ORIE 5740 Final Project Report

Younes Bensouda Koraichi

Pratyush Kothiyal

Professor: Damek Davis

Cornell University

Spring 2022

Abstract: The purpose of this project is to help a Portuguese hotel to increase its revenue by delivering three analyses of the hotel's customer database. After cleaning our data, we first identified three client segments using K-means clustering which categorizes clients by their length of stay, age, and rooms booked. Next, we fit several linear and nonlinear regression models to predict and identify factors that influence non-lodging client expenses, concluding with an XGBoost model that fits best and implies that the length of stay and price of lodging influences non-lodging revenue. Finally, we fit a linear and nonlinear classification model to predict and identify factors that influence a client's bed preference, and find that both models fit each preference better and the contributing factors for best fit are primarily a client's age and length of stay, following intuition as these two factors indicate the desire of a customer to want a more comfortable stay.

I. Introduction

The dataset describes customer data from a four-star hotel located in Lisbon (Portugal), between 2015 and 2018 [1]. It shows 31 variables that identify the 83,590 clients who visited the hotel during this period. The variables are either numeric, boolean or categorical. They belong to one of four categories:

- **Biographical:** Contains personal information of each customer, including name and valid ID number such as passport ID (both hashed for anonymization), age and nationality.
- **Guest History:** Contains historical information of a guest profile, including number of bookings the customer checked in, canceled, and no showed, how many nights the customer stayed, total lodging and non-lodging expenses, average lead time, days since the guest's first and last reservation, number of rooms and guests scaled by length of stay.
- **Reservation:** Contains details of the reservation, including who booked the reservation (travel agent, corporate, direct guest, etc.), and the distribution channel (where the customer made the reservation, be it website, referral, etc.)
- **Booking Preference:** Contains boolean information of whether a guest indicated a room preference, including the floor of the room, bed size, distance from elevator, whether the room is quiet, has a shower, has a bathtub, and has no alcohol in the minibar.

We conduct analyses on this dataset to answer three questions: First, what customer segments can we identify; second, can we predict how much non-lodging expenses a guest spends; and finally, can we predict the bed preference of a customer. Due to our first analysis, we found three distinct categories for customer segments. Our second analysis did not result in a strong modeling relationship, as the best fitted model was a XGBoost model with a very weak

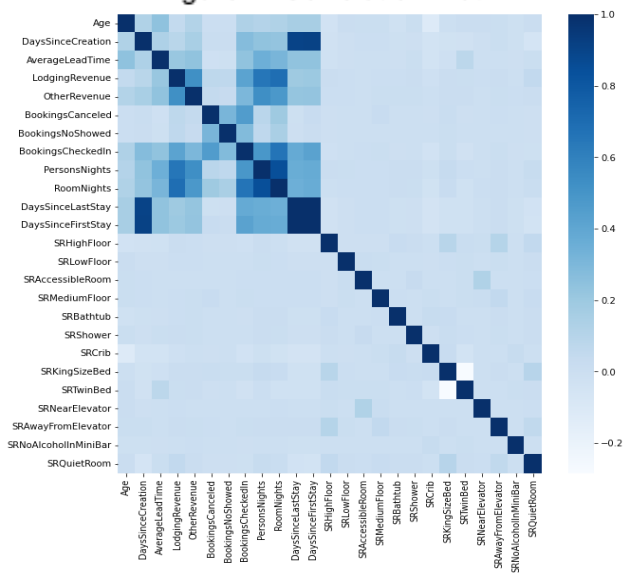
R squared score. Finally, our third analysis found that a Logistic Regression and Random Forest model best classifies bed preference with a reasonable accuracy.

II. Data Cleanup

After investigating our initial variables, we began the task of cleaning and transforming the data. First, we removed the NameHash and DocIDHash variables from the dataset because we consider those variables not useful to our study. We also transformed the categorical variables using one-hot encoding. With the inclusion of these variables, our total number of variables for analysis becomes 224. Finally, we found that 4.5% of the rows have the age missing. We decide to impute the missing ages with the mean of the column, so that we can use these rows for modeling even with this inconvenience.

There exists correlation between pairs of variables that are calculated using similar metrics. For instance, DaysSinceFirstStay and DaysSinceLastStay will be the same for customers who only have stayed at the hotel once. Similarly, PersonsNights and RoomNights both use the number of nights of a reservation, but scales it by the factors of the number of guests and rooms in a reservation respectively. In both cases, it is clear the source of their correlation yet we do not believe they measure similar quantities, and thus we include them in our analyses.

