HOTEL ROOM SALES PREDICTION FOR A TRAVEL AGENCY

by
NAZLI DENIZ TÜRKER

Submitted to the Sabancı Graduate Business School
in partial fulfilment of
the requirements for the degree of Master of Science in Business Analytics

Sabancı University
July 2021

HOTEL ROOM SALES PREDICTION FOR A TRAVEL AGENCY

Approved by:

Approval Date: July 12, 2021

ABSTRACT

HOTEL ROOM SALES PREDICTION FOR A TRAVEL AGENCY

NAZLI DENIZ TÜRKER

Business Analytics, Master's Thesis, July 2021

Thesis Supervisor: Prof. Dr. Abdullah Daşçı

Thesis Co-Supervisor: Prof. Dr. Burçin Bozkaya

Keywords: Tourism Analytics, Sales Prediction, Hotel Sales Prediction

Predicting sales can be extremely beneficial to the tourism industry because it allows planners and managers to foresee future performance. This allows travel agencies to make more informed decisions about facilities, improve their contracts with more favorable terms, and offer better deals to customers in order to maximize their revenue and minimize their loss. Sales prediction enables travel agencies to adjust prices based on facility supply and customer demand, focus on sales to different demographics or change their marketing strategy to attract more customers of a specific segment. In this thesis, we compare various statistical and machine learning models on several datasets containing basic information on hotels, hotel features, and points of interests (PoI) near hotels in order to present a robust and accurate solution to hotel room sales prediction problem based on real-life data from one of the largest travel agencies in the Turkish tourism market. The results show that machine learning regression models have a great potential for hotel sales prediction. Random Forest Regression is outstanding with the highest goodness of fit and Support Vector Regression is good at accuracy values in the majority of the cases. Besides, there is a significant difference between the predictive performances by using All Segments and Two Adults Segment datasets. Additionally, the results with PoI datasets are also as good as the results without PoI datasets.

ÖZET

BIR SEYAHAT ACENTESI İÇIN OTEL ODA SATIŞ TAHMINI

NAZLI DENIZ TÜRKER

İş Analitiği Yüksek Lisans Tezi, Temmuz 2021

Tez Danışmanı: Prof. Dr. Abdullah Daşçı

Tez Eş Danışmanı: Prof. Dr. Burçin Bozkaya

Anahtar Kelimeler: Turizm Analitiği, Satış Tahmini, Otel Satış Tahmini

Gelecekteki performansın ongorulebilmesine olanak tanıması anlamında satış tahmini, turizm endüstrisi için son derece faydalı bir araçtır. Bu sayede seyahat acentelerinin planlayıcıları ve yöneticileri, şirket gelirlerini en üst düzeye çıkarmak ve operasyonel zararı en aza indirmek için anlaşma yapılacak tesisler hakkında daha ölçülü kararlar verebilir, sözleşmeleri daha uygun şartlarla hazırlayabilir ve müşterilerine daha iyi fiyatlar sunabilirler. Doğru bir tahminleme ile tercih edilen profilden daha fazla müşteri çekilebilir, tesislerin kapasitelerine ve müşterilerin taleplerine göre fiyat ayarlaması yapılabilir, satışların farklı demografik özelliklere odaklamasına veya pazarlama stratejilerinde değişiklik yapılmasına imkan sağlanabilir. Bu tezde, Türkiye turizm pazarının en büyük seyahat acentelerinden birinin verilerine dayanarak otellerin oda satış tahmin sorununa sağlam ve doğru bir çözüm sunmak amacıyla otel, otel özellikleri ve otellerin etrafındaki çekim alanı noktaları (PoI) temel bilgileri ile çalıştırılan farklı istatistiksel ve makine öğrenimi modelleri karşılaştırılmıştır. Sonuçlar, makine öğrenimi regresyon modellerinin otel satış tahmini için büyük bir potansiyele sahip olduğunu göstermektedir. Alınan sonuçların büyük bir çoğunluğuna göre Random Forest Regresyonu uyum iyiliği anlamında en yüksek başarı oranına sahipken, doğruluk anlamında en iyi sonuçları Support Vector Regresyonu vermektedir. Ayrıca, Tüm Segmentler ve İki Yetişkin Segment veri kümelerini kullanarak yapılan tahminlerin sonuçları arasında performans anlamında önemli bir fark vardır. Bununla birlikte, PoI eklenen ve PoI eklenmeyen veri kümelerinin birbirlerine göre bir üstünlükleri görülmemektedir.

ACKNOWLEDGEMENTS

First and foremost, I am eternally grateful to my beloved family for all the love, support, and vision they have given to me. They have been always a source of motivation and inspiration for me to become the best version of myself. Additionally, I would like to thank them for encouraging and believing in me to continue my education after a long gap.

I would like to express my sincere gratitude to my advisors Prof. Abdullah Daşçı and Prof. Burçin Bozkaya for their invaluable guidance and mentoring in my thesis. I am thankful for their support and belief in me throughout my studies.

I would like to acknowledge the support and mentoring given by Prof. Ulf Nilsson at Sabancı University during and after my assistantship for him.

My deepest gratitude to Prof. Hiroshi Mamitsuka for his support and guidance during my exchange period at Aalto University.

I am grateful to Sabancı Business School and my professors for awarding me with a scholarship in Masters in Business Analytics program and giving me the opportunity to join the Erasmus+ Programme.

I would like to thank Aalto University and Exchange Office for their hospitality during my exchange studies. I have joyful memories of the time spent in Finland.

I am grateful to the tourism agency for their support and belief in my work and for providing me with the data.

Finally, I would like to say special thanks to all my dear friends for their continuous encouragement, support, and patience during my studies. I am looking forward to spending more time with you from now on.

To My Beloved Family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Tourism, as one of the world's most important industries, refers to any practice that involves people traveling to places other than their usual residence for a short period of time. It is based on people traveling to various locations for a variety of reasons and also consists of a large network of interconnected industries.

A travel agency plays a critical role in this network because it is involved in the entire process of creating and promoting all of these activities. It is a travel agency that organizes and processes all of attractions, access points, facilities, and related services for visitors. An ideal travel agency arranges for travel tickets, travel documents, accommodation, entertainment, and other related services from different suppliers. Because a travel agency is the one that connects the dots to reveal a good picture of a tourism experience, it bears a great deal of responsibility not only to satisfy customers but also to make a profit in order to keep the services running. As a result, planning is critical, and prediction is the first step.

Predicting sales can be extremely beneficial to the tourism industry because it allows planners and managers to foresee the future performance. This allows travel agencies to make more informed decisions about facilities, improve their contracts with more favorable terms, and offer better deals to customers in order to maximize their revenue and minimize their loss. Sales prediction enables travel agencies to adjust prices based on facility supply and customer demand, focus on sales to different demographics, or change their marketing strategy to attract more customers of a specific type. In this thesis, we compare various statistical and machine learning models on several datasets containing basic information on hotels, hotel features, and points of interests (PoI) near hotels in order to present a robust and accurate solution to hotel room sales prediction problem based on real life data from one of the largest travel agencies in the Turkish tourism market. In order to understand the essence of the operations, we analyze datasets stored on the agency's company database. To solve the problem, we create several regression models and compare them.

This thesis is organized as follows. In the next chapter, which is Chapter 2, you would find the results of the literature research which prepares a general structure for our study. Chapter 3 presents data collection, preprocessing and descriptive analysis on different datasets which are used for this study. In Chapter 4, we describe different machine learning algorithms, various parameters and the cross validation results of the built models by using several different values for these parameters. Chapter 5 comprises the evaluation and comparison of the results of the models build with different datasets. Finally, Chapter 6 concludes the thesis, also possible future research directions in this area are discussed in this chapter.

2. LITERATURE REVIEW

In this chapter, we present the results of the literature research which prepares a general structure for our study. First, we made an introduction starting with research methodologies in tourism and hospitality industries. Then we discourse on quantitative studies for prediction in these industries. In the end, we have a high level review of machine learning methodologies in different predictive studies which we use to predict hotel room sales in our study.

Tourism and hospitality are vital sectors that have long been studied qualitatively and quantitatively not just by business but also by academia. As in usual, qualitative research leads to a deeper knowledge of the subject's social, cultural, and political elements. Quantitative research, on the other hand, places an emphasis on concrete and data-driven analyses (Provenzano & Baggio, 2019). Both methodologies are really important to have an insight about the dynamics of the industries and to take the business a step further. Moreover, these methodologies complement and contribute to each other to create better outcomes.

However, with the rapid expansion of globalization and the increasing importance of the tourist industry for countries, data is becoming increasingly crucial in managing the transition. To that end, efforts are being made to analyze the general characteristics of visitors and their consumption behavior using descriptive and inferential statistics in order to reach accurate findings and conclusions (Provenzano & Baggio, 2019). These efforts, for instance, show up as advance booking models which calculate the future reservation increments and combine them into the realized demand in order to predict future demand (Lee, 2018). Although this kind of basic solutions work out in a certain sense, more comprehensive recipes are inevitably required as internal and external dynamics get more sophisticated and the industries reach a bigger scale over time.

As long as the industry planning and management requires more accurate and efficient prediction methodologies, statistics come into use frequently throughout the years. Time series analysis, for instance, steps forward with linear and nonlinear

techniques among statistical approaches. For tough questions which linear models fail to perform well or fall behind to explain, more complicated nonlinear models come to the rescue. Many researchers have resorted to nonlinear approaches such as neural networks as a result of this perspective (Yu, Wang, Gao & Tang, 2017).

In the course of time, many more academicians and business people had a chance to give some thought to this new approach. Consequently, there are enhanced applications of neural networks in tourism demand prediction and the results show that the neural networks outperform rigid statistical models such as multiple regression, moving average, and exponential smoothing in prediction. Yet neural networks have weaknesses such as the necessity for a large number of regulating parameters, the difficulty in establishing a stable solution and the risk of over-fitting. Unlike neural networks that use the empirical risk minimization principle, which is based on measuring the performance of the algorithm on training data because the distribution is unknown, machine learning approaches such as Support Vector Regression uses the structural risk minimization principle, which is based on balancing the complexity of the model against over-fitting and aims to reduce the generalization error upper bound rather than the training error (Chen & Wang, 2007).

Machine learning approaches not only provide a good solution to interpret the results of qualitative studies derived from unstructured data but also ensure that the results are reproducible (Provenzano & Baggio, 2019). When the relative gain of Machine Learning methods is compared to the auto regressive moving average (ARMA) statistical model, for example, it is discovered that the direct approach achieves the relative gain one step ahead of aggregating predictions. As prediction horizons are extended, both systems show considerable increases in prediction accuracy, but there are no significant differences between the two approaches. This discovery demonstrates that machine learning techniques are particularly well adapted to mid and long-term forecasting (Claveria, Monte & Torra, 2016).

Since data is the fuel of qualitative studies, significant attributes are required to build a descriptive model. (Rhee & Yang, 2014) study the relative variations in various hotel features based on different customer groups. According to their research, various factors contribute to the preferability of hotels, and different visitor profiles look for different attributes. They arrive at these conclusions by analyzing the total hotel rating displayed by guests. We can infer from this study that visitor ratings and hotel characteristics are useful sources for developing a descriptive model. Besides, (Xue & Zhang, 2020) also show how geographical patterns of accommodation sites influence the behavioral and consumption habits of various tourist segments. Based on the findings of this study, we can infer that PoI (point-of-interest) settlements

near hotels have a significant impact on accommodation selections.

Starting with basic concepts there are several prediction studies based on linear regression models.(Tingting & Risto, 2016), for instance, build a simple multi-linear regression model with environmental data to predict heat demand in a neighbourhood. They end up with models that present high accuracy and strong robustness. Linear regression models are compared to neural networks to predict water demand by (Pulido-Calvo, Montesinos, Roldán & Ruiz-Navarro, 2007) and they reveal that with the correct adjustments linear regression models may provide results that are as good as neural networks.(Farizal, Qaradhawi, Cornelis & Dachyar, 2020), as well, build multi-linear regression models to predict fast moving products demand and obtain accurate results with a high value of coefficient of determination which means robust and good fit models. These studies show that linear regression models can be as good as more complex models for predictive analysis.

In cases where independent variables are highly correlated, ridge regression can be a good option to study. When predicting daily crude oil prices, (Li, Zhou, Li, Wu & He, 2019) utilize the ridge regression methodology to build a model that outperforms the competition on numerous parameters. In an environment where identifying appropriate predictors is challenging to predict real estate prices, (Jae Joon, Hyun Woo, Kyong Joo & Tae Yoon, 2012) combine ridge regression with a genetic algorithm to provide a better answer.

For more complex issues machine learning algorithms come to the rescue. A decision tree model, along with a neural network model, is found to perform slightly better than the other models in terms of accuracy for predicting electricity energy consumption (Tso & Yau, 2007). (Czajkowski & Kretowski, 2016) investigate several representations of the decision tree technique for regression issues and discover that the tree representation is essential in the final prediction model.

Random forest regression is also a machine learning algorithm which is employed for though cases. In an empirical research, a customer profitability prediction model is created using random forest regression, and this model outperforms the others in terms of prediction performance (Kuangnan, Yefei & Malin, 2016). In another paper, the topic of airline departure delays is explored in depth, and a random forest regression prediction model is developed. According to the results, the model achieves a good performance no matter how sophisticated the conditions are in the airport transportation industry (Guo, Yu, Hao, Wang, Jiang & Zong, 2021). Furthermore, Amazon spot price prediction study which is made by (Khandelwal, Chaturvedi & Gupta, 2020), is another example that uses random forest regression algorithm. They compared non-parametric machine learning models and random

forest regression steps forth with higher accuracy in this study.

Support vector regression is another machine learning algorithm that has been used for prediction models. (Wu, Ho & Lee, 2004) create a support vector regression model to estimate the travel time over a short distance in China during rush hour. When compared to previous models, the findings show that the model decreases both relative mean errors and root mean squared errors considerably. Moreover, (De Leone, Pietrini & Giovannelli, 2015) perform a research on the prediction of photovoltaic energy output and provide a support vector regression model with accurate findings in terms of root mean square error, mean absolute percentage error, and coefficient of determination.

3. DATA COLLECTION, PREPROCESSING AND DESCRIPTIVE ANALYSIS

This chapter presents data collection, preprocessing and descriptive analysis on different datasets that are used for this study. There are five datasets obtained from three different sources:

- Company - Hotel Sales Transactions Dataset
- Company - Hotel Information Dataset
- Company - Hotel Features Dataset
- Google - Hotel Rating and Location Dataset
- External - Location-Based PoI(Point-of-Interest) Dataset

The company database is on a cloud server which we are granted with necessary access to run the queries and models. The database includes hotel information, hotel features and historical sales transactions. These three datasets are used together with 2 others that are obtained from external data sources to create a new constructed dataset which contains different information on hotels. With this aim, we clean and preprocess the data in each dataset and create new variables that are necessary for the study. One of the 2 external datasets is Google - Hotel Rating and Location Dataset that is created by manual data collection from Google Maps website between May 2020 and July 2020, for the hotels that are included in the Company - Hotel Sales Transaction Dataset. The other external dataset is Location-Based PoI(Point-of-Interest) Dataset that is obtained by a commercial company. Details on these and company-based datasets, and the cleaning and prepocessing steps are explained in the following sections.

3.1 Data Sources

Company - Hotel Sales Transactions Dataset consists of data on hotel sales transactions during the period between 01.01.2015 and 31.12.2019. There are 132,246 rows and 18 columns in the dataset. Each row has the information on a unique sales transaction, shown in Table 3.1.

Table 3.1 Company - Hotel Sales Transactions Dataset

Variable Name	**Variable Description**
voucherID	Unique ID per sale
customerID	Unique ID per customer
customerGender	Customer gender
customerAge	Customer age
customerSegment	Customer segment
hotelID	Unique ID per hotel
checkInDate	CheckIn date
checkOutDate	CheckOut date
salesDate	Sales date
salesChannel	Sales channel
salesAmount	Sales amount
profit	Profit amount
discountAmount	Discount amount
adultCount	Number of adult guests
childCount	Number of child guests
roomCount	Number of rooms
nightCount	Length of stay
guestNight	Number of $totalguest \times lengthofstay$

Company - Hotel Information Dataset contains the main information on hotels. It consists of 3,405 rows and 10 columns. Each row has information on a unique hotel, shown in Table 3.2.

Table 3.2 Company - Hotel Information Dataset

Variable Name	**Variable Description**
hotelID	Unique ID per hotel
hotelName	Hotel name
hotelCategoryName	Hotel category
hotelType	Hotel type
hotelChainName	Hotel chain name if part of a chain
longitude	Geographic east–west coordinate
latitude	Geographic north–south coordinate
isCompanyHotel	If the hotel has an agreement or not
isSingleManForbidden	If single man accepted or not
acceptableMinAge	The minimum age to be accepted

Company - Hotel Features Dataset contains the feature information of each hotel. There are 177 distinct features in the dataset. We organise the dataset by grouping features according to their functionalities as shown in Table 3.3 and detecting the number of features of each hotel within each category.

Table 3.3 Company - Hotel Features Dataset

Category Name	**Category Example**
accessibility	Wheel chair camp, elevator etc.
generalServices	Laundry, wifi, safebox etc.
business	Conference hall, meeting room etc.
cafeRestaurant	Cafe, restaurant, bar etc.
entertainment	Cinema, casino, night club etc.
familyHoliday	Play garden, play room, baby sitter etc.
foodBeverage	Patisserie, fresh juice, ice cream etc.
healthBeauty	Sauna, spa, Turkish bath, massage etc.
indoorSports	Fitness, yoga, pilates, table tennis etc.
outdoorSports	Football, basketball, tennis etc.
pet	Pets allowed or not
religious	Prayers room, beach for women etc.
shopping	Mall, jewellery, leather shop etc.
summerHoliday	Beach, aqua park, pier etc.
transportation	Parking lot, car renting, ring service etc.
waterSports	Sailing, surfing, water polo etc.
winterSports	Skiing, snow boarding, lift etc.

