Team 106 Final Project Report
Optimizing Prediction of Canceled Reservations

Isaac Espinoza (iespinoza6: imespinoza@gatech.edu)
Michelle Kwak (mkwak9: mkwak9@gatech.edu)
Vaisag Radhakrishnan (vradhakr8: vradhakr8@gatech.edu)

Submission Date: 16-April-2023
Course Name: ISYE 7406 Data Mining and Statistical Learning

1. Abstract

Online hotel reservations have been on the rise due to the convenience of booking on the internet for the customers, as well as the ease of booking management for the hotels. However, these bookings tend to be cancelled at a higher proportion than any other type of booking due to the ease of cancellations and lack of penalties (fees). It is critical for the businesses to forecast and predict these cancellations to reduce their impact on the revenue. The aim of this research is to help hotels and the hospitality industry in general to improve the forecasting and prediction of cancellations by analyzing various historical transactions made by customers. Many machine learning models were developed as part of this study using data over a two-year period. Ensemble methods like Random Forest and Gradient Boosting methods performed the best. Additionally, it was observed that the features like lead time, number of special requests and average price per room were the most important variables that helped predict the cancellations successfully.

2. Introduction

Online hotel reservations have increased, compared to other traditional methods such as phone calls or travel agencies, due to the development of technologies related to the internet and cost cutting measures by the hospitality industry. However, due to the flexibility of online reservations and low cancellation fees, there are increasing numbers of cancellations every year. Free cancellations are a tremendous benefit for the consumers, but they have a detrimental impact on the hotel's earnings.

The main goal of this project is to reliably predict, based on a hotel reservation's many data collection features, whether it will be canceled. Also, the features will be carefully examined to determine whether they contribute to a reservation's cancellation. According to research by D-edge hospitality solutions, as of 2018, on average, 40% of reservations are canceled before arrival (2019).

2.1. Scientific Research Questions

The team’s goal is to predict whether a reservation will be cancelled or not based on the available features. Additional descriptive questions like the ones below are also considered during this project.

- Is a customer more likely to cancel a reservation when the booking is made days/weeks/months ahead?
- Is a customer more likely to show up when they have stayed with the hotel in the past?
- Does the room price affect cancellation? Do higher prices mean a greater likelihood of cancellation?
- Does season (spring, etc.) or month of the year influence cancellations?
- Does the group size have any impact? Are groups with children more likely to cancel the reservation?

3. Literature Review

The hospitality industry faces difficult challenges in balancing their brand and revenue loss. One method of avoiding revenue loss is overbooking and it requires a good analysis of demand. However, this method comes at risk if their calculations are incorrect and has the potential to damage their hard-earned reputation. Another method is to implement a cancellation policy in the terms of service when booking. This is meant to encourage customers to cancel in advance and reduce the number of idle capacities. However, this could negatively impact prospective customers and they may look for alternative reservations with more flexible options. Other options such as offering a reduced rate but without possibility of cancellation can also have a detrimental effect on the goodwill of the customer towards the hotel chain.

There are several benefits to predicting cancellations. Sanchez et. al state that it plays a crucial role in the operations of modern organizations because it supports a variety of business decisions, from operations, to tactical, to strategic. Some examples include capacity planning, resource planning, advertising, and promotional planning (2020). They implemented Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting Machines (GBM) for tree boosting ensemble methods to develop a model to predict critical cancellations (predictions between 4-7 days before reservation date).

Antonio et. al presented a cancellation prediction system, based on a machine learning model that uses date from a hotel’s Property Management Systems (PMS). Their study involved data from two hotels in a Portuguese hotel chain. The methodology used was XGBoost, a tree boosting machine learning method. They found that guests who were likely to cancel were contacted in advance. The hotel inquired about information that could provide better service. Examples include expected time of arrival, number of children, etc. Without being offered a substantial promotion or offer, guests cancelled much less than guests not contacted (2017).

4. Problem Statements and Data Source

The data for this project is from Kaggle. It is a cleaned dataset from this publication in the Data in Brief journal. There are 36,275 records and 19 columns in the dataset. The dataset includes bookings made by customers in two years 2017-2018. The bookings include both cancellations and non-cancellations.