Figure 1: Correlation Matrix



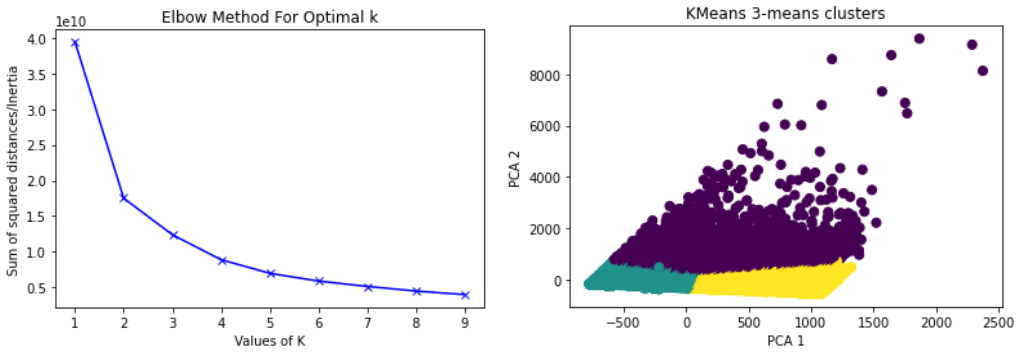
III. Client Segmentation

For this task, we wanted to help the hotel understand their client database better, through client segmentation. It consists of finding groups of customers within the dataset that are similar between customers of the same group, but different between groups.

For dimensionality reduction and visualization purposes, the first step of the client segmentation task was to use PCA. With this method, we were able to reduce the dimensionality of the features from 224 to 2, while keeping 94% of the variance. (67% on the first axis, and 27% on the second).

The clustering algorithm which was chosen was K-Means. We first had to decide the number of clusters K that we needed for the segmentation. For this, we used the Elbow method, which consists of running the KMeans algorithm for several K (here 1 to 9), and compute the sum of squared distances to the center of each cluster divided by the inertia (see figure).

Figure 2: PCA and K-Means Clustering Results



We chose K=3 given the shape of the curve, and then ran the KMeans algorithm on the dataset resulting from PCA. The conclusions drawn from this graph are mixed. We expected to be able to better discern distinct client groups, but we can hardly see them on the plot. However, we can still draw some takeaways from each group, by analyzing the mean of the features within each cluster. Below is a summary of each cluster:

- Young, Short-Term Guests (purple, 42895 points): these bring the smallest revenue, are the youngest, and have the shortest lead time.
- Old Loyal Customers (yellow, 35807 points): these bring high revenue, are old in age and have old guest profiles.
- Corporate Clients (green, 4888 points): these are a very high revenue clientele, often make a corporate booking with a very high lead time and no-show rate, and book the highest number of rooms and people per reservation.

IV. Predicting Client Expenses

The underlying idea behind this goal is to enable the hotel to know how much a client will spend on expenses that are not Lodging Expenses. Those expenses are described by the variable *OtherRevenue* and represent the total amount spent by the client on food, beverage, spa and other expenses. Knowing if a customer will spend a lot of money on other expenses can be helpful for the hotel booking website system for example, to send incentives or offers for those clients who might spend a lot, because those spendings are extremely profitable.

After all the preprocessing was completed, we have a dataset that is completely real-valued and checked for anomalies and missing values. This dataset is now ready to be used for regression modeling. We split the dataset between training (80% of the data) and test set (20% of the data). When using cross validation, the training set was used.

We used a simple linear regression model to investigate whether a linear relationship exists between the predictors and the response variable *OtherRevenue*. We observe a very large testing MSE of 3.0 x 10^12 compared to 8844 on the training set. To improve our base model, we decided to introduce regularization. We set lambda by using an efficient LOOCV cross validation enabled by sklearn on the train set, to find alpha equal to 376. For L2 regularization, our training error is similar to base regression, being 8846, but our testing MSE significantly improved to 6797. L1 regularization had similar results, with a training MSE of 8913

and test MSE of 6796. Furthermore, we found that L1 regularization kept 21 variables: Age, DaysSinceCreation, AverageLeadTime, LodgingRevenue, PersonsNights, RoomNights, SRCrib, 11 Nationality variables, and the 3 Market Segment variables. Although we found improvements for our test error, all linear models had very low R squared scores, none reaching over 0.40 in either training or test sets. Thus, we conclude that a linear relationship does not best model this relationship, and decide to pursue nonlinear models.