Google - Hotel Rating and Location Dataset contains hotel location and rating data collected from Google Maps website by manual search during the period between May 2020 and July 2020, together with hotelID and hotelName obtained from Company - Hotel Sales Transaction Dataset. It consists of 2,300 rows and 6 columns. Each row has information on a unique hotel, shown in Table 3.4.

Table 3.4 Google - Hotel Rating and Location Dataset

Variable Name	**Variable Description**
hotelID	Unique ID per hotel given by the company
hotelName	Hotel name
longitude	Geographic coordinate that specifies the east–west position of the hotel
latitude	Geographic coordinate that specifies the north–south position of the hotel
rating	Average rating score given to the hotel
ratingCount	Total number of rating scores given to the hotel

Location-Based PoI(Point-of-Interest) Dataset, that contains longitude and latitude data of the points with social and economic importance, is obtained from a commercial data provider (Kaya, Alpan, Balcisoy & Bozkaya, 2021). We use the data

to determine the number of PoIs located within 0.5km, 1km, 3km and 5km radii of the location of the hotels for each 12 PoI categories in the dataset, shown in Table 3.5.

Table 3.5 Location Based PoI Categories

Category Name	**Category Description**
autoServices	Auto galleries, auto services, car wash, gas stations
business	Business centers, organized industrial zones, radio, TV, newspaper headquarters
communityServices	Municipal administrations, post offices, police headquarters, courthouses, student dormitories, cultural centers, wedding halls, churches, synagogues, mosques
entertainment	Cafes, bars, cultural centers, art galleries, movie theaters, theaters
financialInstitutions	Bank branches, ATMs
hospitals	Hospitals, medical centers, district polyclinics, dental care centers, dialysis centers, nursing homes, rehabilitation centers, laboratories
parkingServices	Private and public parking services
parksRecreation	Sports center, recreation facilities, parks, stadiums, beaches, bowling centers, sports clubs
restaurants	Restaurants, patisseries
shopping	Shopping malls, supermarkets, grocery stores, pharmacies, electronic stores, textile stores, petrol station markets, bakeries, tourism agencies, jewelry stores, bookshops, cosmetic stores, hairdressers
transferHubs	Train stations, bus stations, ports, airports
travelDestinations	Hotels, car rental agencies, tourist attractions

The constructed dataset, which is hotel based and includes the variables shown in Table 3.6 and Table 3.7, is created based on the five datasets mentioned previously.

3.2 Data Cleaning and Preprocessing

The datasets obtained from the company and the location-based PoI dataset are not ready to be used for the purpose of training regression models for room sales prediction. To prepare the data, we clean and process the data as described in detail as follows:

Table 3.6 Constructed Dataset

The custom-made dataset to be used in the regression models for annual room sales prediction of each hotel

Variable Name	**Variable Description**
listPricePerAdult	Average list price per adult - Continuous variable calculated from Company - Sales Transaction Dataset
discountRatio	Average sales discount ratio - Continuous variable calculated from Company - Sales Transaction Dataset
isHolidayHotel	Holiday hotel or not - Binary variable converted from hotelType feature in Company - Information Dataset
isChainHotel	Chain hotel or not - Binary variable converted from hotelChainName feature in Company - Information Dataset
isSingleManForbidden	Single man forbidden or not - Binary variable imported from Company - Information Dataset
hotelCategory2Stars	2 stars hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategory3Stars	3 stars hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategory4Stars	4 stars hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategory5Stars	5 stars hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategoryBoutique	Boutique hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategoryResort	Resort hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
hotelCategorySpecial	Special document hotel or not - Binary variable converted from hotelCategoryName feature in Company - Information Dataset
rating	Hotel average rating score on Google - Continuous variable imported from Google - Hotel Information Dataset
ratingCount	Total number of votes on Google - Continuous variable imported from Google - Hotel Information Dataset

Company - Hotel Sales Transactions Dataset which has 132,246 rows end up to be as 130,187 rows after elimination of 2,059 records due to missing (2 rows) and incorrect (2,057 rows) data. 1,602 of these 2,057 rows of data are the sales with a sales date later than check-in date, 16 of them are the sales with a sales amount of 0 (zero) or lower than 0, and the rest of them (439 observations) are the sales with a profit amount of 0 (zero) or lower than 0.

discountAmount column, which has both positive (14,832 observations) and negative numbers (100,264 observations), is converted to absolute values for consistency.

guestNight column, which has positive numbers for only 17,941 of the observations, is edited with the results obtained by calculation shown in Equation 3.1 for data

Table 3.7 Constructed Dataset

Variable Name	**Variable Description**
accessibility	Number of features in accessibility category
generalServices	Number of features in general services category
business	Number of features in business category
cafeRestaurant	Number of features in cafe and restaurant category
entertainment	Number of features in entertainment category
familyHoliday	Number of features in family holiday category
foodBeverage	Number of features in food and beverage category
healthBeauty	Number of features in health and beauty category
indoorSports	Number of features in indoor sports category
outdoorSports	Number of features in outdoor sports category
pet	Pets are allowed or not
religious	Number of features in religious category
shopping	Number of features in shopping category
summerHoliday	Number of features in summer holiday category
transportation	Number of features in transportation category
waterSports	Number of features in water sports category
winterSports	Number of features in winter sports category

correction:

$$guestNight = (adultCount + childCount) \times nightCount \tag{3.1}$$

After cleaning the Company - Hotel Sales Transactions Dataset we process the data to extract the data we use for regression models. Firstly, 3 new features are created for each transaction by using existing columns.

roomNight is the total room-night count as shown in Equation 3.2

$$roomNight = roomCount \times nightCount \tag{3.2}$$

listPricePerAdult is the average room-night price per adult before discounts as shown in Equation 3.3.

$$listPricePerAdult = \frac{salesAmount + discountAmount}{roomNight \times adultCount} \tag{3.3}$$

discountRatio is the average discount ratio of the transaction as shown in Equation

3.4.

$$discountRatio = \frac{discountAmount}{salesAmount + discountAmount} \tag{3.4}$$

In order to evaluate the impact of using segment specific data, in total 3 custom-made datasets are constructed to run the predictive models with. The first one is All Segments Dataset which consists of the information for the hotels that are sold to all 4 customer segments, that are Two Adults, Family, Group, Single, in 2019. The second one is Two Adults Segment Dataset that consists of the information for the hotels that are sold only to two adults customer segment in 2019. The third one is Family Segment Dataset that consists of the information for the hotels that are sold only to family customer segment, that are the customers which have at least one child with them, in 2019. The following features are calculated to construct these datasets: For All Segments Dataset, only 2019 transaction data, which has 25,194 rows, are used to calculate:

- Average listPricePerAdult
- Average discountRatio
- Sum of roomNight count

for each hotel.

For Two Adults Segment Dataset, only 2019 and two adults segment transaction data, which has 13,548 rows, are used to calculate:

- Average listPricePerAdult
- Average discountRatio
- Sum of roomNight count

for each hotel.

For Family Segment Dataset, only 2019 and family segment transaction data, which has 7,539 rows, are used to calculate:

- Average listPricePerAdult
- Average discountRatio
- Sum of roomNight count

for each hotel.

From 3,405 hotel records in Company - Hotel Information Dataset only 2,231 of them are also in Company - Hotel Sales Transaction Dataset, only 1,749 of them are also in Company - Hotel Feature Dataset, only 1,134 of them are sold in 2019, only 987 of them are sold to Two Adults segment in 2019, and only 673 of them are sold to Family segment. Therefore 2,271 of the hotels in Company - Hotel Information Dataset are eliminated for All Segments Dataset, 2,418 of them are eliminated for Two Adults Segment Dataset, and 2,732 of them are eliminated for Family Segment Dataset.

After cleaning, preprocessing and creating previously mentioned 3 columns for our constructed datasets from company-based data, we add isHolidayHotel and isSingleMenForbidden columns from Company - Hotel Information Dataset. Additionally, we convert hotelType, hotelCategoryName and hotelChainName into binary variables and add following columns, which are explained in Table 3.6 previously, to the constructed datasets based on the information in these columns:

- isHolidayHotel
- isChainHotel
- hotelCategory2Stars
- hotelCategory3Stars
- hotelCategory4Stars
- hotelCategory5Stars
- hotelCategoryBoutique
- hotelCategoryResort
- hotelCategorySpecial

Finally, we add rating and ratingCount from Google-based dataset to complete the base datasets to be used for predictive models.

For All Segments Dataset; four hotels that have 'Undefined' hotelCategoryName, ten hotels with missing rating and ratingCount data, ten hotels that have isCompanyHotel = 0 are eliminated.

For Two Adults Segment Dataset; eight hotels with missing rating and ratingCount data, nine hotels that have isCompanyHotel = 0 are eliminated.

For Family Segment Dataset; one hotel that have 'Undefined' hotelCategoryName, three hotels with missing rating and ratingCount data.

In all datasets there are many outliers for annual roomNight values of hotels, which is the dependent variable for our models, as shown in the boxplot in Figure 3.1 and in the summary statistics in Figure 3.2 Therefore we eliminate the hotels with the value of annual roomNight smaller or equal to 5, to eliminate hotels with very small number of annual roomNight values, and the hotels with the value of annual roomNight larger than mean + 3 * standard deviation of all hotels' annual roomNight values. The number of data points eliminated by outlier elimination in each dataset are as follows:

- All Segments Dataset - 398 hotels
- Two Adults Segment Dataset - 411 hotels
- Family Segment Dataset - 232 hotels

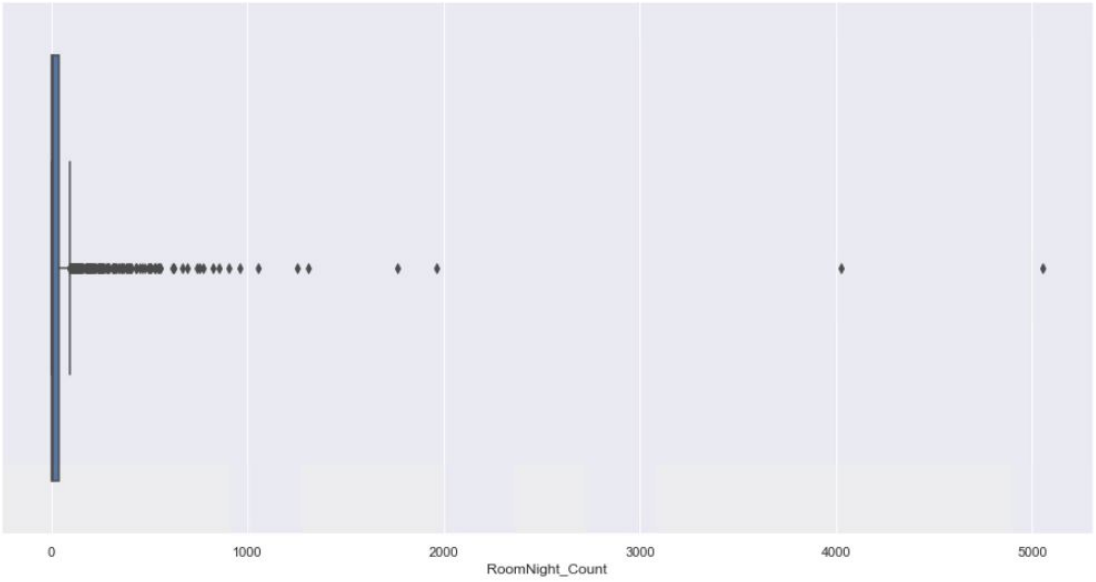

Figure 3.1 All Segments Dataset Annual roomNight Boxplot Before Outlier Elimination

```
count    1110.000000
mean       63.248649
std       241.810382
min         1.000000
25%         4.000000
50%        10.000000
75%        40.000000
max      5056.000000
Name: RoomNight_Count,
```

Figure 3.2 All Segments Dataset Annual roomNight Summary Statistics Before Outlier Elimination

After all data cleaning and preprocessing we have 712 rows in All Segments Dataset, 559 rows in Two Adults Segment Dataset, and 323 rows in Family Segment Dataset to be used in predictive models. The boxplot of the annual RoomNight values of hotels, which is the dependent variable, for All Segments dataset after outlier elimination is shown in the Figure 3.3. Additionally, the summary statistics of the dependent variable for All Segments dataset, Two Adults Segment dataset, and Family Segment dataset are shown in figures 3.4, 3.5, and 3.6 respectively.

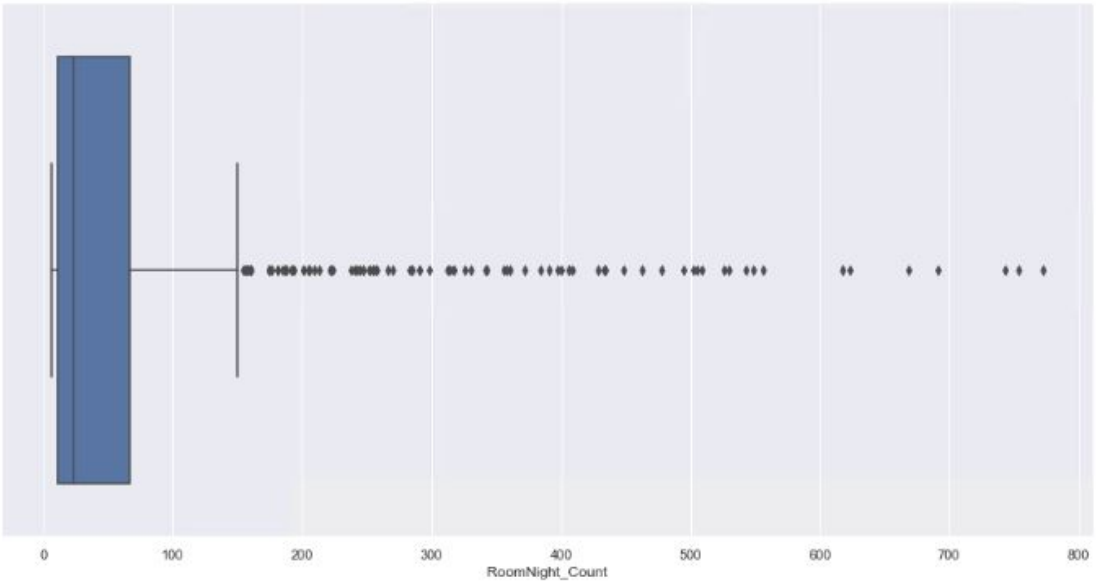

Figure 3.3 All Segments Dataset Annual roomNight Boxplot After Outlier Elimination

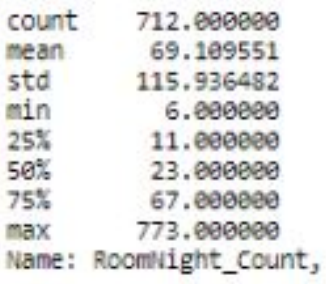

Figure 3.4 All Segments Dataset Annual roomNight Summary Statistics After Outlier Elimination

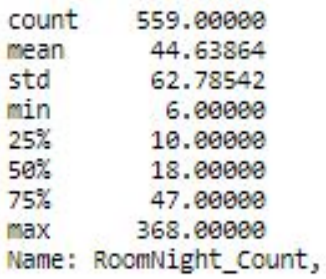

Figure 3.5 Two Adults Segment Dataset Annual roomNight Summary Statistics After Outlier Elimination

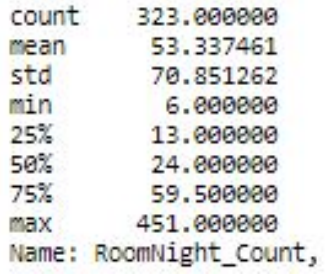

Figure 3.6 Family Segment Dataset Annual roomNight Summary Statistics After Outlier Elimination

Moreover, we use location-based PoI data to assess the impact of PoIs around the location of hotels in sales prediction by adding new columns to the base datasets. We detect number of PoIs in 12 categories within 0.5km, 1km, 3km, and 5km radii of each hotel by using Haversine which is an equation that calculates the length of an arc between two locations on longitude and latitude (Maria, Budiman, Haviluddin & Taruk, 2020).