The dataset was a clean dataset with only a few exceptions. There were no missing values or outliers in any of the fields. The only field that required cleanup was the booking date field created by combining the date, month, and year fields. Date cleanup was carried out only for a few rows in the booking date field where some dates fell on February 29th, 2018. As 2018 was not a leap year, this was considered an error and cleaned up to make the booking date February 28th, 2018.

Table 1 shows the various variables used as part of the final models created. The response variable is the booking_status, that shows whether the booking was cancelled or not. The response variable was converted into a binary variable for ease of working with it in creating the models. The aim of this project is to predict this booking_status field using the various variables in *Table 1*. In addition to this, multiple variables were created to visualize, understand, and interpret the data better.

Feature Name	Description	**Type**
no_of_adults	Number of adults	Numerical
no_of_children	Number of children	Numerical
no_of_weekend_nights	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel	Numerical
no_of_week_nights	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel	Numerical
type_of_meal_plan	Type of meal plan chosen	Categorical
required_car_parking_space	Does the customer require a parking space? (0 - No, 1- Yes)	Binary
room_type_reserved	Type of room reserved by the customer.	Categorical
lead_time	Number of days between the date of booking and the arrival date	Numerical

arrival_year	Year of booking	Date
market_segment_type	Market segment of booking - Complimentary, Online, Offline, Aviation and Corporate booking	Categorical
repeated_guest	Whether customer has already stayed with the hotel before	Binary
no_of_previous_cancellations	Number of historical cancellations by the customer	Numerical
no_of_previous_bookings_not_canceled	Number of times the customer showed up after booking	Numerical
avg_price_per_room	Average price of the room	Numerical
no_of_special_requests	Total number of special requests made by the customer (e.g., high floor, view from the room, etc.)	Numerical
season	The season of the year. Values include winter, spring, summer etc.	Categorical
booking_status	Flag indicating if the booking was canceled or not.	Binary

Table 1: Shows all the variables used to create machine learning models for cancellation prediction.

5. Methodologies Used

The team carried out an initial exploratory data analysis to understand the data better. Many key insights in data mining projects are usually observed during this phase. Past the exploratory data analysis, the research was focused on how to predict the response variable booking status. After the exploratory data analysis, as with any machine learning research project, we continued by creating additional features by one-hot encoding variables. One-hot encoding is a technique to code the categorical variables into numerical codes such as 0,1. This helps improve the machine learning algorithms to improve its prediction accuracy. Post this, the full data was split into train and test data with 70% training and 30% test data. All machine learning algorithms were trained on this 70% train data and tested on the rest of the 30% of data. There were 36,275 records in the dataset split into 25,560 training and 10,715 testing datasets. This step is essential to avoid over-fitting in the model, leading to results that look good but do not perform well.

The team implemented various data mining and statistical techniques to predict a cancellation based on all the features of a transaction. Starting from simple classification methods like KNN, Logistic Regression, Decision Trees, Generalized Additive Models, Naïve Bayes etc., to more complex ensemble methods like Random Forest and Gradient Boosting methods. Although methods like KNN, logistic regression and decision trees produce models that clearly explain the pattern, since weightage is given to predict a cancellation more accurately; ensemble methods work better in general. Random Forest, for example, combines results from multiple decision trees by randomly selecting a selected number of features and a selected number of records for each decision tree. Random Forest then combines all results and takes the majority class predicted by all the trees involved. This improves the predictive capacity of the model. This kind of ensemble method is best suited for a classification application like this due to its ability to predict more accurately than other methods like, say, a single decision tree. Nevertheless, a drawback of these ensemble methods is that one cannot really explain how they arrive at their classification results.

Cross validation was carried out on the models that would most benefit to ensure that the best hyper parameters were used to train the model. Cross validation is a technique that, simply stated, trains the models based on various splits of training data for picking up the best hyper parameters. As an example, for K-Nearest Neighbor (KNN), a technique that uses k neighbors of the test data to predict the class, choosing a value of k is vital. Cross validation was carried out in this case to determine the best k out of the various k-values. K-fold cross validation can also be used during the splitting of the training/testing datasets so the model can run on these multiple splits to try and obtain more accurate results once they are averaged out.

