASHLIGHT APPS?* Retrieved from Wired: https://www.wired.com/2014/10/iphone-apps/

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. *SOCIALCOM PASSAT '12 Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conferecne on Privacy, Security, Risk, and Trust* (pp. 144-153). Amsterdam: IEEE.

Singh, V. K., Bozkaya, B., & Pentland, A. (2015). Money Walks: Implicit Mobility Behavior and Financial Well-Being. *Plos One*.

Sobolevsky, S., Sitko, I., Des Combes, R. T., Hawelka, B., Arias, J. M., & Ratti, C. (2014). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in Spain. *Big Data (BigData Congress), 2014 IEEE International Congress* (pp. 136-143). IEEE.

Sverdlove, H., & Cilley, J. (2012, October). *Pausing Google Play: More Than 100,000 Android Apps May Pose Security Risks With Mobile Security Survey.* Retrieved July 7, 2018, from Carbon Black: https://www.bit9.com/download/reports/Pausing-Google-Play-October2012.pdf

Tavor, T., & Garyn-Tal, S. (2016). Further examination of the demographic and social factors affecting risk aversion. *Financial Markets and Portfolio Management, 30*(1), 95-110.