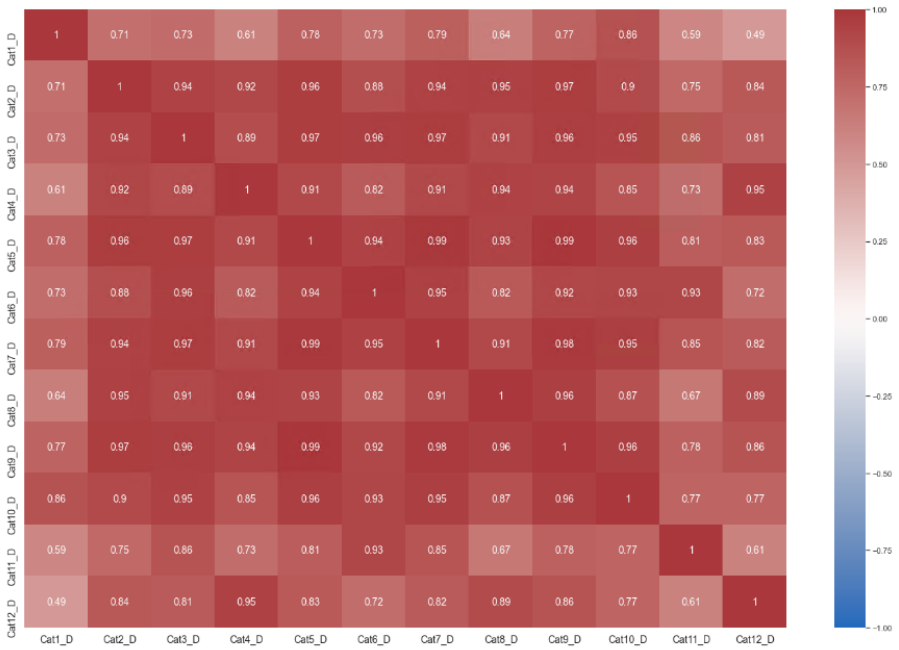

Figure 3.7 Correlation Between PoI 5Km. Categories

Due to the high correlation between the number of PoIs in different categories within 5km radii, shown in Figure 3.7, only the PoI numbers within 0.5km, 1km, and 3km radii are used. Besides, the correlation matrix of the number of PoIs in different categories within 3km radii is shown in the Figure A.1 in Appendix A. Seven more datasets are created to be used in model training process by adding different groups of PoI data into All Segments, which covers all 4 segments (Two Adults, Family, Group, Single), Two Adults Segment and Family Segment datasets. The datasets created to be used in predictive analysis and PoI data added to each of them are as follows:

- All Segments Dataset
- All Segments Dataset with 0.5km PoIs
- All Segments Dataset with 1km PoIs
- All Segments Dataset with 3km PoIs
- Two Adults Dataset
- Two Adults Dataset with 0.5km PoIs
- Two Adults Dataset with 1km PoIs
- Two Adults Dataset with 3km PoIs
- Family Segment Dataset

- Family Dataset with 0.5km PoIs

More preprocessing steps are applied to these ten datasets to prepare them to be used in the model training. These steps are explained in detail as follows:

- Independent variables of the datasets have both continuous and binary variables. The continuous variables are normalized with MinMaxScaler, which scales the data between zero and one by using minimum and maximum values with the Equation 3.5.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.5}$$

- Datasets split into train and test datasets by 70%-30%. This means that only 70% of the data are used to train the models and 30% of them are used to test the models.

3.3 Descriptive Analysis

In this section, we report some descriptive analysis results of the Company - Hotel Sales Transactions Dataset we use to create the customised dataset, and of the customised dataset itself.

Company - Hotel Sales Transactions Dataset, which consists of sales transaction records between 01.01.2015 and 01.01.2019, has 130,187 rows of data in total. Number of hotel sales transactions per year is shown in Figure 3.8. The percentage rates and number of transactions of the annual distribution of the data are as follows:

- 2015 13.9% - 18,045
- 2016 17.6% - 22,848
- 2017 23.5% - 30,553
- 2018 25.8% - 33,547
- 2019 19.4% - 25,194

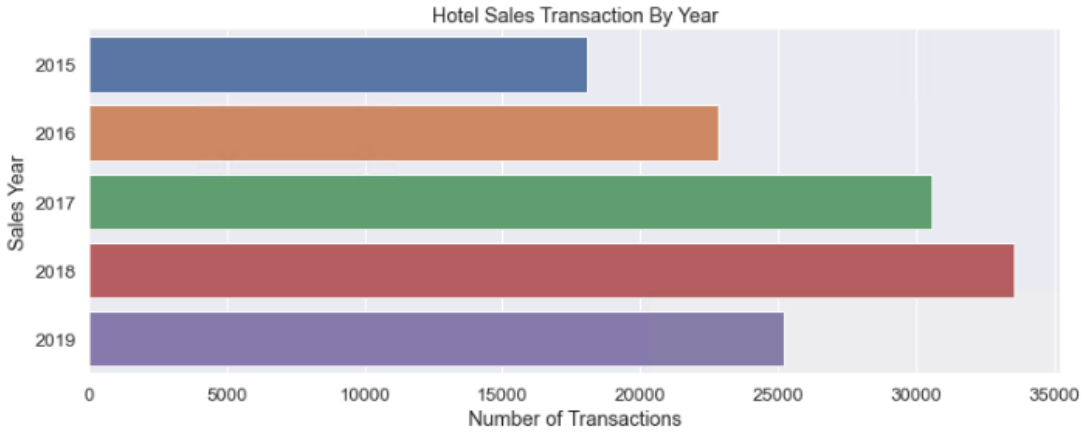

Figure 3.8 Hotel Sales Transaction by Year

There are four customer segments in the dataset. We use only 2019 data for our study, thus we report the descriptive analysis results for only 2019 data. More than half of the transactions in 2019 belong to the two adults segment, the next largest one in the dataset is the family segment with almost 30% of the data. Number of hotel sales transactions per customer segment for 2019 data is shown in Figure 3.9. The percentage rates and number of transactions of the customer segment distribution of the data are as follows:

- Two Adults 53.8% - 13,548
- Family 29.9% - 7,539
- Group 9.0% - 2,255
- Single 7.4% - 1,852

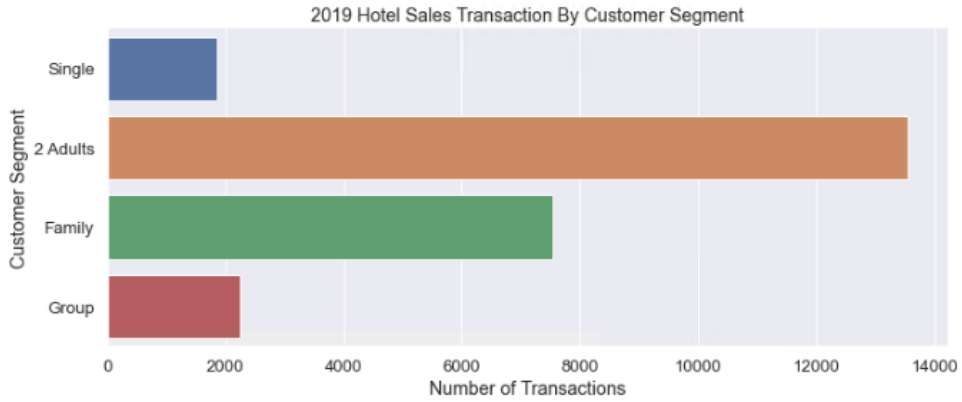

Figure 3.9 2019 Hotel Sales Transaction by Customer Segment

Dependent variable in our regression model is roomNight, which is the annual room-

night sales per hotel. It is calculate per hotel by summing up all RoomNight values for each transaction through out a year. The RoomNight value of transactions has a min value of 1, max value of 30, mean of 3.56, median of 3 with a standard deviation of 2.07. More than 80% of the transactions have less than or equal to 5 as RoomNight value. Frequency distribution of different RoomNight values for 2019 data is shown in Figure 3.10.

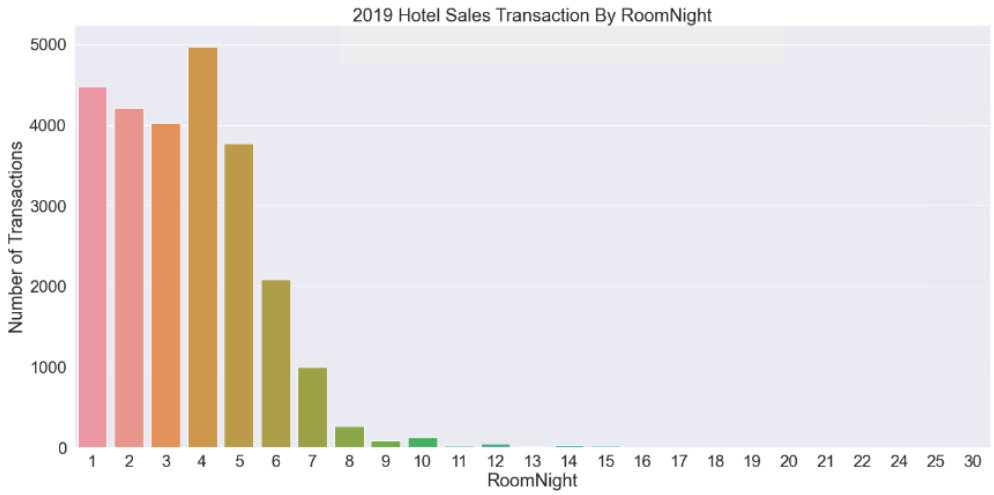

Figure 3.10 2019 Hotel Sales Transaction by RoomNight

2019 RoomNight values of transactions per customer segment are shown in Figure 3.11 and boxplots of them are shown in Figure 3.12. Top 3 values of RoomNight for two adults segment, which is the largest segment in the data, are 4, 1, and 2. Whereas, they are 4, 5, and 3 for family segment, respectively. The boxplots show that group segment has higher range of RoomNight values than other segments, in general.

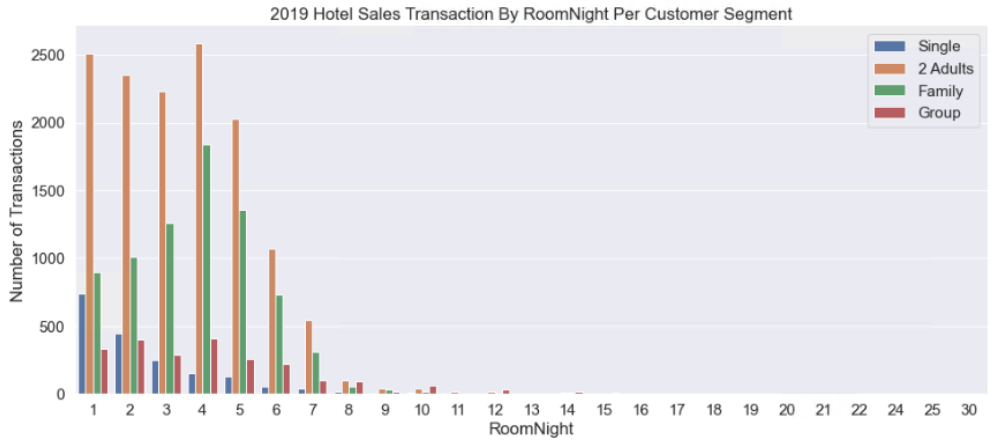

Figure 3.11 2019 Hotel Sales Transaction by RoomNight per Customer Segment

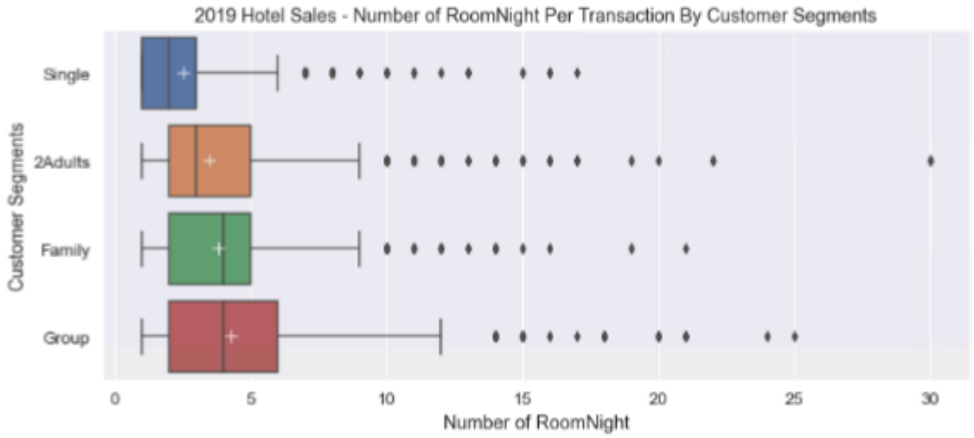

Figure 3.12 2019 Hotel Sales Number of RoomNight per Transaction by Customer Segments

One of the independent continuous variables we use in the regression models is the ListPricePerAdult, which is the room price per adult before the discount. The boxplots of 2019 ListPricePerAdult by customer segment are shown in Figure 3.13 and the scatterplot of average ListPricePerAdult vs. total sales of RoomNight in 2019 per hotel is shown in Figure 3.14. The scatterplot shows that there is no clear relationship between average list price and rooms sold.

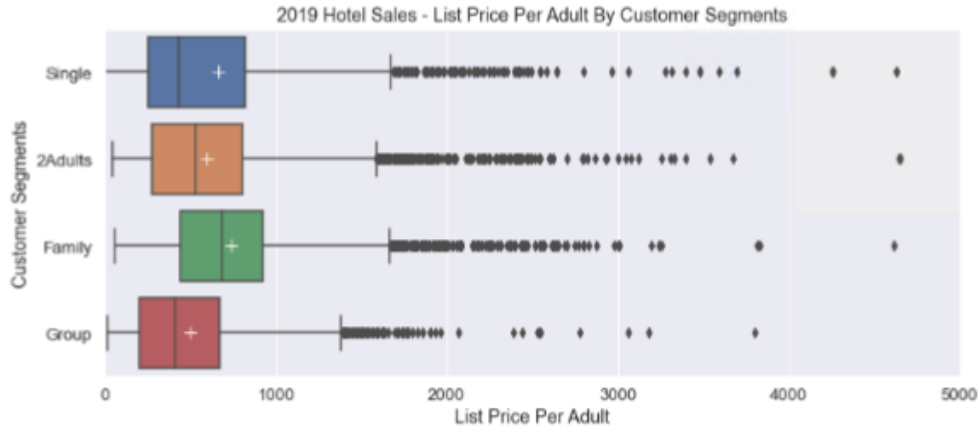

Figure 3.13 2019 Hotel Sales List Price per Adult by Customer Segments

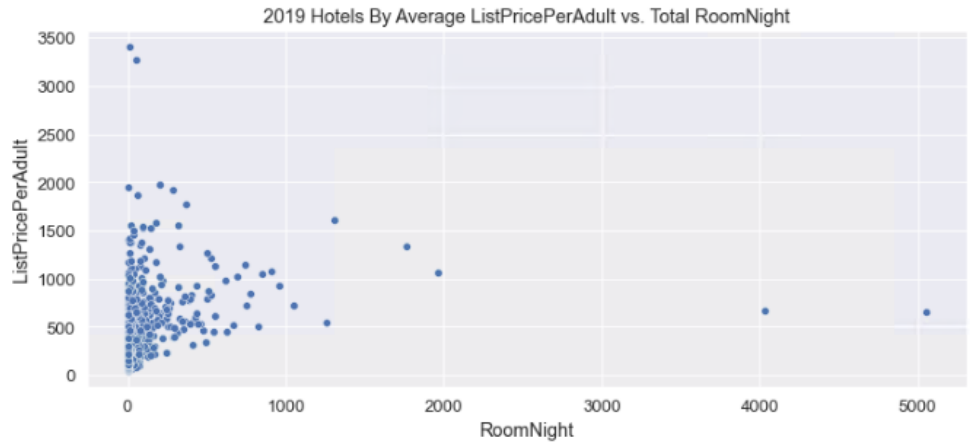

Figure 3.14 2019 Hotels by Average ListPricePerAdult vs Total RoomNight

Another independent continuous variable we use in the regression models is the DiscountRatio, which gives the ratio of the discount applied to the sales transaction on ListPricePerAdult. The boxplots of 2019 DiscountRatio per customer segment are shown in Figure 3.15. The discount ratios of family and two adults segments seem to be higher than single and group segments, in general. Moreover, the scatterplots of average DiscountRatio vs. total sales of RoomNight in 2019 per hotel and average DiscountRatio vs. ListPricePerAdult are shown in Figure 3.16 and Figure 3.17, respectively. The scatterplots show that there are no clear relationships between average discount ratio and rooms sold, and average discount ratio and average list price.