6. Analysis and Results

6.1. Exploratory Data Analysis

Upon analyzing the booking status against the variable lead time, it was observed that, on average, bookings with longer lead time tend to be cancelled more. However, some customers still honored their reservation when bookings were made more than 200 days (about six and a half months) in advance. *Figure 1* shows the boxplots of booking status vs lead time and average price per room. As shown in the figure, it can be observed that, contrary to our hypothesis, the average price does not influence the booking status.

Additionally, analyzing the cancellations over time, it was clear that during the holiday season, i.e., November, December and January, the percentage of cancellations was much less than in other months of the year. This can be observed in *Figure 2.* This may be because people that choose to go on vacation/holiday during this time do not tend to cancel their vacations compared to other times of the year. This might be due to it being the holiday season for most people. It is also likely that many of these winter bookings are business/commercial bookings for conferences, etc., that take advantage of the hotels "slow seasons." These events are well planned and unlikely to be cancelled.

It was also observed that during the winter months, the number of bookings is also reduced.

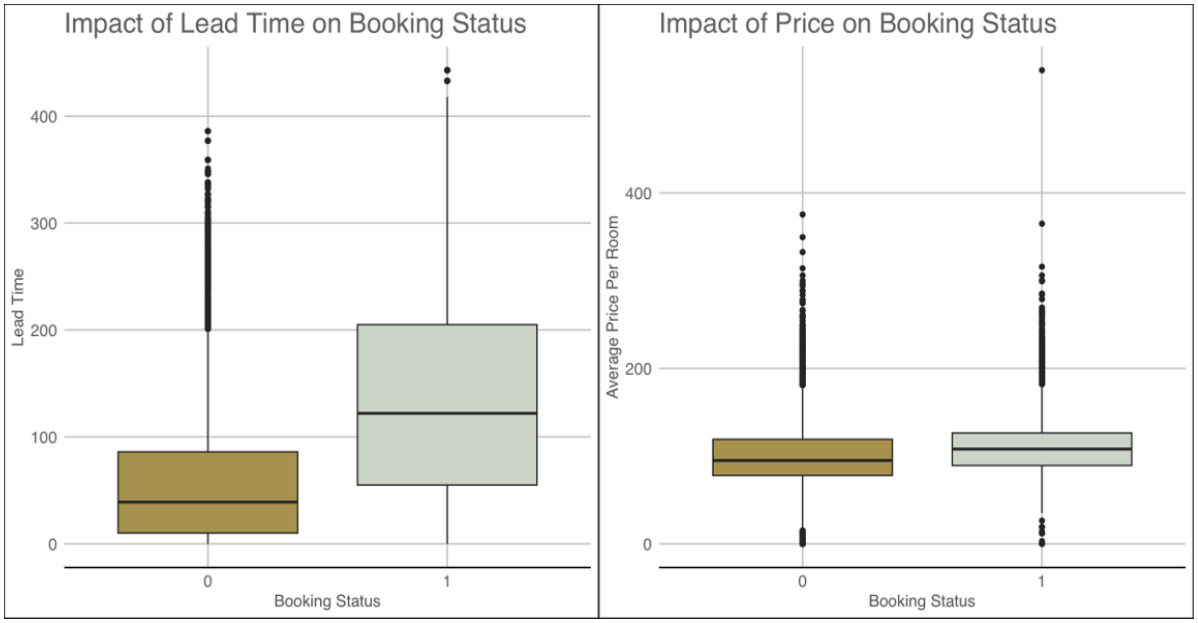

Figure 1: Boxplots showing the impact of lead time and price on the booking status.

