Predictive Analytics: Project Report

Morgan Gant

Rachael Doyle

April 23, 2023

**Abstract**

When looking to go on a trip, having a place to stay is one of the most important things to consider. Most people book hotel reservations to ensure a room(s). However, life happens, and schedules are always changing. Canceling a trip might not affect the traveler, but it can negatively impact a hotel who had the room(s) reserved. With the information we have obtained from an online hotel booking channel, we will analyze the given variables, engineer more variables, build models, tune the model parameters, and evaluate the accuracy of predicting whether a hotel guest will cancel their reservation or not.

**Table of Contents\***

[**Project Introduction** 3](#_Toc132799967)

[**Data Description** 3](#_Toc132799968)

[**Exploratory Analysis** 5](#_Toc132799969)

[**Questions to Consider** 9](#_Toc132799970)

[**Feature Engineering** 9](#_Toc132799971)

[**Feature Selection** 11](#_Toc132799972)

[**Hyper Parameter Tuning** 12](#_Toc132799973)

[**Predictive Modeling/Ensemble Method** 13](#_Toc132799974)

[**Conclusion** 14](#_Toc132799975)

[**Implications and Further Questions** 15](#_Toc132799976)

# **Project Introduction**

People can stay in hotels for many different reasons, whether it be for a week-long vacation at a nice resort, for a weekend getaway for sports, or even just a single night for a place to rest, but usually these hotel rooms are booked via a reservation. The reservation system is in place so that someone may reserve a room for however many nights they need and keep a record of the price. Some hotels have rewards systems in place to try and get more customers to consistently stay with them and use these reservation systems. Despite the many reservations hotels receive, they also are losing money in the way of cancellations. In 2019 on average 40% of the reservations in hotels were cancelled before the reservation date.

With this high number of cancellations hotels are wondering if there is a way to predict the possibility of a reservation being cancelled. What factors come into play with cancellation, how many days are leading up to the reservation, the amount of people staying in the room, or even the price of the room being stayed in. Using these features, along with a few others seen in the data set we plan to engineer variables, use hyper parameter tuning, and build models that will predict the likelihood of a reservation being cancelled. The goal of this project is to build a model that will best predict the likelihood of a hotel reservation being cancelled, and show which factors are the primary cause for these cancellations.

# **Data Description**

For this project, a dataset of hotel reservation information was analyzed. This dataset was taken from Kaggle, an online site that contains public data sets. This data set consists of 36,276 rows and 19 columns/variables with no missing values. Each row represents information given from a guest when making a reservation online.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable:** | **Type** | **Description:** | **Missing Values:** |
| **Booking\_ID** | Unique Identifier | Unique identifier of each booking | 0 |
| **no\_of\_adults** | Integer | Number of adults | 0 |
| **no\_of\_children** | Integer | Number of Children | 0 |
| **no\_of\_weekend\_nights** | Integer | Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel | 0 |
| **no\_of\_week\_nights** | Integer | Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel | 0 |
| **type\_of\_meal\_plan** | Categorical | Type of meal plan booked by the customer: | 0 |
| **required\_car\_parking\_space** | Integer | Does the customer require a car parking space? (0 - No, 1- Yes) | 0 |
| **room\_type\_reserved** | Categorical | Type of room reserved by the customer. The values are ciphered (encoded) by INN Hotels. | 0 |
| **lead\_time** | Integer | Number of days between the date of booking and the arrival date | 0 |
| **arrival\_year** | Integer | Year of arrival date | 0 |
| **arrival\_month** | Integer | Month of arrival date | 0 |
| **arrival\_date** | Integer | Date of the month | 0 |
| **market\_segment\_type** | Categorical | Market segment designation. | 0 |
| **repeated\_guest** | Integer | Is the customer a repeated guest? (0 - No, 1- Yes) | 0 |
| **no\_of\_previous\_cancellations** | Integer | Number of previous bookings that were canceled by the customer prior to the current booking | 0 |
| **no\_of\_previous\_bookings\_not\_canceled** | Integer | Number of previous bookings not canceled by the customer prior to the current booking | 0 |
| **avg\_price\_per\_room** | Float | Average price per day of the reservation; prices of the rooms are dynamic. (in euros) | 0 |
| **no\_of\_special\_requests** | Integer | Total number of special requests made by the customer (e.g. high floor, view from the room, etc) | 0 |
| **booking\_status** | Categorical | Flag indicating if the booking was canceled or not. | 0 |