Figure 3.15 2019 Hotel Sales Discount Ratio per Transaction by Customer Segments

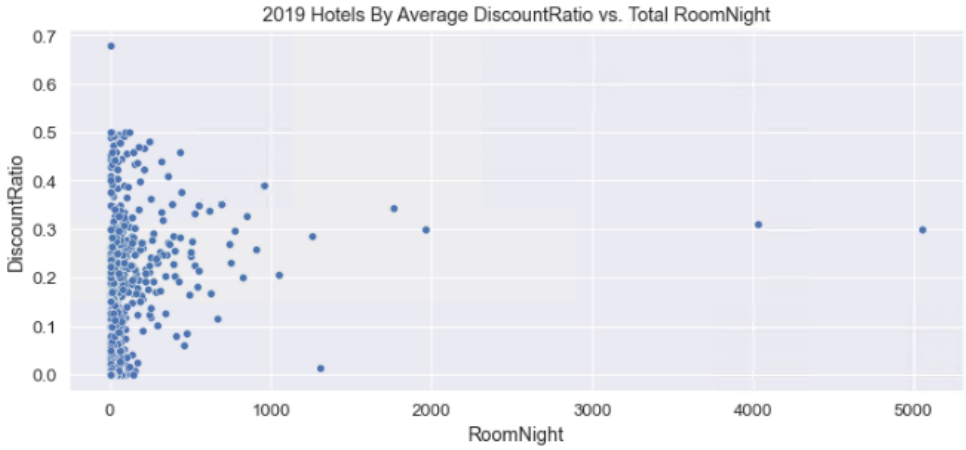

Figure 3.16 2019 Hotels by Average DiscountRatio vs Total RoomNight

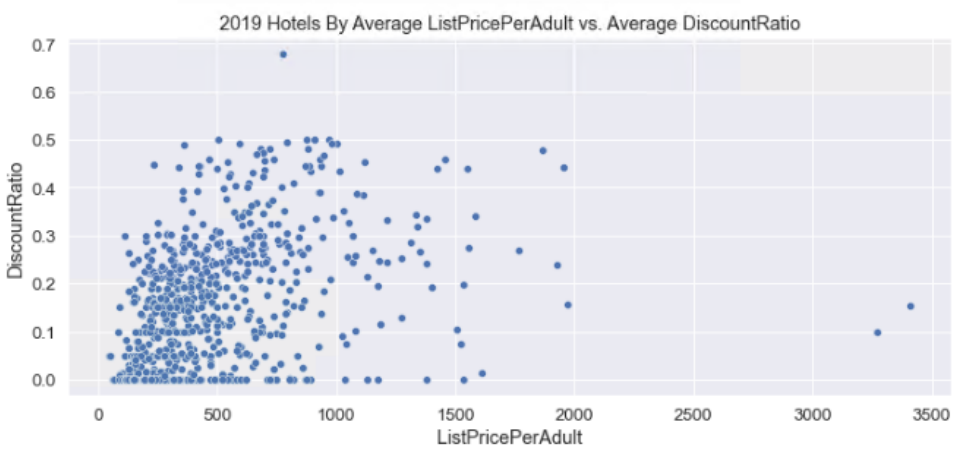

Figure 3.17 2019 Hotels by Average ListPricePerAdult vs Average DiscountRatio

There are seven different hotel categories in hotels sold in 2019. The data shows that more than 50% of the transactions belong to 5Star hotels and 2Star hotels are only a few of them. Number of hotel sales transactions per hotel category for 2019 data is shown in Figure 3.18. The percentage rates and number of transactions of the hotel category distribution of the data are as follows:

- 5Star 53.6% - 13,480
- 4Star 12.2% - 3,077
- Special 7.5% - 1,876
- Boutique 5.1% - 1,283
- 3Star 3.0% - 763
- Resort Hotel 2.8% - 706

- 2Star 0.1% - 20

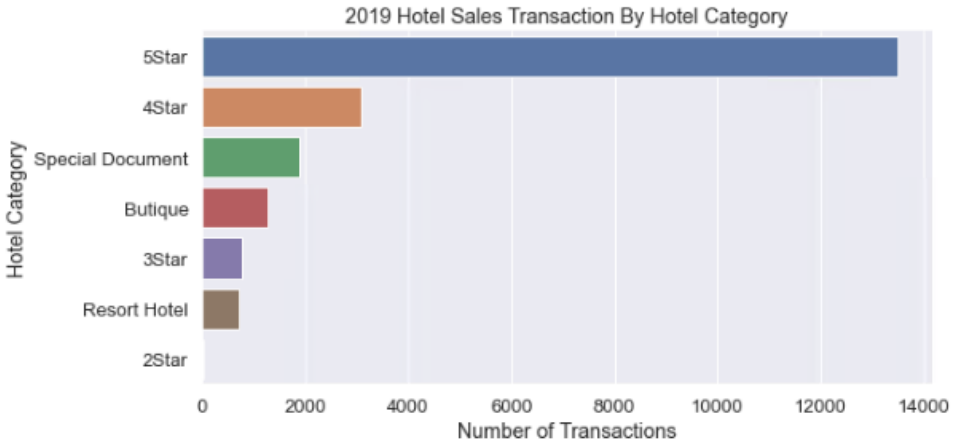

Figure 3.18 2019 Hotel Sales Transaction by Hotel Category

Although the majority of the number of transactions in 2019 belongs to 5Star hotels, Figure 3.19 shows that Resort Hotel category has the highest average RoomNight value. Moreover, Figure 3.20 shows that average RoomNight for Resort Hotels are the highest for the Single segment and almost the same as the 5Star hotels for the Group segment.

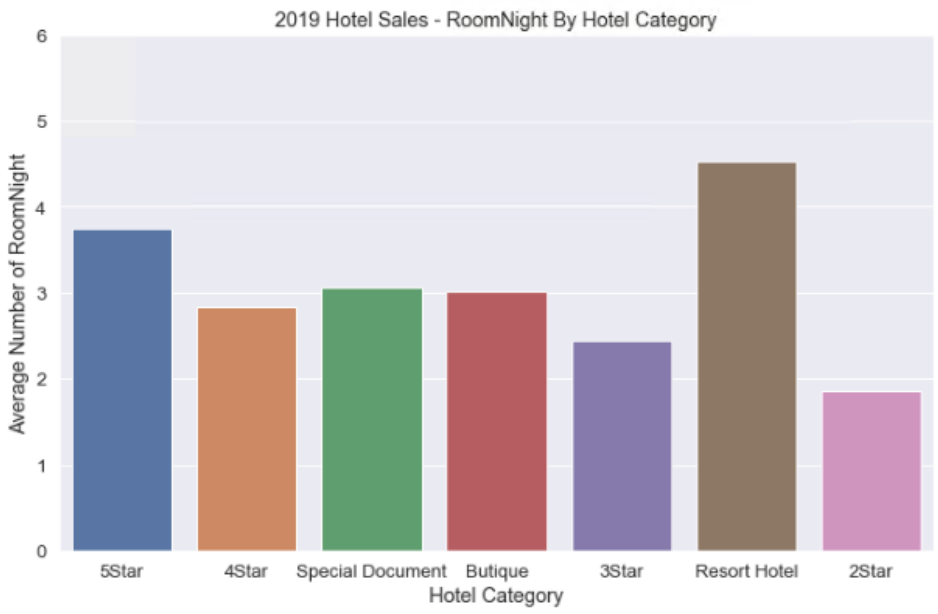

Figure 3.19 2019 Hotel Sales RoomNight by Hotel Category

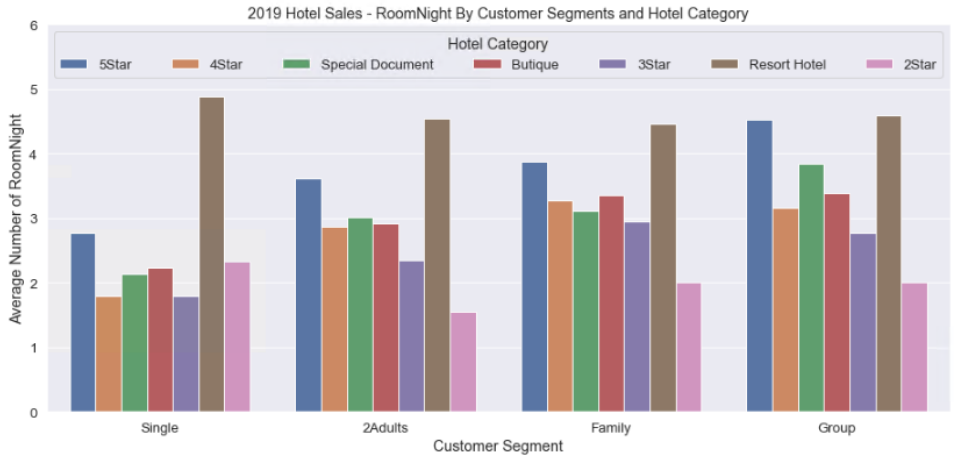

Figure 3.20 2019 Hotel Sales RoomNight by Customer Segments and Hotel Category

Additionally, there are a few binary variables in the dataset. The ratio of each group for some of the interesting ones of these variables and the average RoomNight values of each group are as follows:

- IsHolidayHotel: 57.1% of the hotels are labeled as 1 and the average annual RoomNight value of these hotels in 2019 is 104.20 whereas it is only 8.70 for the other group
- IsChainHotel 30.9% of the hotels are labeled as 1 and the average annual RoomNight value of these hotels in 2019 is 85.37 whereas it is 53.36 for the other group
- IsSingleManForbidden 19.2% of the hotels are labeled as 1 and the average annual RoomNight value of these hotels in 2019 is 190.55 whereas it is only 33.02 for the other group

Descriptive statistics of continuous variables of the constructed dataset are shown in Figure 3.21. Also, the correlation matrix of these features is shown in Figure 3.22. From the correlation matrix it can be seen that some of the hotel feature groups, such as indoor_sports, outdoor_sports, entertainment, family_holiday, summer_holiday, water_sports, are highly correlated with each other. It is interesting to see there is not any high correlation between hotel feature groups and other features such as Rating, Rating_Count, ListPricePerAdult, DiscountRatio, and RoomNight_Count.

	transportation	water_sports	accessibility	general_services	business
count	712.0	712.0	712.0	712.0	712.0
mean	1.0	2.0	0.0	4.0	0.0
std	1.0	2.0	0.0	3.0	1.0
min	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	1.0	0.0
50%	0.0	1.0	0.0	4.0	0.0
75%	1.0	2.0	0.0	6.0	1.0
max	3.0	11.0	2.0	15.0	2.0

	cafe_restaurant	entertainment	family_holiday	food_beverage	summer_holiday
count	712.0	712.0	712.0	712.0	712.0
mean	2.0	5.0	3.0	6.0	5.0
std	2.0	5.0	3.0	3.0	5.0
min	0.0	0.0	0.0	0.0	0.0
25%	1.0	0.0	0.0	4.0	0.0
50%	2.0	4.0	2.0	6.0	3.0
75%	2.0	8.0	5.0	7.0	8.0
max	10.0	22.0	15.0	15.0	24.0

	health_beauty	indoor_sports	outdoor_sports	shopping	winter_sports
count	712.0	712.0	712.0	712.0	712.0
mean	3.0	3.0	1.0	0.0	0.0
std	3.0	3.0	2.0	1.0	1.0
min	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0	0.0
50%	3.0	1.0	0.0	0.0	0.0
75%	6.0	5.0	2.0	0.0	0.0
max	18.0	14.0	9.0	5.0	4.0

	Rating	Rating_Count	ListPricePerAdult	DiscountRatio*	RoomNight_Count
count	712.0	712.0	712.0	71200.0	712.0
mean	4.0	1036.0	432.0	14.0	69.0
std	0.0	1265.0	345.0	14.0	116.0
min	2.0	2.0	59.0	0.0	6.0
25%	4.0	244.0	205.0	0.0	11.0
50%	4.0	653.0	322.0	11.0	23.0
75%	4.0	1364.0	565.0	23.0	67.0
max	5.0	16191.0	3411.0	68.0	773.0

Figure 3.21 Descriptive Statistics of Constructed Dataset

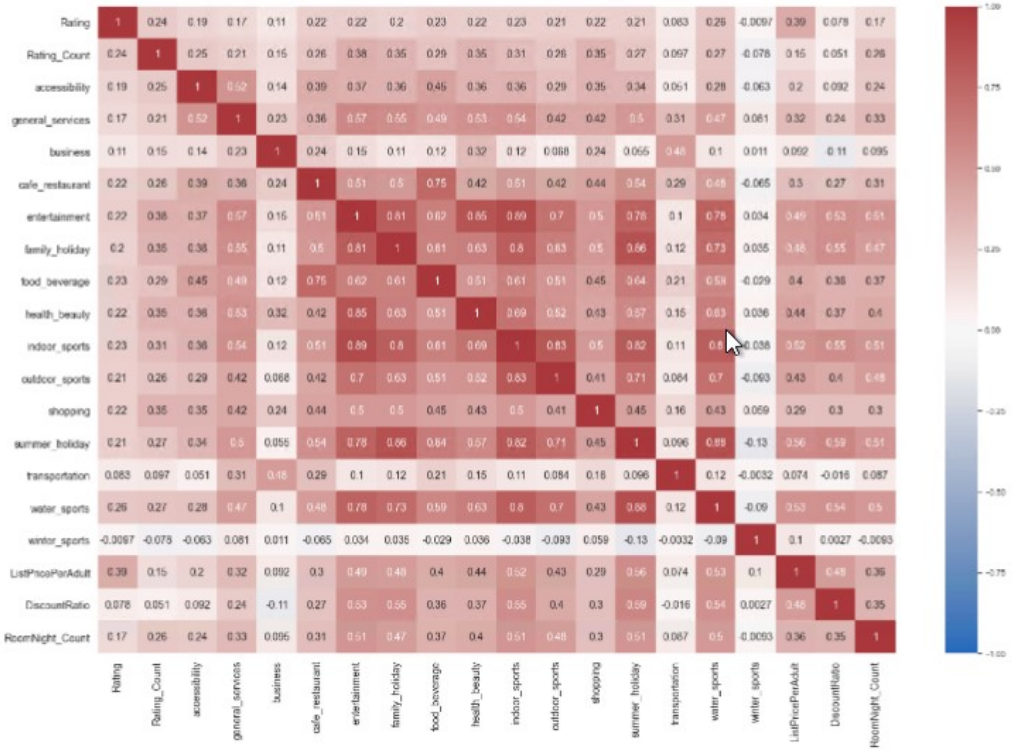

Figure 3.22 Correlation Matrix of Constructed Dataset

4. MODEL TRAINING AND CROSS VALIDATION

In this chapter, we describe the different machine learning algorithms, parameters and cross validation results of the built models.

We train the models to find the best performing one to predict annual total room sales of each hotel. When training the models, different combinations of constructed and preprocessed datasets, that are explained in detail in the previous chapter, are used. The data are randomly split into train and test datasets by using 70-30 ratio, respectively, and models are trained by using only 70% of each dataset. The number of rows in the train and test datasets are as follows:

- All Segments Train Dataset: 498
- All Segments Test Dataset: 214
- Two Adults Segment Train Dataset:391
- Two Adults Segment Test Dataset: 168
- Family Segment Train Dataset: 226
- Family Segment Test Dataset: 97

While training the models we use repeated k-fold cross validation with five folds and three repetitions. Cross validation is a statistical approach for analyzing and comparing learning algorithms that divides data into two segments: one for learning or training a model and the other for validating it (Refaeilzadeh, Tang & Liu, 2009). In repeated cross validation methodology, the cross validation procedure is repeated by the defined number of times and the average of these cross validation results is given as the final result. For our model training process, repeated cross validation is used to get a more generalized result. Moreover, five-fold cross validation is selected over ten-fold, which is considered as the most common one for data mining and machine learning (Refaeilzadeh et al., 2009), due to the size of our datasets.

Furthermore, in order to achieve the best estimators for each model, the hyper-parameters in the models are optimised to maximize the R^2, by using GridSearchCV algorithm in scikit-learn library (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot & Duchesnay, 2011).

We train models on all of ten datasets, which are described in the previous chapter, for three main segments defined as All Segments, Two Adults Segment, and Family Segment. Then we compare the results of each dataset in accordance with having PoI features or not.

We describe the models and the best results achieved from the optimized models by using the best R^2 value in the following sections. Moreover, all results achieved by the optimized models are available on Appendix A.

4.1 Linear Regression

"Regression is the study of dependence"(Weisberg, 2005). It is a statistical learning technique which aims to model the relation between a quantitative response and one or more independent factors. When there is only one independent variable x for response y it is called simple linear regression. Multiple linear regression is used when there are more than one independent variable x_p for response y.

We train the Linear Regression Model by using repeated k-fold cross validation with five folds and three repeats. To achieve the best model, we repeated running the model with all the features that have p_value lower than 0.30 by each time dropping the features with a p_value higher than or equal to 0.30, until there are no features with a p_value higher than or equal to 0.30 left in the model. The results of best models are shown in Table 4.1, Table 4.2, Table 4.3, Table 4.4 and Table 4.5.

By using Two Adults Segment Dataset rather than All Segments Dataset validation R^2 value increase from 0.06 to 0.17 and MAE decrease from 58.68 to 34.38. R^2 and MAE are not as good as Two Adults Segment's ones for Family Segment Dataset.

Table 4.1 Linear Regression Best Cross Validation Results on All Segments Without PoI, With 0.5Km PoI, With 1Km PoI and With 3Km PoI Datasets

			Best Results
Linear Regression	MAE	Train	53.75
		Validation	58.68
	R2	Train	0.35
		Validation	0.06

Table 4.2 Linear Regression Best Cross Validation Results on Two Adults Segment Without PoI, With 0.5Km PoI and With 1Km PoI Datasets

			Best Results
Linear Regression	MAE	Train	32.38
		Validation	34.38
	R2	Train	0.31
		Validation	0.17

Table 4.3 Linear Regression Best Cross Validation Results on Two Adults Segment With 3Km PoI Dataset

			Best Results
Linear Regression	MAE	Train	32.26
		Validation	34.25
	R2	Train	0.31
		Validation	0.16

Table 4.4 Linear Regression Best Cross Validation Results on Family Segment Without PoI Dataset

			Best Results
Linear Regression	MAE	Train	39.35
		Validation	41.93
	R2	Train	0.25
		Validation	0.10

Table 4.5 Linear Regression Best Cross Validation Results on Family Segment With 0.5Km PoI Dataset

			Best Results
Linear Regression	MAE	Train	39.29
		Validation	42.43
	R2	Train	0.26
		Validation	0.10

4.2 Ridge Regression

In cases where the independent variables are highly correlated, ridge regression is a method of calculating the coefficients of multiple-regression models. Ridge regression is a prominent parameter estimate approach for dealing with the collinearity issue that commonly arises in multiple linear regression (McDonald, 2009). By decreasing the parameters, the model avoids collinearity and minimizes complexity by decreasing the coefficients with the cost function.

Regularization enhances the problem's conditioning and minimizes the variation of the estimations (Pedregosa et al., 2011). In Ridge regression there is a parameter that defines the regularization strength. When training the model, we use an open-source scikit-learn library. In this machine learning library, GridSearchCV algorithm is a powerful tool for hyper-parameter tuning to find the best regularization parameter that results in the best performance of a model on a dataset. In scikit-learn GridSearchCV algorithm the hyper-parameter which defines the regularization strength is called 'alpha'. We optimize the model for R^2, with 16 different alpha values that are listed below:

#alpha=[1e-5,1e-4,1e-3,1e-2,1e-1,1,10,100,200,300,400,500,600,700,800,1000]

According to the R^2-based results of 160 different models we run, with 10 different datasets and 16 different alpha values, the best results are achieved while using alpha = 100. Using All Segments Dataset and Family Segment Without PoI Dataset give the highest validation R^2, 19%. However, MAE is the lowest for Two Adults Segment Dataset with 33.97. Moreover, including PoI data to the datasets does not have an impact on the model R^2 in neither All Segments Dataset, nor Two Adults Segment Dataset, and result is in even lower for validation R^2 in Family Segment Dataset.

The best estimators and results achieved according to the best R^2 value are shown in Table 4.6, Table 4.7, Table 4.8 and Table 4.9.