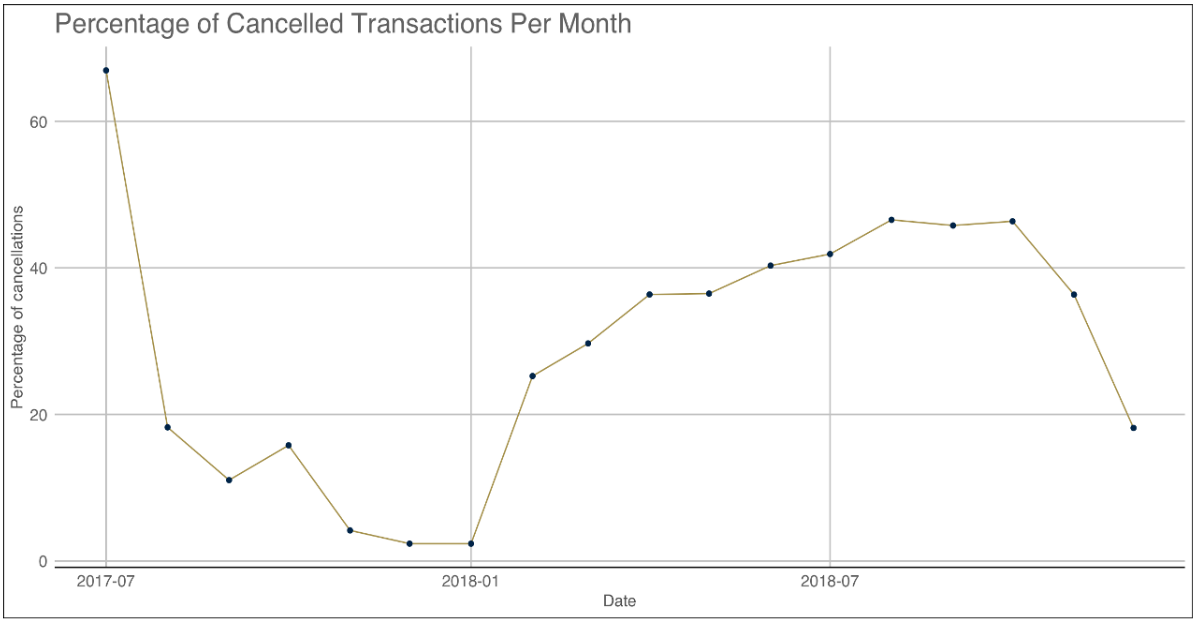

Figure 2: Timeline of monthly cancellation percentages.

Figure 3 shows the percentage of cancellations and non-cancellations based on the arrival day of the week. Bookings tend to be cancelled more on Sundays than any other day of the week. Close to 20% of all cancellations occur on Sundays. On the other hand, Thursday arrivals had the lowest cancellation percentage compared to other days at 12%.

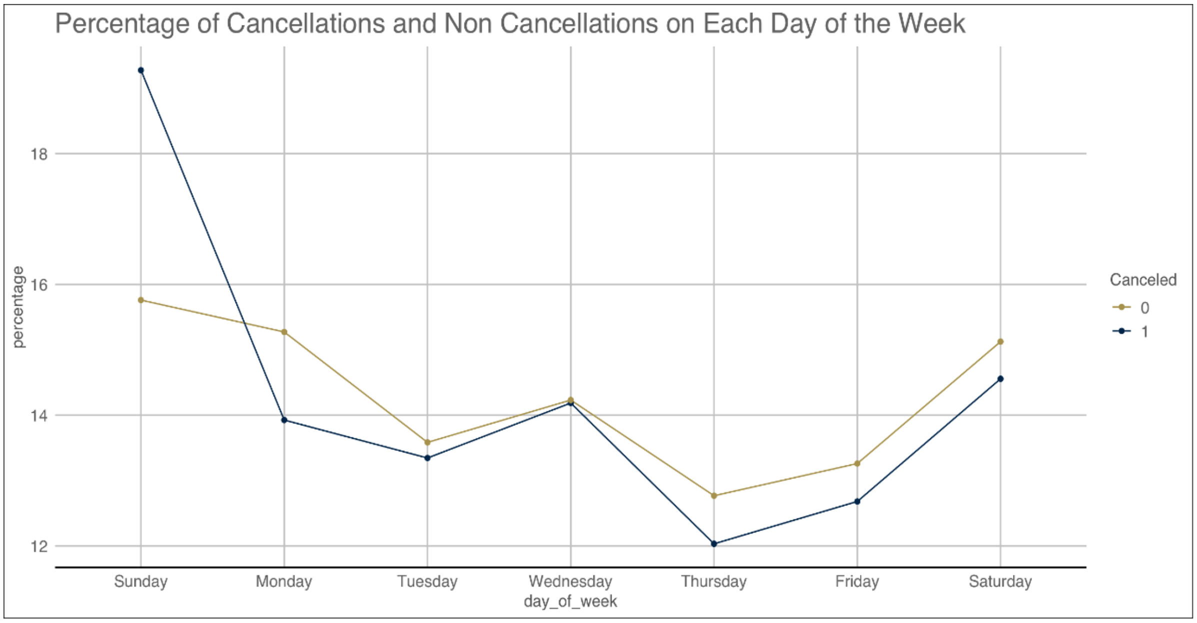

Figure 3: Percentage of cancellations and non-cancellations on each day of the week.