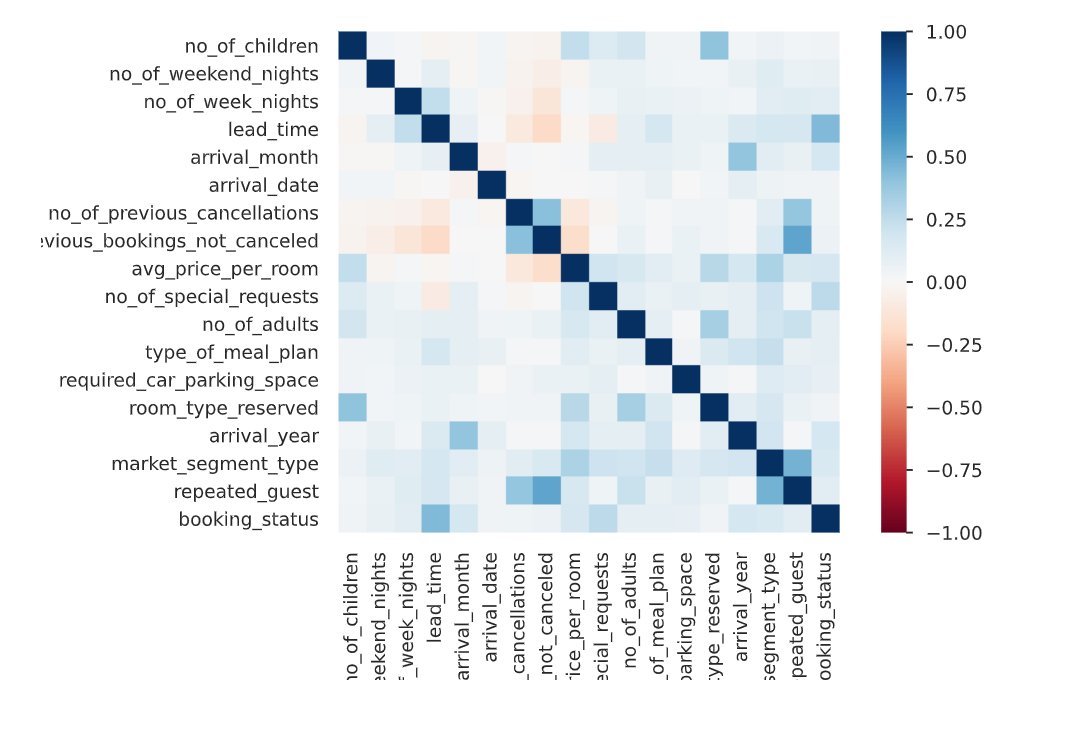
# **Exploratory Analysis**

*Figure 1*



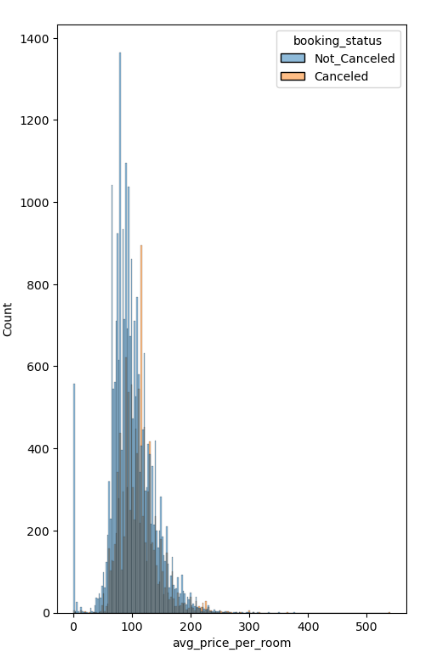
The above graph displays the target variable we are focusing on, the number of reservations that are cancelled or not cancelled. There is almost double the number of non-cancellations compared to cancellations, telling us the data is imbalanced.