For nonlinear models, we decided to build a Random Forest and XGBoost model. In both cases, we used cross validation on the training set, with 2 folds fitted on a set of parameters shifting the maximum depth of trees for both models, and learning rate, L1 regularization on lead weights, and percentage of features used per tree for XGBoost. Our Random Forest model with a maximum depth of 10 yields a training error of 5147 and test error of 6259. The results are better than all the previous models, but we are still slightly overfitting. We compare these results to XGBoost, with alpha set to 1, the percentage of features used per tree to 30%, the learning rate to 0.1 and the maximum depth of trees to 5. It provides similar results, with a larger training error of 6140 and test error of 6233. The top features for this model are PersonsNights, LodgingRevenue, and BookingsCheckIn, which follows intuition as customers who frequently stay at the hotel or have a large party have many opportunities to spend money on hotel amenities and extra services. Between these two models, we choose XGBoost because it minimizes the test MSE. Although the test R squared of 0.44 does not indicate a very strong correlation, we believe that it is still helpful for this task in identifying important factors that contribute to a client's non-lodging expense.

V. Predicting Client Lodging Preferences

Among the features included in the dataset are many boolean values that indicate the lodging preferences of the customer. We attempt here to predict the lodging preferences of

customers based on their demographic data. Understanding this information can aid the hotel in guided suggestions for lodging preferences, thus increasing customer satisfaction. Our scope for this report is to identify customers likely to prefer one of two bed preferences: King Size and Twin Bed. We use two classification models, one linear and one non-linear, to model this relationship.

Our chosen linear model is a logistic regression model using 2 fold cross validation, with L2 regularization using Grid Search. The model performs well in predicting Twin Bed classification, with a training and test accuracy of 86%. In contrast, the model appears to perform poorly when predicting King Size Bed, with a training and test accuracy of 64%. This may indicate that the underlying relationship between the King Size Bed classification and the predictor variables is more nonlinear compared to the Twin Bed classification.

We chose to use the Random Forest model for our nonlinear model, with a 2 fold cross validation for each factor, and found a max depth of 30 to be sufficient for Twin Bed and 50 for King Size Bed. Compared to Logistic Regression, the Random Forest model fits better for both Twin and King Size test sets, with a higher test accuracy of 89% and 74% respectively. We observe that for both features, the top two most impactful features are DaysSinceCreation and Age. These features are representative of the Old Loyal Customer segment found in Section I. It follows intuition that as these customers are more frequent, they would be more likely to note their bed preference.

VI. Conclusion

In our analysis, we succeed in two of our three analyses. We identified three distinct groups of guests: young, last minute customers that bring in low revenue, old loyal customers that bring in high revenue, and corporate clients that reserve multiple rooms and also bring in high revenue. We also found that a Random Forest model can best predict a customer’s bed

preference with a reasonable accuracy above 70%, and top factors in bed preference are the age of a guest and length of stay.

Although we believe our second analysis, which involved building a regression model to predict a client's non-lodging expenses, is helpful in identifying impactful factors that predict non-lodging expenses, there is room to improve the model to obtain a higher R squared score. In future studies, there are two strategies to improve this model. First, one may transform the response variable, as it may differ by several orders of magnitude. Second, adding additional non-correlated variables, such as average customer rating or additional guest metrics, may provide better context and increase the regression scores.

Table 1: Client Expense Regression Results

	Training		Test	
Model	MSE	R Squared	MSE	R Squared
Linear Regression	8844	0.348	3001×10^9	-2706×10^5
Ridge Regression	8846	0.348	6797	0.387
Lasso Regression	8913	0.342	6796	0.387
Random Forest	5147	0.62	6259	0.43
XGBoost	**6140**	**0.55**	**6233**	**0.44**

Table 2: Bed Preference Classification Model Accuracy

	Twin Bed		King Size Bed	
Twin Bed	Train	Test	Train	Test
Logistic Regression	0.856	0.859	0.639	0.639
Random Forest	**0.924**	**0.886**	**0.994**	**0.737**

Bibliography

1. Antonio, Nuno, et al. “A Hotel's Customers Personal, Behavioral, Demographic, and Geographic Dataset from Lisbon, Portugal (2015–2018).” *Data in Brief*, Elsevier, 24 Nov. 2020, https://www.sciencedirect.com/science/article/pii/S2352340920314645?via=ihub.