The market segment variable showed what type of booking was made. This field shows whether the booking was made online, offline, through a corporate, complimentary, or aviation. Additional research on the numbers across market segments showed that most reservations were made online. This was not surprising as the number of online reservations is on the rise due to technological advancements. Due to this, the percentage of cancellations is also high for this market segment. 60% of all cancellations were from bookings that were made online. See *Figure 4* shows the percentage of cancellations and non-cancellations across market segments. Complimentary and Corporate bookings had low cancellations. This is expected to be this way as these bookings are made by businesses and do not get cancelled in general compared to individual bookings.

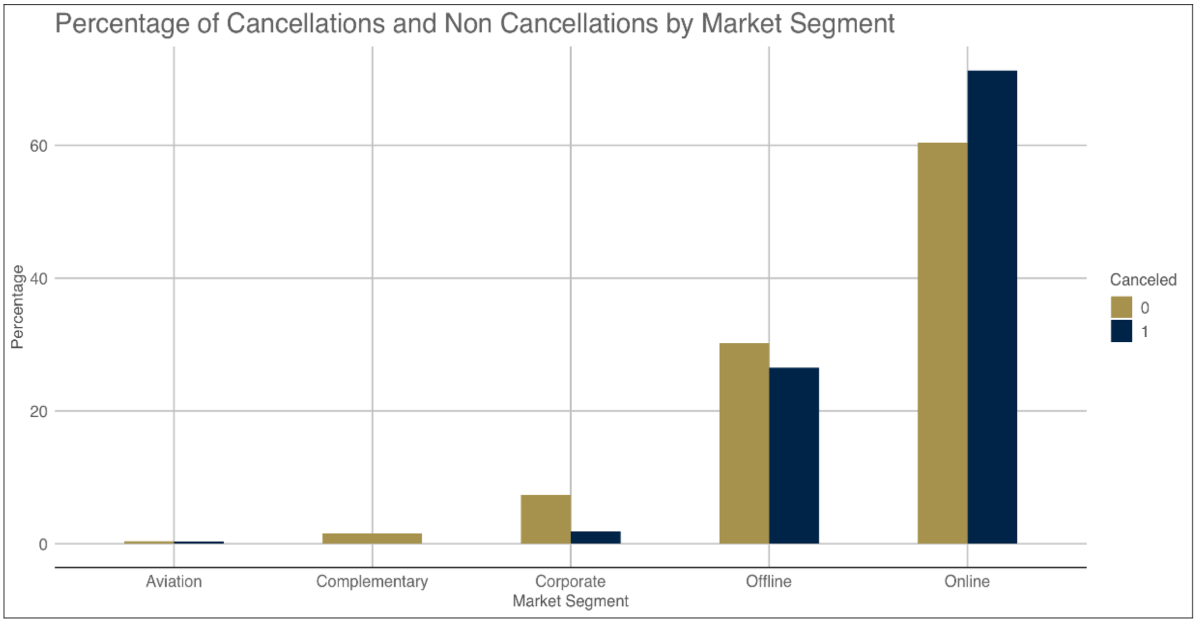

Figure 4: Percentage of cancellations and non-cancellations on each day of the week.