*Figure 2*



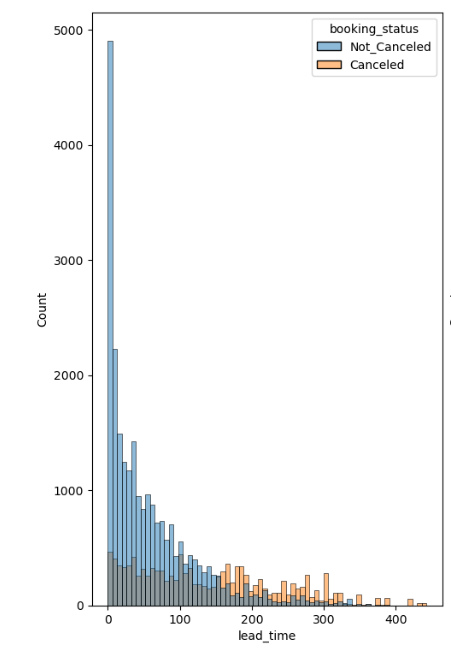
The above chart displays the correlation values between each of the variables. Most of the data is not correlated, which is seen in the lightly shaded boxes, however there are a few correlations that may be worth noting, as seen with the darker shaded boxes. Overall, we would not expect there to be issues with correlation based on the heat map and frequency of lighter colors.

*Figure 3*



*Figure 3* shows the distribution of the average price per room and the count of how many reservations were cancelled and not cancelled, corresponding with the prices. This is slightly skewed right, with most of the data being in the range 50 to 200 for the average room price.

*Figure 4*



The above graph shows the distribution of lead time and the count of how many reservations were cancelled and not cancelled. As a reminder, the lead time is the amount of time the reservation was scheduled before the day it was needed. The data is right-skewed, as lead time increased there were more cancellations than non-cancellations. This variable would be beneficial for using a boxcox transformation when engineering features.

# **Questions to Consider**

* How many variables should we include in our initial models?
* What tuning parameters should we include in the GridSearchCV and RandomizedSearchCV?
* Our focus is F1 score for performance but when testing, is there something that captures the performance better?
* Should there be costs associated with misclassifications?

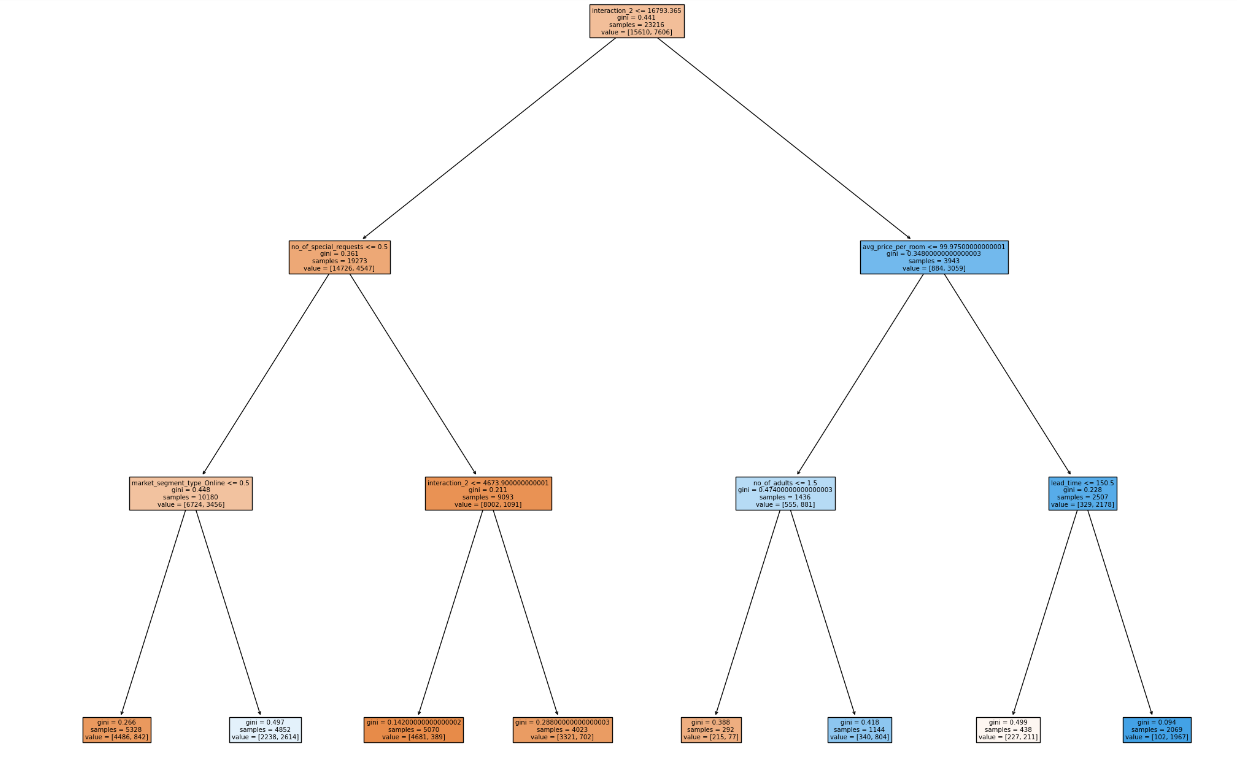
These are the things we will be thinking about and reflecting back to throughout the study