Table 4.6 Ridge Regression Best Cross Validation Results on All Segments Without PoI, With 0.5Km PoI, With 1Km PoI and With 3Km PoI Datasets

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.57	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.52
		Validation		57.77		57.76
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24
		Validation		0.19		0.19

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.88	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.64
		Validation		57.82		57.89
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24
		Validation		0.19		0.19

Table 4.7 Ridge Regression Best Cross Validation Results on Two Adults Segment Without PoI, With 0.5Km PoI, With 1Km PoI and With 3Km PoI Datasets

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	32.18	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.11
		Validation		33.97		33.98
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.30	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30
		Validation		0.18		0.18

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.13	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.11
		Validation		33.98		34.01
	R2	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30
		Validation		0.18		0.18

Table 4.8 Ridge Regression Best Cross Validation Results on Family Segment Without PoI Dataset

			Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	39.22
		Validation		41.91
	R2	Train		0.24
		Validation		0.19

Table 4.9 Ridge Regression Best Cross Validation Results on Family Segment With 0.5Km PoI Dataset

			Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	39.14
		Validation		42.03
	R2	Train		0.25
		Validation		0.10

4.3 Decision Tree Regression

Unlike other approaches that use a set of features (or bands) jointly to perform prediction in a single decision step, the Decision Tree (DT) is based on a multistage or hierarchical decision scheme or a tree like structure.

The tree is made up of a root node (which contains all data), a number of internal nodes (splits), and a number of terminal nodes (leaves). Each node in the decision tree structure makes a binary choice that divides one or more classes from the other classes. In general, processing is done by traveling down the tree until the leaf node is reached. This is referred to as a top–down strategy. (Xu, Watanachaturaporn, Varshney & Arora, 2005).

There are more hyper-parameters that are used to tune DT Regression than Ridge Regression in GridSearchCV algorithm from scikit-learn library. The hyper-parameters we optimised in DT model are as follows:

- max_features: The maximum number of features to take into account when looking for the optimum split. Sqrt means the square root of number of features and log2 means the log of number of features to the base 2.
- max_depth: The maximum depth of a tree
- min_samples_split: The minimum number of samples necessary to split an internal node
- min_samples_leaf: The minimum number of samples required to be present at a leaf node.

(Pedregosa et al., 2011).

We optimize the model for the best R^2 value with the set of hyper-parameters below:

- #max_features = [sqrt,log2]
- #max_depth = [3,5,8]
- #min_samples_split = [20,40]
- #min_samples_leaf = [10,20]

According to the R^2-based results of 240 different models we run, with 10 different datasets and 2x3x2x2 = 24 different hyper-parameter sets, using Two Adults Segment Dataset rather than All Segments Dataset decreases MAE by 38% (from 57.15 to 34.93). Furthermore, although including PoI data with the Two Adults Dataset increases R^2 from 0.12 to 0.14 and decreases MAE from 34.93 to 33.91, it does not affect results for All Segments Dataset and Family Segment Dataset.

The graph of the best estimator for All Segment without PoI Dataset is shown in Figure A.2 in Appendix A. Moreover, The best estimators and results achieved are shown in tables 4.10 through 4.15.

Table 4.10 Decision Tree Regression Best Cross Validation Results on All Segments Without PoI Dataset

			Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	52.50
		Validation		57.15
	R2	Train		0.31
		Validation		0.14

Table 4.11 Decision Tree Regression Best Cross Validation Results on All Segments With 1Km PoI Dataset

			Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	49.29
		Validation		57.61
	R2	Train		0.38
		Validation		0.11

Table 4.12 Decision Tree Regression Best Cross Validation Results on Two Adults Segment Without PoI Dataset

			Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	32.99
		Validation		34.93
	R2	Train		0.27
		Validation		0.12

Table 4.13 Decision Tree Regression Best Cross Validation Results on Two Adults Segment With PoI 1Km and With 3Km PoI Dataset

			0.5Km PoI		1Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	31.99	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	31.68
		Validation		33.91		34.07
	R2	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.32	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.33
		Validation		0.14		0.14

Table 4.14 Decision Tree Regression Best Cross Validation Results on Family Segment Without PoI Dataset

			Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40	37.17
		Validation		42.68
	R2	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40	0.29
		Validation		0.08

Table 4.15 Decision Tree Regression Best Cross Validation Results on Family Segment With 0.5Km PoI Dataset

			Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	38.46
		Validation		42.77
	R2	Train	#max_depth=3 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	0.25
		Validation		0.05

4.4 Random Forest Regression

Random Forest (RF) is an ensemble of unpruned classification or regression trees created by using bootstrap samples of the training data and random feature selection in tree induction. Prediction is made by aggregating (majority vote or averaging) the

predictions of the ensemble (Svetnik, Liaw, Tong, Culberson, Sheridan & Feuston, 2003).

RF Regression hyper-parameters that are used in GridSearchCV algorithm from scikit-learn library are as follows:

- n_estimators: The total number of trees in the forest
- max_features: The maximum number of features to take into account when looking for the optimum split.
- max_depth: The maximum depth of a tree
- min_samples_split: The minimum number of samples necessary to split an internal node
- min_samples_leaf: The minimum number of samples required to be present at a leaf node.

(Pedregosa et al., 2011).

We optimize the model for the best R^2 value with the set of hyper-parameters below:

- #n_estimators = [1000]
- #max_features = [sqrt,log2]
- #max_depth = [3,5,8]
- #min_samples_split = [20,40]
- #min_samples_leaf = [10,20]

According to the R^2-based results of 240 different models we run, with 10 different datasets and 2x3x2x2 = 24 different hyper-parameter sets, using Two Adults Segment Dataset rather than All Segments Dataset decreases MAE by 39% (from 54.24 to 32.92). However, there is no significant impact on MAE values in neither All Segments Dataset, nor Two Adults or Family Segment datasets.

The best estimators and results achieved are shown in tables 4.16 through 4.21.

Table 4.16 Random Forest Regression Best Cross Validation Results on All Segments Without PoI Dataset

			Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20 #n_estimators=1000	51.29
		Validation		54.24
	R2	Train		0.34
		Validation		0.23

Table 4.17 Random Forest Regression Best Cross Validation Results on All Segments With PoI 0.5Km and With 1Km PoI Dataset

			0.5Km PoI		1Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	49.41	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	49.16
		Validation		53.81		53.73
	R2	Train		0.38		0.38
		Validation		0.23		0.23

Table 4.18 Random Forest Regression Best Cross Validation Results on Two Adults Without PoI Segment Dataset

			Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	29.41
		Validation		32.92
	R2	Train		0.39
		Validation		0.21

Table 4.19 Random Forest Regression Best Cross Validation Results on Two Adults Segment With PoI 0.5Km Dataset

			Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	29.07
		Validation		32.95
	R2	Train		0.40
		Validation		0.21

Table 4.20 Random Forest Regression Best Cross Validation Results on Family Segment Without PoI Dataset

			Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	37.93
		Validation		40.75
	R2	Train		0.27
		Validation		0.14

Table 4.21 Random Forest Regression Best Cross Validation Results on Family Segment With 0.5Km PoI Dataset

			Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	37.68
		Validation		40.76
	R2	Train		0.29
		Validation		0.15

4.5 Support Vector Regression

A Support Vector Machine (SVM) is a kind of supervised learning method used for classification and regression. This is a tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding overfit to the data (Jakkula, 2006).

There are three hyper-parameters that are used for SVM Regression in GridSearchCV algorithm from scikit-learn library. They are as follows:

- kernel: Indicates the kind of kernel, which is the core function that takes low dimensional input space and transforms it into a higher-dimensional space to be used in the algorithm. Kernel can be Linear, Polynomial, and RBF (Radial Basis Function).
- C: The regularization parameter which is inversely proportional to strength, that specifies how much error can be tolerated, it helps to control over-fitting.
- gamma: The kernel coefficient

(Pedregosa et al., 2011).

We optimize the model for the maximum R^2 value with the set of hyper-parameters below:

- #kernel = [linear,poly,rbf]
- #C = [100,10,1.0,0.1,0.01]
- #gamma = [scale,auto]

According to the R^2-based results of 300 different models we run, with 10 different datasets and 3x5x2 = 30 different hyper-parameter sets, using Two Adults Segment Dataset rather than All Segments Dataset improves MAE value 40% (from 50.23 to 29.89) and has a slight improvement in R^2 with 0.02 points. While including PoI data to All Segments datasets improves R^2 34% (from 0.17 to 0.26), there is no significant impact on R^2 or MAE values in Two Adults or Family Segment datasets.

The best estimators and results achieved with them are shown in tables 4.22 through 4.27.

Table 4.22 Support Vector Regression Best Cross Validation Results on All Segments Without PoI Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	37.08
		Validation		50.23
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.34
		Validation		0.17

Table 4.23 Support Vector Regression Best Cross Validation Results on All Segments With PoI 1Km Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	37.69
		Validation		50.71
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.34
		Validation		0.26

Table 4.24 Support Vector Regression Best Cross Validation Results on Two Adults Segment Without PoI Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=10 #gamma=scale #kernel=poly	22.19
		Validation		29.89
	R2	Train	#c=10 #gamma=scale #kernel=poly	0.39
		Validation		0.19

Table 4.25 Support Vector Regression Best Cross Validation Results on Two Adults Segment With PoI 0.5Km Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=poly	14.90
		Validation		31.24
	R2	Train	#c=100 #gamma=scale #kernel=poly	0.57
		Validation		0.20

Table 4.26 Support Vector Regression Best Cross Validation Results on Family Segment Without PoI Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	24.27
		Validation		39.01
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.35
		Validation		0.03

Table 4.27 Support Vector Regression Best Cross Validation Results on Family Segment With 0.5Km PoI Dataset

			Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	24.80
		Validation		38.73
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.33
		Validation		0.04

5. RESULTS

Our study reveals a comprehensive understanding of the capabilities of different regression models in hotel room sales prediction. In this section, we evaluate the dataset - model harmony and the models from the perspectives of goodness of fit (R^2) and accuracy (MAE).

The models are compared according to test results achieved by using without PoI and with PoI data, and the best models for All Segments, Two Adults Segment and Family Segment datasets are presented in tables 5.1 through 5.3.

Additionally we calculate the feature importance scores by using the impurity-based feature importance of random forests, which uses the mean of impurity decrease accumulation inside each tree, for all datasets. For All Segments Dataset the importance of the features that contribute to the regression models are shown in Figure 5.1, Figure 5.2, Figure 5.3 and Figure 5.4. The top five important features are common for both without and with PoI datasets, even though the order is different. These are summer_holiday, entertainment, indoor_sports, outdoor_sports and water_sports. Although the best R^2 values, without PoI=0.36, 0.5Km PoI=0.39, 1Km PoI=0.40 and 3Km PoI=0.39, are achieved by Random Forest Regression (RF Regression) models as shown in Figure 5.5, MAE values of these models, without PoI=53.92, 0.5Km PoI=52.90, 1Km PoI=52.73 and 3Km PoI=53.07, are not the lowest ones as shown in Table 5.1. The lowest MAE values are achieved by Support Vector Regression (SV Regression) models. Although adding PoI data does not change R^2 or decrease MAE for Linear Regression, Ridge Regression and SV Regression, it causes increases in R^2 and decreases in MAE for Decision Tree Regression (DT Regression) and RF Regression models. Adding PoI data to RF Regression model causes an increase in R^2 from 0.36 to up to 0.40.