6.2. Results

Multiple data mining methods were used to compare and then obtain the best results for this study. The measure used to compare between models is the mean classification error. Mean classification error shows the mean number of observations in the test data classified incorrectly. Mean classification error was chosen as the parameter as this is the most intuitive measure to compare accuracy across models. *Figure 5* shows the results obtained from various models that were trained and tested as part of this study. This shows that the random forest with hyperparameters tuned showed the least mean classification error. For our specific model different “mtry” and number of trees values were implemented to tune the model properly. “mtry” adds randomness to the random forest model since it controls the number of variables/features a decision tree can consider at any specific time. The number of trees is also important to ensure that the model will provide results in a reasonable amount of time. For our best results, the number of trees chosen was 500 with an “mtry” value of five. Random forest is known for its predictive performance, and it proved to be the best for our dataset. Random forest combined with the hyperparameters tuned chooses the best model out of multiple models. Ensemble methods, in general proved to work the best for this dataset. Ensemble methods, like Random Forest followed by Gradient Boosting method, performed well. As discussed before, ensemble methods are designed to get better predictive performance. Other methods (KNN, Logistic Regression, GAM (Generalized Additive Model) with splines, Naïve Bayes, etc.) shown in *Figure 5* are arranged in increasing order of mean classification error.

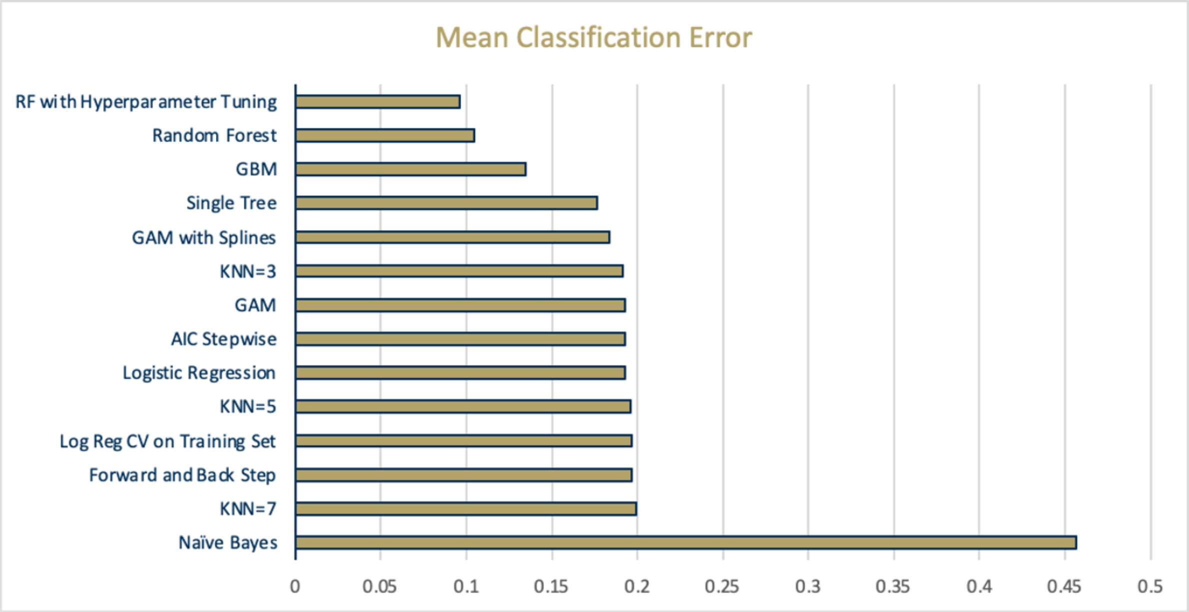

Figure 5: Mean Classification Error results obtained from various machine learning models tried.

As part of this research, one of the other important results that the team was able to find was that the variables lead time, number of special requests and average price per room were the 3 most important variables in determining whether a booking will be cancelled. Figure 6 shows the top five variables arranged in descending importance. This shows that the hypothesis of more lead time leading to cancellations still stands true and should be investigated to understand the influence of lead times on cancellations. The number of special requests is another variable that plays a significant role in bookings being cancelled or not. This may be due to the customers being unhappy about their experience, wanting more done by the hotels, leading to cancellations. Average price per room is the next feature that showed high importance in determining the outcome of the hotel booking. Even though we saw that, on average, the price does not influence the cancellations, this shows that the average price per room is still an important feature in predicting the cancellation or non-cancellations when associated with other independent variables.

K-fold Cross validated results from KNN, and logistic regression are shown in the appendix. This was done to further explore the KNN and logistic regression models since we expected them to also perform well. The 10-fold cross-validation was done on the testing/training splits. Both models performed better than the baseline models after the cross-validation, not to the level of the ensemble methods but still better.