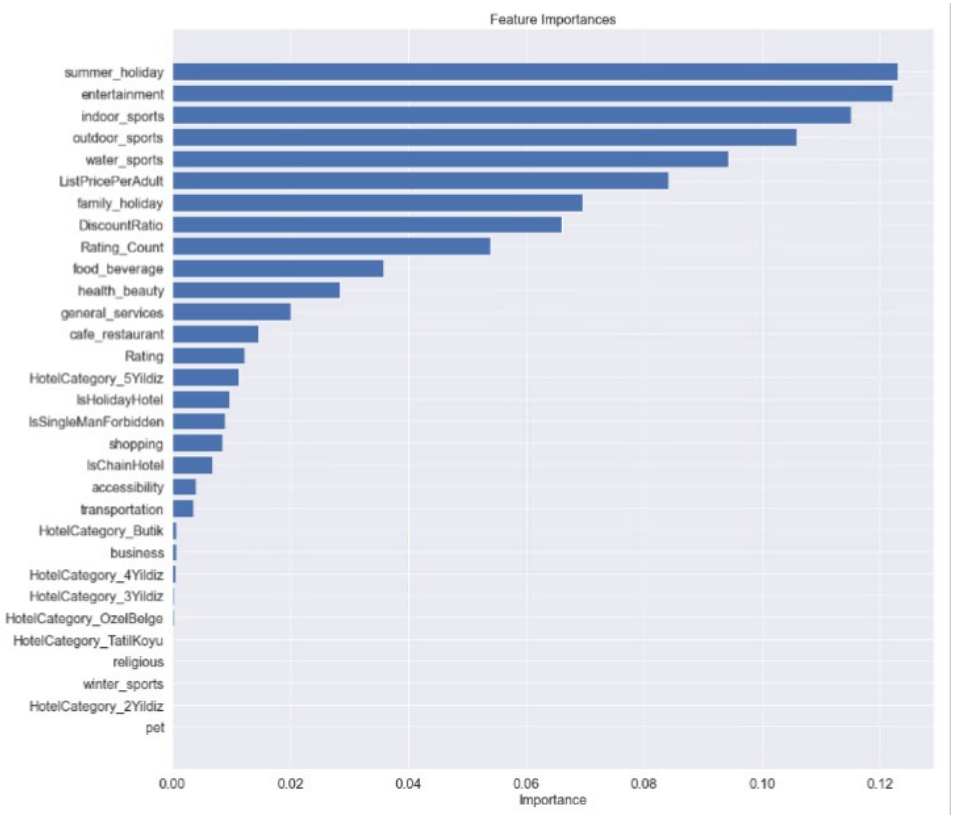

Figure 5.1 All Segments Without PoI Dataset Feature Importance

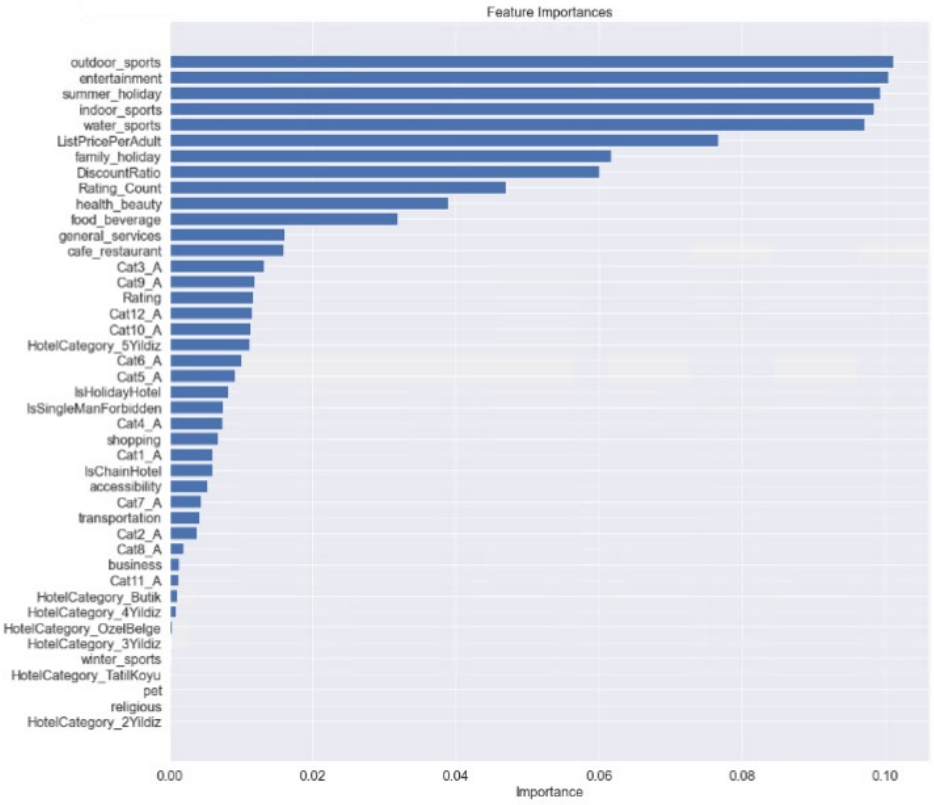

Figure 5.2 All Segments 0.5Km PoI Dataset Feature Importance

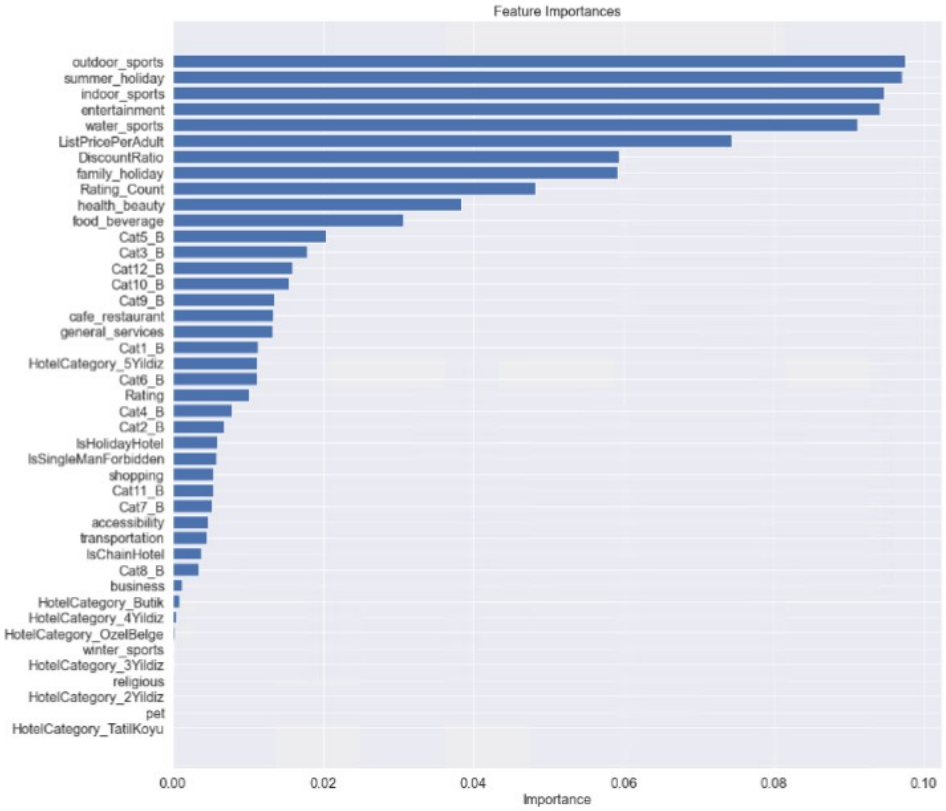

Figure 5.3 All Segments 1Km PoI Dataset Feature Importance

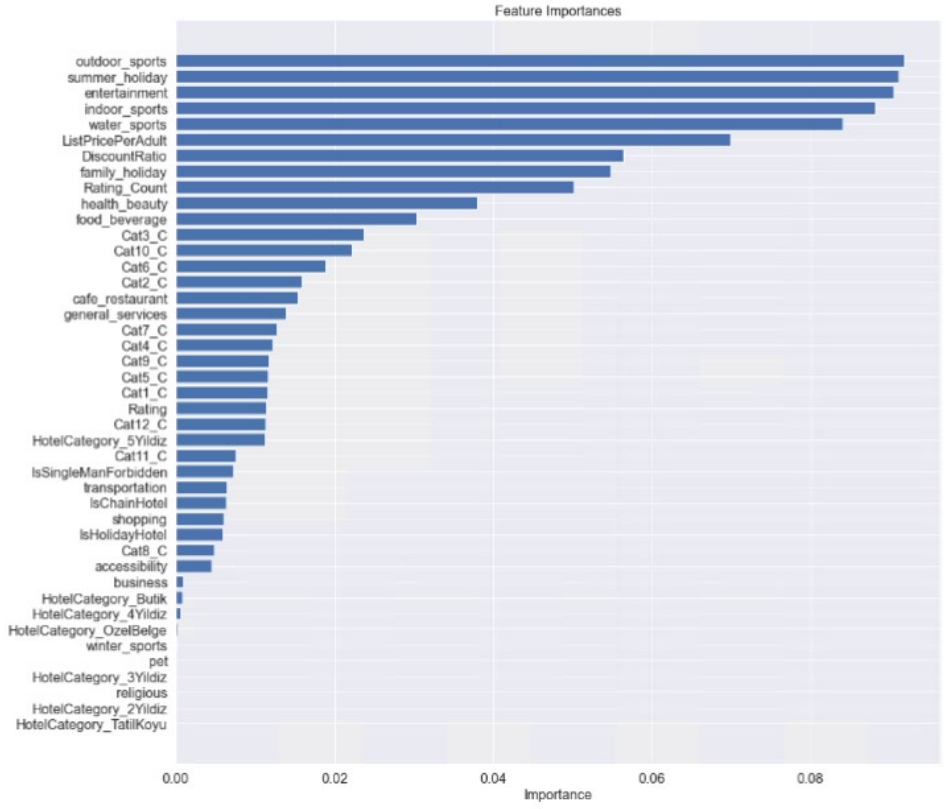

Figure 5.4 All Segments 3Km PoI Dataset Feature Importance

Table 5.1 All Segments Dataset Test Results Comparison With Different Models

			All Segments			
			Without PoI	0.5Km PoI	1Km PoI	3Km PoI
Linear Regression	MAE	Train	53.85	53.85	53.85	53.85
		Test	59.08	59.08	59.08	59.08
	R2	Train	0.34	0.34	0.34	0.34
		Test	0.31	0.31	0.31	0.31
	MSE	Train	9,034.39	9,034.39	9,034.39	9,034.39
		Test	8,740.90	8,740.90	8,740.90	8,740.90
Ridge Regression	MAE	Train	56.11	56.16	56.24	56.32
		Test	56.83	56.97	57.09	57.00
	R2	Train	0.25	0.25	0.25	0.25
		Test	0.31	0.31	0.31	0.31
	MSE	Train	10,273.63	10,263.57	10,261.72	10,273.86
		Test	8,781.35	8,779.18	8,781.25	8,787.54
DT Regression	MAE	Train	52.86	49.25	48.16	51.21
		Test	58.29	55.20	57.50	55.00
	R2	Train	0.31	0.40	0.38	0.35
		Test	0.29	0.29	0.30	0.26
	MSE	Train	9,420.65	8,263.20	8,476.51	8,929.59
		Test	9,101.99	8,991.37	8,943.00	9,444.91
RF Regression	MAE	Train	50.94	48.91	48.58	47.09
		Test	53.92	52.90	52.73	53.07
	R2	Train	0.35	0.39	0.40	0.44
		Test	0.36	0.39	0.40	0.39
	MSE	Train	8,952.76	8,341.79	8,256.56	7,635.48
		Test	8,120.51	7,795.08	7,679.01	7,771.34
SV Regression	MAE	Train	37.70	37.69	37.69	37.73
		Test	50.49	50.48	50.71	50.88
	R2	Train	0.35	0.34	0.34	0.34
		Test	0.26	0.26	0.26	0.26
	MSE	Train	8,950.63	9,073.04	9,050.96	9,058.71
		Test	9,484.21	9,474.54	9,456.55	9,458.01

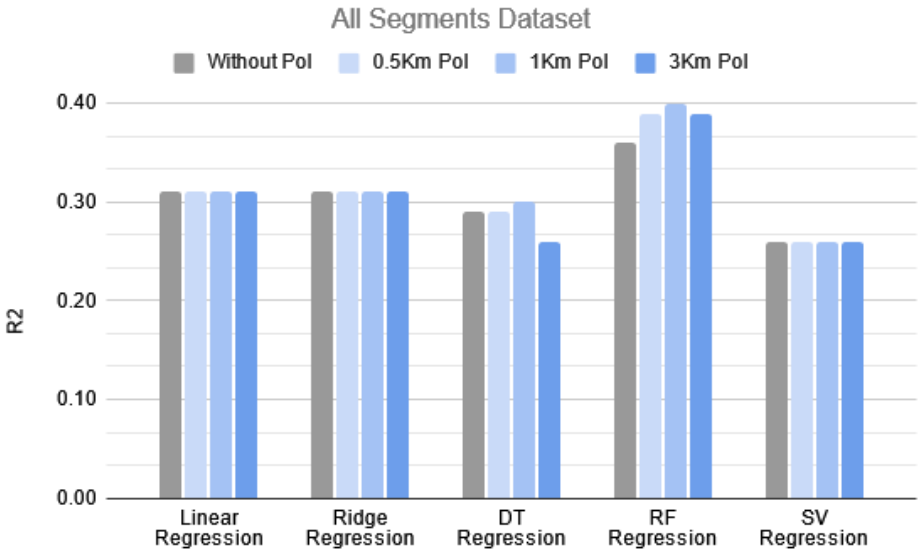

Figure 5.5 All Segments Dataset R^2 Comparison

For Two Adults Segment Dataset the features that contribute the most to the regression models are slightly different from All Segments Dataset. While entertainment, indoor_sports and outdoor_sports are in top five important features like in All Segments Dataset, ListPricePerAdult and Rating_Count features are added to the top five list as shown in Figure 5.6, Figure 5.7, Figure 5.8 and Figure 5.9. The features appear in top five do not change, except their ranking, by including PoI data to Two Adults Segment Dataset. As in All Segments Dataset, the best R^2 values, without PoI = 0.34, 0.5Km PoI=0.35, 1Km PoI=0.34 and 3Km PoI=0.34, are achieved by Random Forest Regression as shown in Figure 5.10. Although the Random Forest Regression models' R^2 values for Two Adults Segment datasets are not as high as the ones for All Segments datasets, MAE values of these models, without PoI = 30.82, 0.5Km PoI=30.90, 1Km PoI=30.96 and 3Km PoI=31.05, are around 40% lower than All Segments Dataset ones, which indicate higher accuracy levels as shown in Table 5.2. Adding PoI data to the models does not have a significant impact on the results except Support Vector Regression. Adding PoI data to the models does not change the R^2 or MAE for Linear Regression, Ridge Regression and RF Regression much but it causes a decrease in MAE for DT Regression and an increase in R^2 from 0.15 to up to 0.22 for SV Regression. However, adding PoI data in SV regression also cause increases in MAE values.