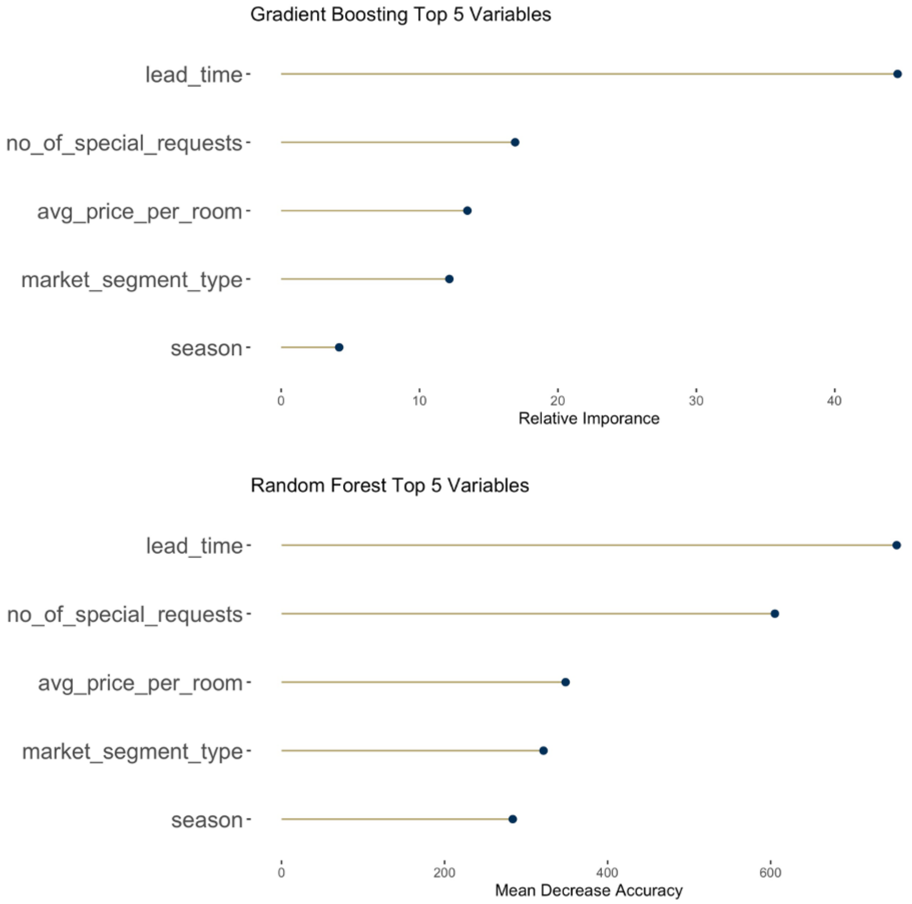

Figure 6: Variable Importance Results. The top graph shows the top 5 variables from gradient boosting method and the bottom from random forest.

7. Future Work and Conclusion

7.1. Future Work and Lessons Learned

During our work, we found out that ensemble methods like random forest, especially when used in conjunction with cross validation, can be computationally intensive. Better fine-tuning methods need to be found to make running this model efficient. If not, these ensemble methods might not be viable for a large-scale real-world prediction set up. But when executed properly, ensemble methods also seem to perform the best. Cross-Validation was also shown to be particularly important in optimizing a model's performance. Further models could be constructed to improve upon the existing ones tried in this project. Models with many more features or features engineered to produce the best predictions (like the top three discovered in this project) could be tried out as an extension of this project. The variables deemed important by the ensemble methods could be used as a starting point to create new models. We also learned that there are methods to deal with skewed data when the data available requires it. Oversampling and undersampling can be indispensable for highly skewed datasets. In our project we did not feel it necessary to make use of those tools but investigated them, nevertheless. Another possibility for future work is to pursue more performance metric observations. For example, ROC is a popular metric of model fitness for classification. To ensure the findings of our study, Precision Recall Curve(PRC) could also be applied for confirmation of a balanced dataset and strength in results. We could also take the findings from this study and expand on other variables, combining multiple methods. Finally, although our results were positive, a larger data set could be used to implement this project to train and validate the model.