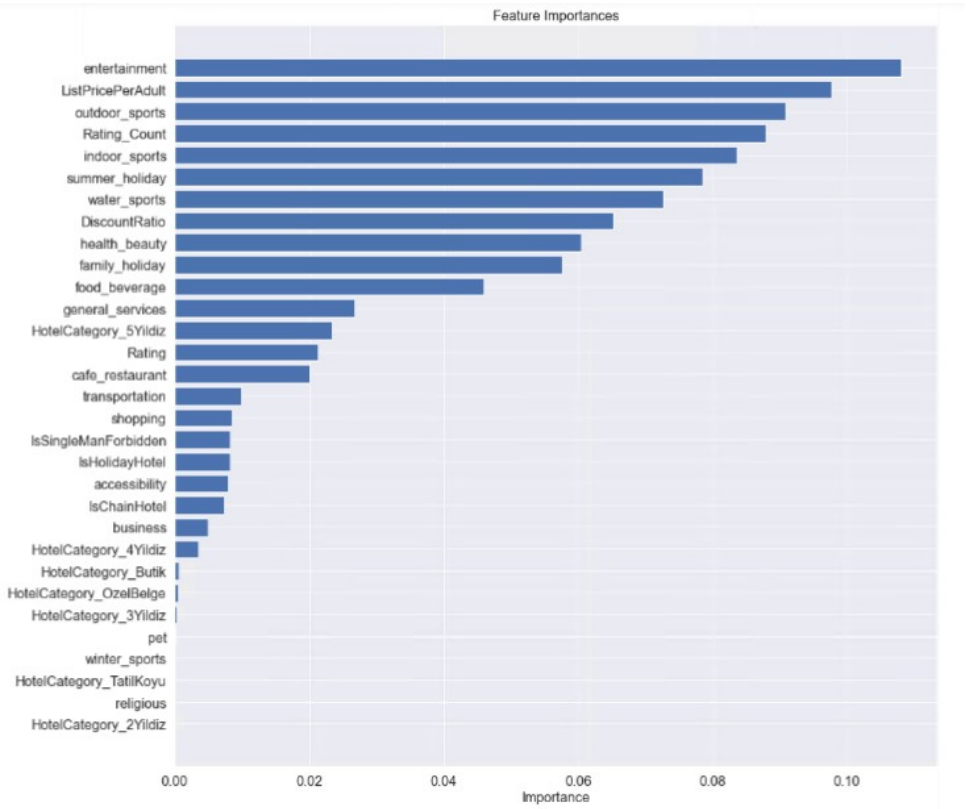

Figure 5.6 Two Adults Segment Without PoI Dataset Feature Importance

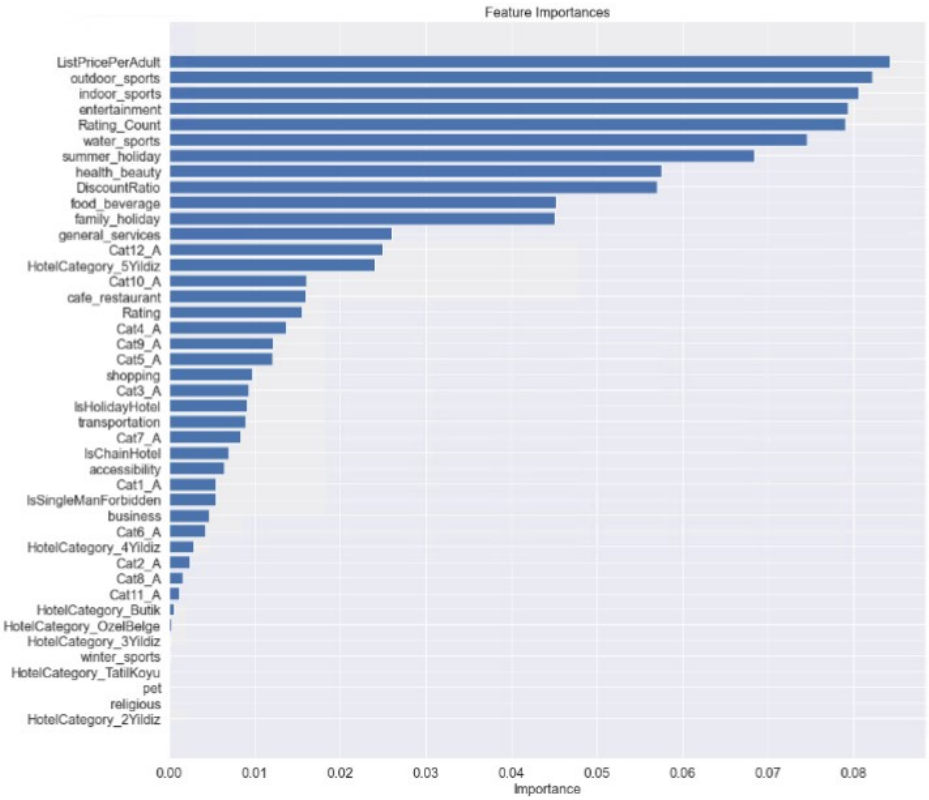

Figure 5.7 Two Adults Segment 0.5Km PoI Dataset Feature Importance

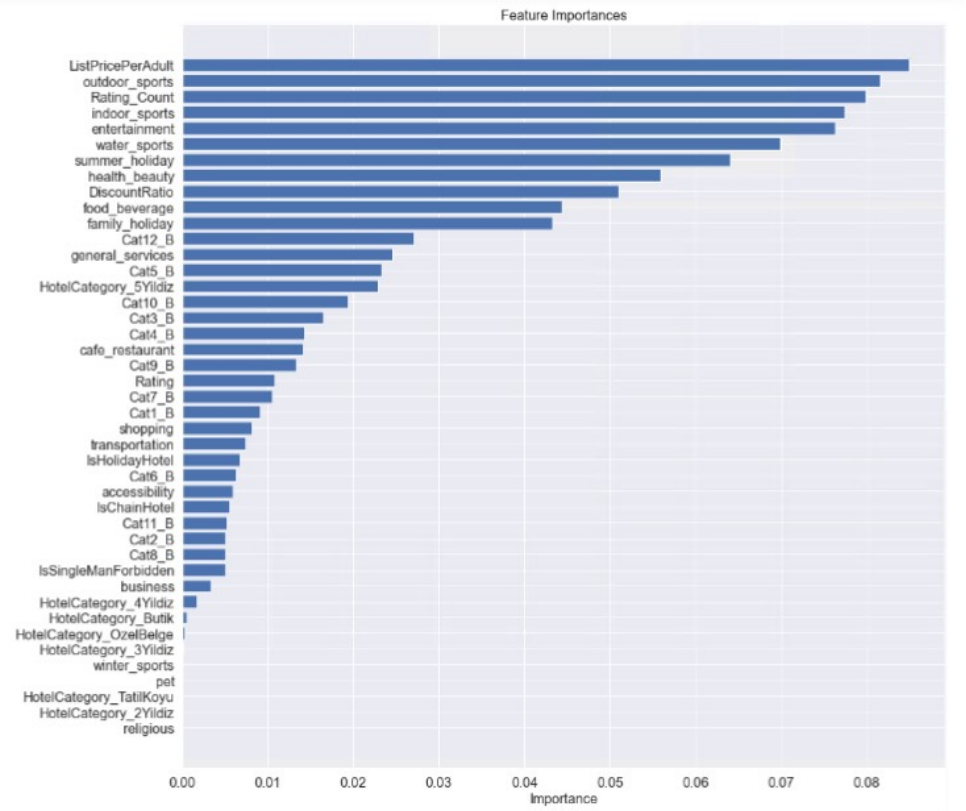

Figure 5.8 Two Adults Segment 1Km PoI Dataset Feature Importance

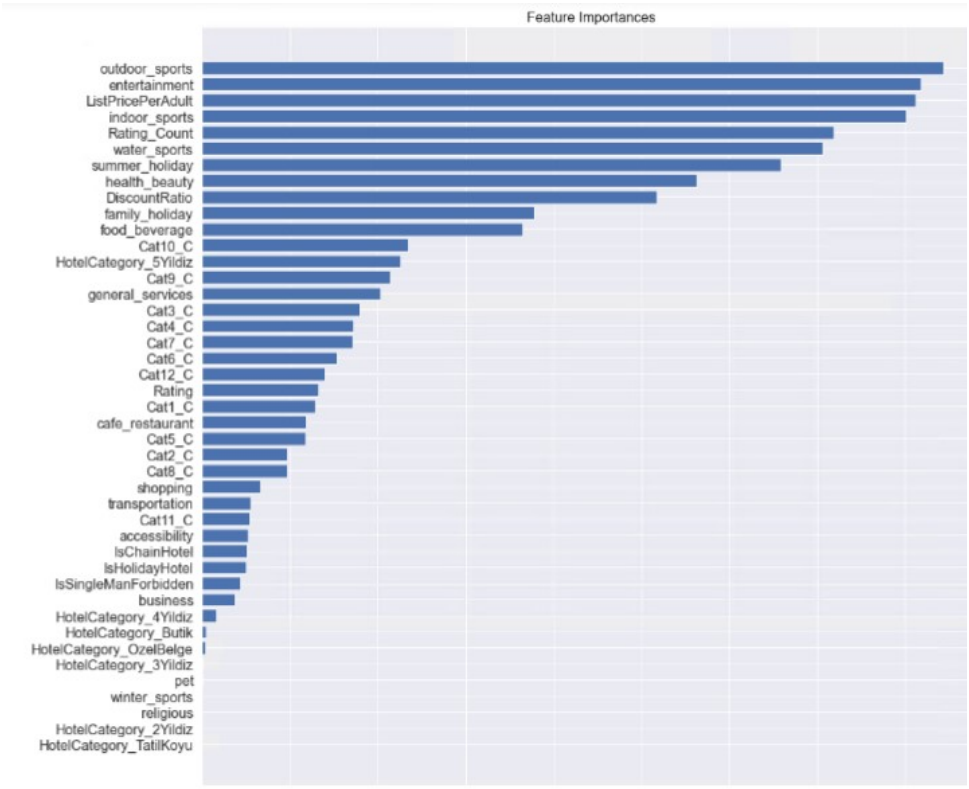

Figure 5.9 Two Adults Segment 3Km PoI Dataset Feature Importance

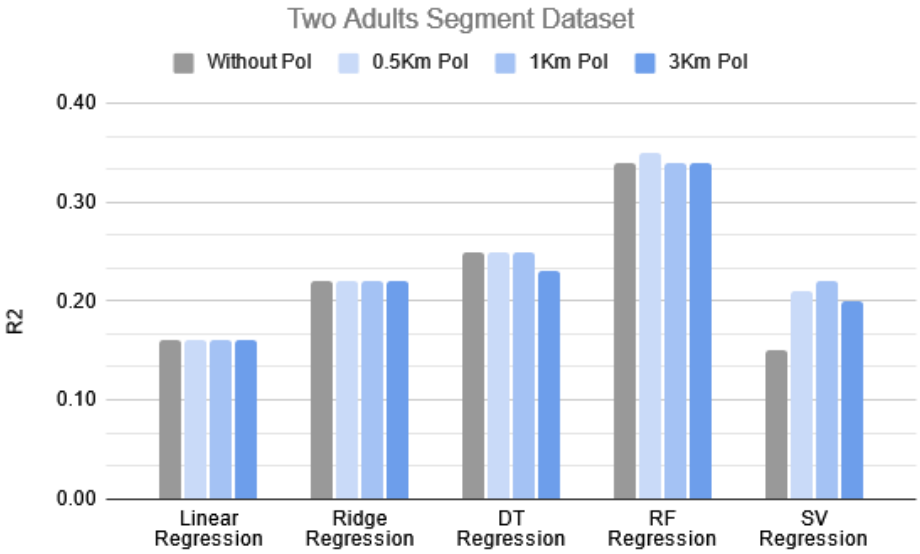

Figure 5.10 Two Adults Segment Dataset R^2 Comparison

Table 5.2 Two Adults Segment Dataset Test Results Comparison With Different Models

			Two Adults Segment			
			Without PoI	0.5Km PoI	1Km PoI	3Km PoI
Linear Regression	MAE	Train	32.53	32.53	32.53	32.38
		Test	34.92	34.92	34.92	34.72
	R2	Train	0.30	0.30	0.30	0.30
		Test	0.16	0.16	0.16	0.16
	MSE	Train	2,798.91	2,798.91	2,798.91	2,812.36
		Test	3,161.27	3,161.27	3,161.27	3,150.39
Ridge Regression	MAE	Train	32.27	32.20	32.53	32.22
		Test	34.43	34.41	34.98	34.51
	R2	Train	0.30	0.30	0.30	0.30
		Test	0.22	0.22	0.22	0.22
	MSE	Train	2,815.00	2,810.29	2,811.01	2,811.20
		Test	2,938.00	2,934.27	2,932.50	2,938.28
DT Regression	MAE	Train	34.11	31.77	31.76	31.34
		Test	32.83	30.70	30.72	31.84
	R2	Train	0.25	0.34	0.34	0.36
		Test	0.25	0.25	0.25	0.23
	MSE	Train	2,984.34	2,641.84	2,627.37	2,583.08
		Test	2,825.27	2,805.41	2,805.49	2,880.65
RF Regression	MAE	Train	29.12	29.08	28.65	27.94
		Test	30.82	30.90	30.96	31.05
	R2	Train	0.40	0.41	0.42	0.45
		Test	0.34	0.35	0.34	0.34
	MSE	Train	2,386.49	1,372.26	2,313.93	2,221.66
		Test	2,498.86	2,463.22	2,475.89	2,496.83
SV Regression	MAE	Train	22.46	15.74	15.70	15.81
		Test	30.41	31.30	31.35	31.48
	R2	Train	0.39	0.55	0.55	0.55
		Test	0.15	0.21	0.22	0.20
	MSE	Train	2,459.56	1,803.97	1,785.92	1,792.52
		Test	3,186.15	2,957.36	2,944.92	3,025.77

For Family Segment Dataset the features that contribute more to the regression models are shown in Figure 5.11 and Figure 5.12. The top five important features for without PoI are summer_holiday, ListPricePerAdult, Rating_Count, outdoor_sports and indoor_sports and for 0.5KM PoI are almost the same with different order as ListPricePerAdult, Rating_Count, summer_holiday, water_sports and indoor_sports. The only difference in top five important features of with 0.5Km PoI dataset from without PoI dataset ones is having water_sports feature rather than outdoor_sports. The best R^2 for without PoI dataset is achieved both with Ridge Regression and Random Forest Regression models with the value of 0.19, as shown in Figure 5.13. The best R^2 for with 0.5Km PoI dataset is achieved with Random Forest Regression model with the value of 0.20. Family Segment Dataset has the weakest goodness of fit in terms of R^2 values as shown in Table 5.3.

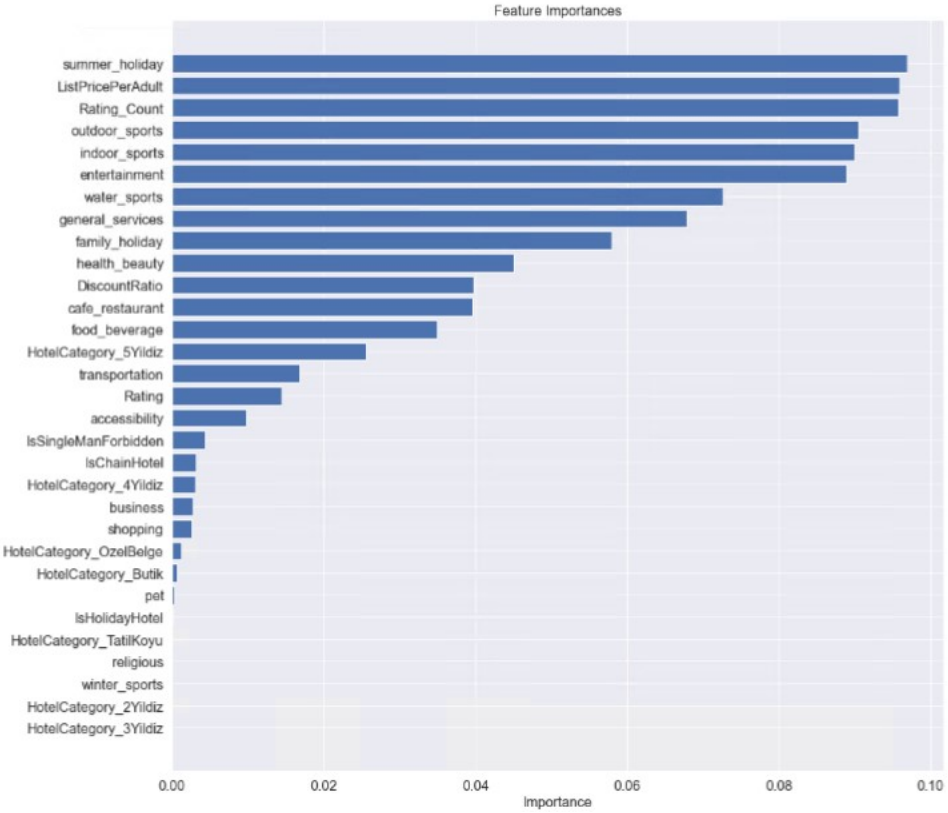

Figure 5.11 Family Segment Without PoI Dataset Feature Importance

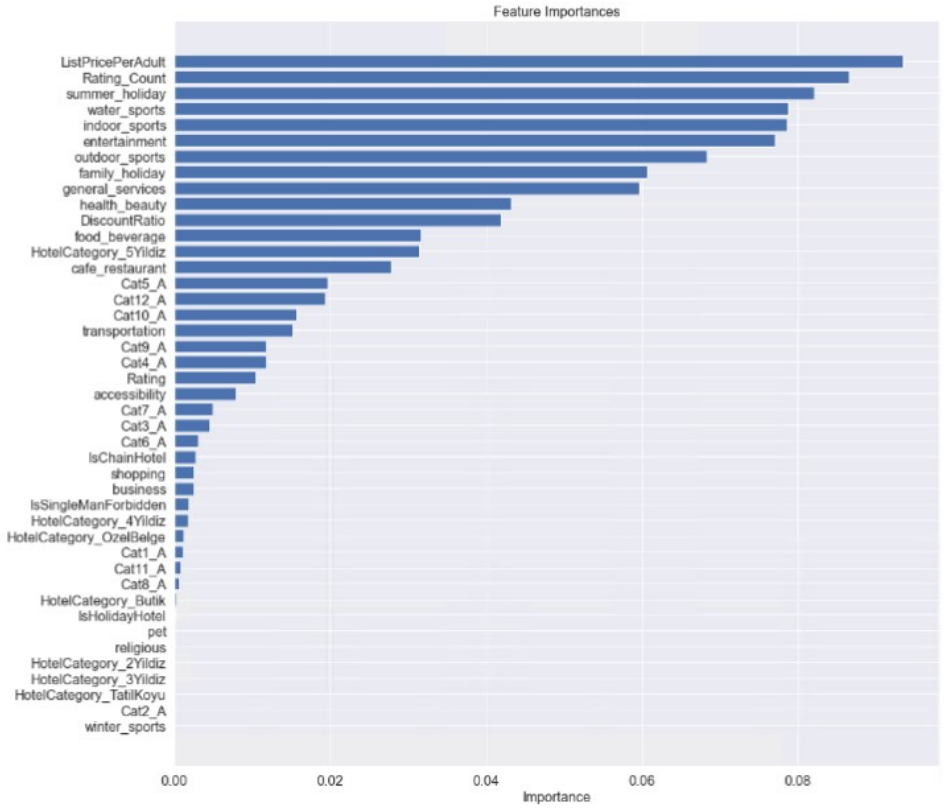

Figure 5.12 Family Segment 0.5Km PoI Dataset Feature Importance

Table 5.3 Family Segment Dataset Test Results Comparison With Different Models

			Family Segment	
			Without PoI	0.5Km PoI
Linear Regression	MAE	Train	39.54	39.46
		Test	43.53	43.76
	R2	Train	0.24	0.25
		Test	0.12	0.11
	MSE	Train	3,876.33	3,797.41
		Test	4,212.07	4,277.49
Ridge Regression	MAE	Train	39.22	39.24
		Test	41.91	41.51
	R2	Train	0.24	0.24
		Test	0.19	0.19
	MSE	Train	3,879.42	3,857.28
		Test	3,899.32	3,874.77
DT Regression	MAE	Train	37.17	36.98
		Test	42.68	46.25
	R2	Train	0.27	0.29
		Test	0.08	0.09
	MSE	Train	3,696.67	3,611.95
		Test	4,401.47	5,236.86
RF Regression	MAE	Train	37.72	37.32
		Test	42.46	41.84
	R2	Train	0.28	0.30
		Test	0.19	0.20
	MSE	Train	3,665.65	3,567.60
		Test	3,900.88	3,855.24
SV Regression	MAE	Train	24.98	25.47
		Test	40.32	39.86
	R2	Train	0.34	0.32
		Test	0.06	0.06
	MSE	Train	3,381.70	3,447.72
		Test	4,493.12	4,522.46

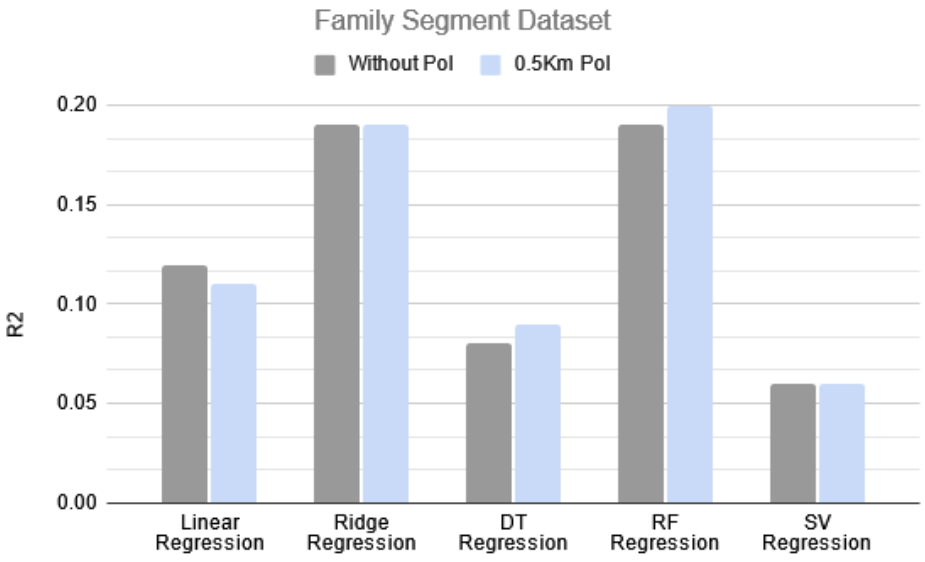

Figure 5.13 Family Segment Dataset R^2 Comparison

Taking everything into account, the results show that Random Forest Regression is the outstanding one with the highest goodness of fit and Support Vector Regression is good at accuracy values in the majority of the cases. Besides, there is a difference between the predictive performances of the models that are run by using All Segments and Two Adults Segment datasets. The MAE value decreases around 40% when the models are run by using Two Adults Segment datasets rather than All Segments datasets. However, R^2 also decreases between 6-15% in Random Forest Regression model, which is the best model according to R^2, and that means a deterioration in the goodness of fit. By comparing the results with and without PoI datasets, it can be seen that the results with PoI datasets are also as good as the results without PoI datasets and they are even better for Random Forest Regression and Decision Tree Regression models of All Segments datasets.

Moreover, the results indicate that, the top five important features, which are outdoor_sports, summer_holiday, entertainment, indoor_sports, water_sports for All Segments datasets, entertainment, ListPricePerAdult, outdoor_sports, Rating_Count, indoor_sports for Two Adults Segment datasets are common for both without and with PoI datasets, even though the order changes. Although PoI related features are not listed in top five important features list of any dataset, list_price and rating_count features appear to be more important in segment specific datasets, namely Two Adults and Family, than All Segments Dataset.

6. CONCLUSION

This study is based on the real-life data provided by a Turkish travel agency and gives us a great opportunity to explore the application of regression models supported by machine learning in annual hotel room sales prediction by looking at previous sales transactions of hotels with different features. Although using real-life data is great to have a grasp of basics and dynamics of the industry, the study requires a staging data correction process. Additionally, despite all efforts, the constructed datasets still include both explained and unexplained abnormalities because of specific business decisions, human errors or imprecise data collection.

The models are built to test the basic types of regression algorithms and compare their results in terms of goodness of fit and accuracy. The results show that the machine learning models improve the performance of prediction comparing to the multivariate linear regression. Moreover, the MAE value is less for Two Adults Segment datasets comparing to All Segments datasets, but also R^2 decreases for Random Forest Regression model for instance. The reason of declining MAE while using a specific segment dataset rather than using All Segments Dataset might be because of the fact that the customers in a specific segment group have a tendency to resemble each other in terms of their purchasing behavior.

In general, the top five important features are common, except their order, for both without and with PoI datasets. However, the importance of price and Rating_Count features increases for segment specific datasets.

Although we receive slightly different results for each dataset which are constructed according to consumer segments and PoI distances, Random Forest Regression model is the outstanding one with the highest goodness of fit in majority of the cases. When splitting nodes, the Random Forest Algorithm looks for the best feature from a random subset of features rather than the most essential one. Aggregating these predictions makes the Random Forest Regression model a powerful tool for forecasting. Besides, Support Vector Regression model is good at accuracy values in majority of the cases in this study.

On the other hand, the Random Forest Regression and the likes of machine learning models require more attention in the application process. Controlling the learning process, namely the hyper-parameter tuning, is extremely important in order to optimize the computational cost and the performance of the model.

As a conclusion, our study indicates that machine learning supported regression models have a great potential for hotel sales prediction. Additionally there are some basic takeaways for the travel agency such as important features on decision making and the benefit of segmentation. As declared in results, there are hotel feature groups which are better to be considered while making decision on hotel selection for supplier agreements. Hotels which have outdoor_sports, summer_holiday, entertainment, indoor_sports, water_sports related features should be prioritised while adding new suppliers to the agency. Moreover, sales strategies are better to be built by considering specific customer segments rather than all customer segments together to capture the benefit that can be obtained from the resemblance of preferences of customers.