The Data Mining and Statistical Learning class provided us with a comprehensive approach to evaluate models and their application to several types of datasets. We started with some of the more basic, well-known, and vastly used models that can be explained and be easily implemented (linear, logistic regression, etc.) and then moved to more complex ones (ensemble, SVM) that can provide results but are more difficult or impossible to explain how they achieve their better results. The class did this in a gradual manner that made it easy to follow along. The homework provided and the R code made available along with the teacher recordings, were extremely helpful in allowing us to follow along and reinforcing the lessons for each week. Many of the lessons learned from the class were applied throughout this project and the final homework.

7.2. Conclusion

We tried multiple data mining and machine learning techniques as part of this project. As shown in figure 5 and explained in the results section, the ensemble methods performed better than the other methods. Perhaps because they can be simpler to optimize. It was also observed that the most important variables were lead time, number of special requests and average price per room were the most important variables in determining whether a reservation will be cancelled or not. The greater the number of days between booking and arrival, the higher the chances of a change of plans. For this reason, hotels should make sure to contact customers who booked their rooms well in advance, at some points before their arrival to ensure that they will still show up on the date of arrival and rebook the rooms in case the customer informs them that they will no longer be going. Depending on how far in advance the booking is made, the number of pre-arrival contacts can be optimized so they do not become intrusive. This will lead to less revenue loss for the business due to cancelled bookings and a better customer experience.

8. Appendix

Figure 8 shows cross validated results for KNN and Logisitic Regression. Mean classification error is the measure used here as well. This shows that KNN had better results across the board in comparison to the logistic regression method. However, this is still more than the ensemble methods as seen in the results section.

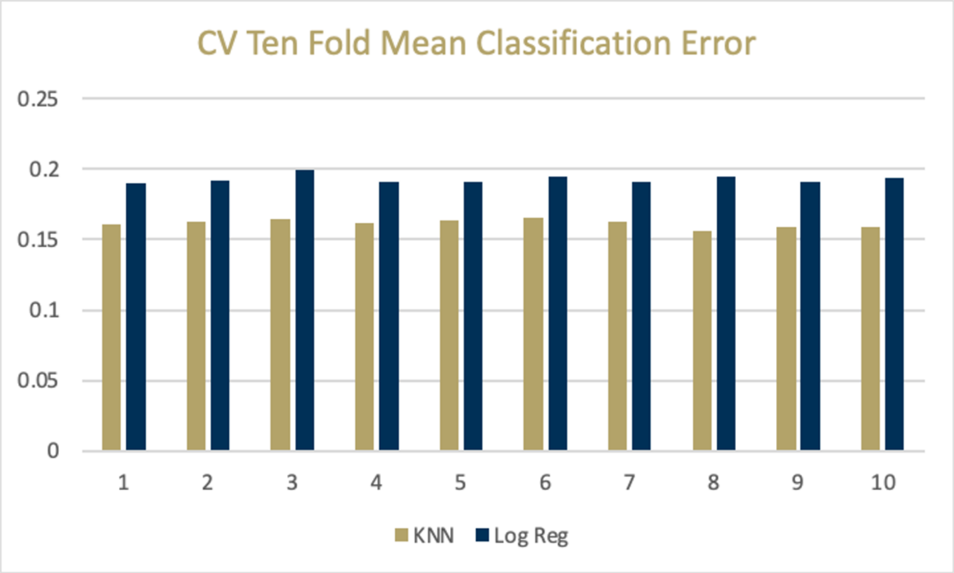

Figure 8: Ten-fold cross validated mean classification error for both KNN and logistic regression.

9. Bibliography and Credits

Antonio, N., de Almeida, A., & Nunes, L. (2017, December). Predicting hotel bookings cancellation with a machine learning classification model. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1049-1054). IEEE.

Global Cancellation Rate of Hotel Reservations Reaches 40% on Average. Hospitality Technology (2019, April 19). Retrieved March 2, 2023, from https://hospitalitytech.com/global-cancellation-rate-hotel-reservations-reaches-40-average#:~:text=04%2F23%2F2019-,Global%20Cancellation%20Rate%20of%20Hotel%20Reservations%20Reaches%2040%25%20on%20Average,Europe%20between%202014%20and%202018.

Sánchez, E. C., Sánchez-Medina, A. J., & Pellejero, M. (2020). Identifying critical hotel cancellations using artificial intelligence. Tourism Management Perspectives, 35, 100718.