In the future work, a better fit and accurate results might be accomplished with a better quality and enriched dataset. Such as a dataset resulted from a more precise and enriched data collection, better grouping of the hotel features, and taking external effects on hotel preferences of customers into account. As an example to the enriched dataset, the data on the hotel options offered to the customers and their preferences could be gathered. Adding this data as a new feature on the dataset gives the ability to evaluate customer preferences of the hotels at first hand and this might have a positive impact on the predictive model. Additionally, there is a potential bias on the data because of some sales agreements, which are the results of the specific business decisions, between some of the hotels and the agency. These agreements bring out a large amount of sales regardless of the features of these hotels, which can not be captured by using only the predictive models. The data can be reorganized and preprocessed to eliminate this effect by determining these agreement bound hotels and their impact on sales or to group hotels according to having these specific agreements or not and running models separately for each hotel groups. Moreover, the hotel features are grouped heuristically and without considering the degree of relevance on sales and any expert opinion in the scope of this study. We think the predictive models can be improved by including the domain knowledge in hotel feature grouping process and even by eliminating irrelevant features from the data. Finally, hotel transaction dates are not considered in this study. Grouping transactions by the time of the year and running models with these different groups of datasets might give an additional insight on the seasonality of the customers' hotel preferences.

BIBLIOGRAPHY

Chen, K.-Y. & Wang, C.-H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, *28*(1), 215–226.

Claveria, O., Monte, E., & Torra, S. (2016). Combination forecasts of tourism demand with machine learning models. *Applied economics letters*, *23*(6), 428–431.

Czajkowski, M. & Kretowski, M. (2016). The role of decision tree representation in regression problems – an evolutionary perspective. *Applied soft computing*, *48*, 458–475.

De Leone, R., Pietrini, M., & Giovannelli, A. (2015). Photovoltaic energy production forecast using support vector regression. *Neural computing applications*, *26*(8), 1955–1962.

Farizal, Qaradhawi, Y., Cornelis, C. I., & Dachyar, M. (2020). Fast moving product demand forecasting model with multi linear regression. *AIP Conference Proceedings*, *2227*(1), 040028.

Guo, Z., Yu, B., Hao, M., Wang, W., Jiang, Y., & Zong, F. (2021). A novel hybrid method for flight departure delay prediction using random forest regression and maximal information coefficient. *Aerospace Science and Technology*, *116*, 106822.

Jae Joon, A., Hyun Woo, B., Kyong Joo, O., & Tae Yoon, K. (2012). Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting. *Expert Systems with Applications*, *39*(9), 8369–8379.

Jakkula, V. (2006). Tutorial on support vector machine (svm). *School of EECS, Washington State University*, *37*.

Kaya, E., Alpan, E., Balcisoy, S., & Bozkaya, B. (2021). Quantifying insurance agency channel dynamics using premium sales big data and external factors. *Big Data*, *9*(2), 116–131.

Khandelwal, V., Chaturvedi, A. K., & Gupta, C. P. (2020). Amazon ec2 spot price prediction using regression random forests. *IEEE transactions on cloud computing*, *8*(1), 59–72.

Kuangnan, F., Yefei, J., & Malin, S. (2016). Customer profitability forecasting using big data analytics: A case study of the insurance industry. *Computers Industrial Engineering*, *101*, 554–564.

Lee, M. (2018). Modeling and forecasting hotel room demand based on advance booking information. *Tourism management (1982)*, *66*, 62–71.

Li, T., Zhou, Y., Li, X., Wu, J., & He, T. (2019). Forecasting daily crude oil prices using improved ceemdan and ridge regression-based predictors. *Energies*, *12*(19).

Maria, E., Budiman, E., Haviluddin, H., & Taruk, M. (2020). Measure distance locating nearest public facilities using haversine and euclidean methods. *Journal of Physics: Conference Series*, *1450*, 012080.

McDonald, G. C. (2009). Ridge regression. *WIREs Computational Statistics*, *1*(1), 93–100.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos,

A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Provenzano, D. & Baggio, R. (2019). Quantitative methods in tourism and hospitality: a perspective article. *Tourism Review*, *ahead-of-print*.

Pulido-Calvo, I., Montesinos, P., Roldán, J., & Ruiz-Navarro, F. (2007). Linear regressions and neural approaches to water demand forecasting in irrigation districts with telemetry systems. *Biosystems engineering*, *97*(2), 283–293.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation, 532–538.

Rhee, H. & Yang, S.-B. (2014). How does hotel attribute importance vary among different travelers? an exploratory case study based on a conjoint analysis. *Electronic Markets*, *25*.

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, *43*(6), 1947–1958.

Tingting, F. & Risto, L. (2016). Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. *Applied Energy*, *179*, 544–552.

Tso, G. K. & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761–1768.

Weisberg, S. (2005). *Applied Linear Regression*, (pp. 1–2). Hoboken, New Jersey: John Wiley Sons, Inc.

Wu, C.-H., Ho, J.-M., & Lee, D. (2004). Travel-time prediction with support vector regression. *IEEE transactions on intelligent transportation systems*, *5*(4), 276–281.

Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. *Remote Sensing of Environment*, *97*(3), 322–336.

Xue, L. & Zhang, Y. (2020). The effect of distance on tourist behavior: A study based on social media data. *Annals of Tourism Research*, *82*, 102916.

Yu, Y., Wang, Y., Gao, S., & Tang, Z. (2017). Statistical modeling and prediction for tourism economy using dendritic neural network. *Computational intelligence and neuroscience*, *2017*, 7436948–9.

APPENDIX A

Model Training Cross Validation Results for All Data Sets and Additional Figures

Table A.1 Linear Regression on All Segments Dataset

			Without PoI	0.5Km PoI
			Best Results	Best Results
Linear Regression	MAE	Train	53.75	53.75
		Validation	58.68	58.68
	R2	Train	0.35	0.35
		Validation	0.06	0.06
	MSE	Train	8,840.33	8,840.33
		Validation	11,228.56	11,228.56

			1Km PoI	3Km PoI
			Best Results	Best Results
Linear Regression	MAE	Train	53.75	53.75
		Validation	58.68	58.68
	R2	Train	0.35	0.35
		Validation	0.06	0.06
	MSE	Train	8,840.33	8,840.33
		Validation	11,228.56	11,228.56

Table A.2 Ridge Regression on All Segments Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.57	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.52
		Validation		57.77		57.76
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24
		Validation		0.19		0.19
	MSE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	10,347.00	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	10,334.51
		Validation		10,807.77		10,796.18

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.88	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	56.64
		Validation		57.82		57.89
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.24
		Validation		0.19		0.19
	MSE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	10,331.05	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	10,341.76
		Validation		10,793.09		10,804.30

Table A.3 Decision Tree Regression on All Segments Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	52.50	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	49.29
		Validation		57.15		57.61
	R2	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.31	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	0.38
		Validation		0.14		0.11
	MSE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	9,425.98	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	8,520.58
		Validation		11,016.98		11,316.51

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	50.81	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	51.08
		Validation		57.41		57.21
	R2	Train	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	0.34	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	0.34
		Validation		0.09		0.10
	MSE	Train	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	9,039.03	#max_depth=8 #max_features=sqrt #min_samples_leaf=20 #min_samples_split=20	9,055.33
		Validation		11,321.95		11,472.06

Table A.4 Random Forest Regression on All Segments Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20 #n_estimators=1000	51.29	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	49.41
		Validation		54.24		53.81
	R2	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20 #n_estimators=1000	0.34	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	0.38
		Validation		0.23		0.23
	MSE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20 #n_estimators=1000	9,050.22	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	8,506.46
		Validation		10,107.02		10,056.38

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	49.16	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	47.15
		Validation		53.73		53.92
	R2	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	0.38	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	0.44
		Validation		0.23		0.23
	MSE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	8,432.31	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	7,688.62
		Validation		9,996.70		10,093.86

Table A.5 Support Vector Regression on All Segments Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	37.08	#c=100 #gamma=scale #kernel=rbf	37.11
		Validation		50.23		50.31
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.34	#c=100 #gamma=scale #kernel=rbf	0.33
		Validation		0.17		0.17
	MSE	Train	#c=100 #gamma=scale #kernel=rbf	8,970.01	#c=100 #gamma=scale #kernel=rbf	9,089.48
		Validation		11,339.36		11,342.77

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	37.69	#c=100 #gamma=scale #kernel=rbf	37.17
		Validation		50.71		50.31
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.34	#c=100 #gamma=scale #kernel=rbf	0.33
		Validation		0.26		0.17
	MSE	Train	#c=100 #gamma=scale #kernel=rbf	9,050.96	#c=100 #gamma=scale #kernel=rbf	9,102.58
		Validation		9,456.55		11,307.07

Table A.6 Linear Regression on Two Adults Segment Dataset

			Without PoI	0.5Km PoI
			Best Results	Best Results
Linear Regression	MAE	Train	32.38	32.38
		Validation	34.38	34.38
	R2	Train	0.31	0.31
		Validation	0.17	0.17
	MSE	Train	2,745.06	2,745.06
		Validation	3,130.20	3,130.20

			1Km PoI	3Km PoI
			Best Results	Best Results
Linear Regression	MAE	Train	32.38	32.26
		Validation	34.38	34.25
	R2	Train	0.31	0.31
		Validation	0.17	0.16
	MSE	Train	2,765.06	2,776.35
		Validation	3,130.20	3,166.33

Table A.7 Ridge Regression on Two Adults Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	32.18	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.11
		Validation		33.97		33.98
	R2	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	0.30	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30
		Validation		0.18		0.18
	MSE	Train	#alpha=100 #fit_intercept=True #normalize=False #solver=lsqr	2,790.12	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	2,784.82
		Validation		3,102.98		3,102.79

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.13	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	32.11
		Validation		33.98		34.01
	R2	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.30
		Validation		0.18		0.18
	MSE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	2,785.67	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	2,785.41
		Validation		3,103.20		3,106.00

Table A.8 Decision Tree Regression on Two Adults Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	32.99	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	31.96
		Validation		34.93		33.91
	R2	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.27	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.32
		Validation		0.12		0.13
	MSE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	2,919.47	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	2,717.12
		Validation		3,323.59		3,178.48

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	31.99	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	31.68
		Validation		33.91		34.07
	R2	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.32	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	0.33
		Validation		0.14		0.14
	MSE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	2,713.44	#max_depth=3 #max_features=log2 #min_samples_leaf=20 #min_samples_split=20	2,671.25
		Validation		3,153.34		3,175.28

Table A.9 Random Forest Regression on Two Adults Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	29.41	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	29.07
		Validation		32.92		32.95
	R2	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	0.39	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	0.40
		Validation		0.21		0.21
	MSE	Train	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	2,435.66	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	2,379.07
		Validation		3,032.54		3,038.52

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	28.77	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	28.31
		Validation		32.78		32.95
	R2	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	0.42	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	0.43
		Validation		0.21		0.21
	MSE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	2,333.71	#max_depth=8 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=20 #n_estimators=1000	2,270.08
		Validation		3,007.71		3,021.82

Table A.10 Support Vector Regression on Two Adults Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
SV Regression	MAE	Train	#c=10 #gamma=scale #kernel=poly	22.19	#c=100 #gamma=scale #kernel=poly	14.90
		Validation		29.89		31.24
	R2	Train	#c=10 #gamma=scale #kernel=poly	0.39	#c=100 #gamma=scale #kernel=poly	0.57
		Validation		0.19		0.20
	MSE	Train	#c=10 #gamma=scale #kernel=poly	2,450.61	#c=100 #gamma=scale #kernel=poly	1,733.94
		Validation		3,222.41		3,102.02

			1Km PoI		3Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=poly	14.90	#c=100 #gamma=scale #kernel=poly	15.00
		Validation		31.48		31.35
	R2	Train	#c=100 #gamma=scale #kernel=poly	0.57	#c=100 #gamma=scale #kernel=poly	0.57
		Validation		0.19		0.19
	MSE	Train	#c=100 #gamma=scale #kernel=poly	1,732.10	#c=100 #gamma=scale #kernel=poly	1,729.86
		Validation		3,134.92		3,127.92

Table A.11 Linear Regression on Family Segment Dataset

			Without PoI	0.5Km PoI
			Best Results	Best Results
Linear Regression	MAE	Train	39.35	39.29
		Validation	41.93	42.43
	R2	Train	0.25	0.26
		Validation	0.10	0.10
	MSE	Train	3,834.97	3,753.06
		Validation	4,307.39	4,286.12

Table A.12 Ridge Regression on Family Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
Ridge Regression	MAE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	39.22	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	39.14
		Validation		41.91		42.03
	R2	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.24	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	0.25
		Validation		0.19		0.10
	MSE	Train	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	3,879.42	#alpha=10 #fit_intercept=True #normalize=False #solver=lsqr	3,806.97
		Validation		3,899.32		4,370.20

Table A.13 Decision Tree Regression on Family Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
DT Regression	MAE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40	37.17	#max_depth=3 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	38.46
		Validation		42.68		42.77
	R2	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40	0.29	#max_depth=3 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	0.25
		Validation		0.08		0.05
	MSE	Train	#max_depth=3 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40	3,633.84	#max_depth=3 #max_features=sqrt #min_samples_leaf=10 #min_samples_split=40	3,812.21
		Validation		4,484.30		4,550.47

Table A.14 Random Forest Regression on Family Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
RF Regression	MAE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	37.93	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	37.68
		Validation		40.75		40.76
	R2	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	0.27	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	0.29
		Validation		0.14		0.15
	MSE	Train	#max_depth=5 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	3,699.37	#max_depth=8 #max_features=log2 #min_samples_leaf=10 #min_samples_split=40 #n_estimators=1000	3,635.14
		Validation		4,214.85		4,181.59

Table A.15 Support Vector Regression on Family Segment Dataset

			Without PoI		0.5Km PoI	
			Best Estimator	Best Results	Best Estimator	Best Results
SV Regression	MAE	Train	#c=100 #gamma=scale #kernel=rbf	24.27	#c=100 #gamma=scale #kernel=rbf	24.80
		Validation		39.01		38.73
	R2	Train	#c=100 #gamma=scale #kernel=rbf	0.35	#c=100 #gamma=scale #kernel=rbf	0.33
		Validation		0.03		0.04
	MSE	Train	#c=100 #gamma=scale #kernel=rbf	3,317.86	#c=100 #gamma=scale #kernel=rbf	3,397.79
		Validation		4,849.22		4,816.31

Figure A.1 Correlation Between PoI 3Km. Categories

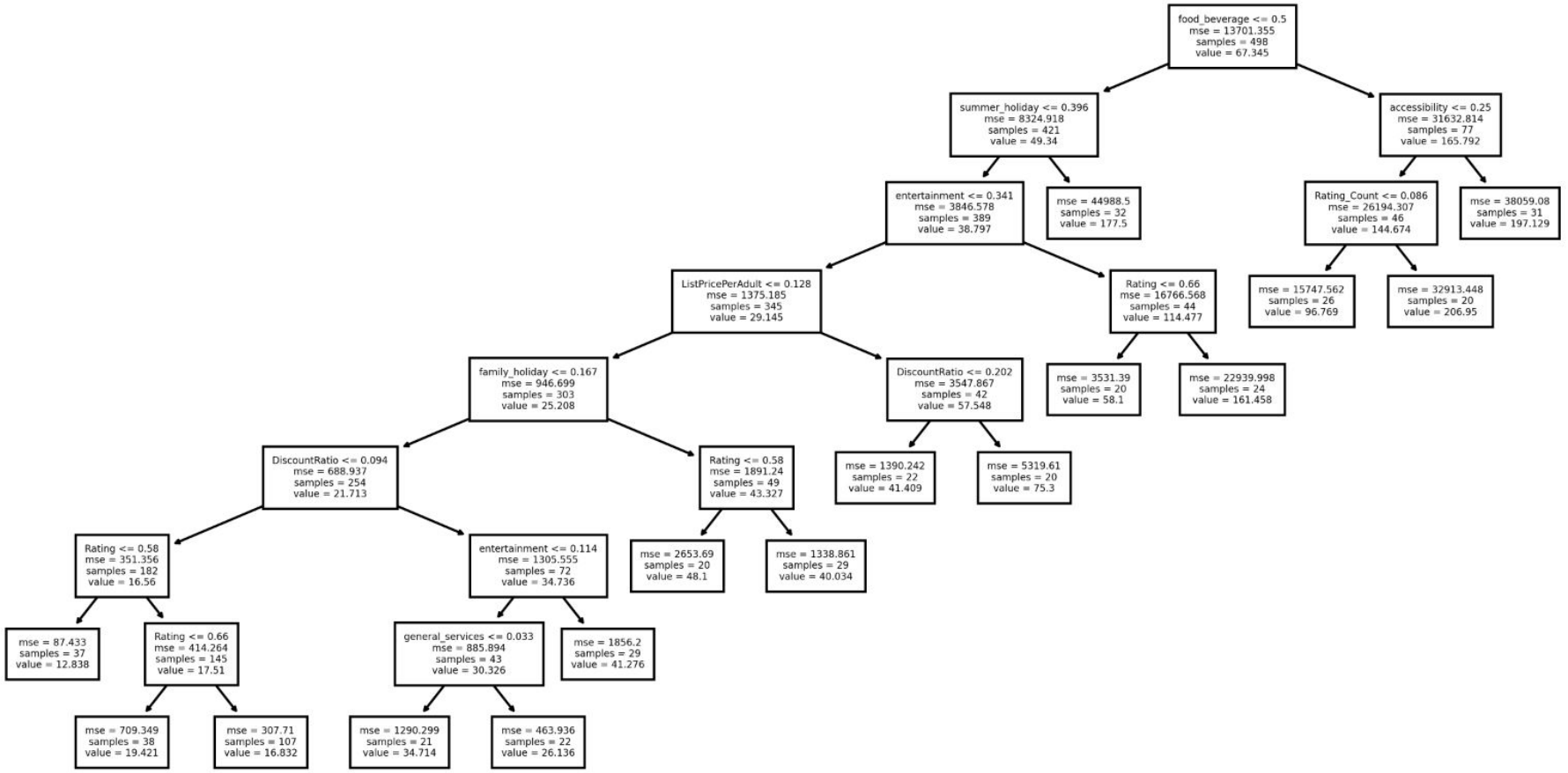

Figure A.2 Decision Tree Regression Best Cross Validation Result on All Segments Without PoI Dataset