# **Feature Engineering**

To start, we engineered new features that may be important in creating the highest performing predictive model when predicting on booking status. The engineered variable descriptions are below:

|  |  |  |
| --- | --- | --- |
| **Variable:** | **Description:** | **Missing Values:** |
| **interaction\_1** | Lead time multiplied by the number of special requests | 0 |
| **interaction\_2** | Lead time multiplied by the average price per room | 0 |
| **interaction\_3** | Average price per room multiplied by the number of special requests | 0 |
| **interaction\_4** | Taken from a decision tree and setting stipulations for interaction\_2, number of special requests and market segment type online | 0 |
| **lead\_time^2** | Lead time squared | 0 |
| **no\_of\_weekend\_nights^3** | Number of weekend nights cubed | 0 |
| **par\_child** | Number of adults multiplied by the number of children | 0 |
| **diff\_night** | Difference between number of weekend nights and weekday nights | 0 |
| **market\_segment\_type\_Online** | Dummy variable created by a type of market segment in the data set | 0 |
| **market\_segment\_type\_Offline** | Dummy variable created by a type of market segment in the data set | 0 |
| **market\_segment\_type\_Corporate** | Dummy variable created by a type of market segment in the data set | 0 |
| **market\_segment\_type\_Complementary** | Dummy variable created by a type of market segment in the data set | 0 |
| **market\_segment\_type\_Aviation** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 7** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 6** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 5** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 4** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 3** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 2** | Dummy variable created by a type of market segment in the data set | 0 |
| **room\_type\_reserved\_Room\_Type 1** | Dummy variable created by a type of market segment in the data set | 0 |
| **type\_of\_meal\_plan\_Not Selected** | Dummy variable created by a type of market segment in the data set | 0 |
| **type\_of\_meal\_plan\_Meal Plan 3** | Dummy variable created by a type of market segment in the data set | 0 |
| **type\_of\_meal\_plan\_Meal Plan 2** | Dummy variable created by a type of market segment in the data set | 0 |
| **type\_of\_meal\_plan\_Meal Plan 1** | Dummy variable created by a type of market segment in the data set | 0 |

*Figure 5*



*Figure 5* is an example of how we designed one of the interactions. This Decision tree diagram represents interaction\_4. The text is small, but you start by looking to see where most of the data went and by following the black arrows, it shows the path where most of the data is. We use variables in the orange boxes and the stipulations they provide to create the interaction.

### **Feature Selection**

*Figure 6*

Chart

Description automatically generated

*Figure 6 shows the* computed average of top 10 variable appearances using RFECV and XGBoost, Random Forest and Gradient Boost as the estimators. When finding significant variables, we considered the original data as well as the engineered features. Some features were replaced, for example room type became a dummy variable and was split into variables based on requested room. In the final modeling/ensemble stage we used the top nine features to build the models.

# **Hyper Parameter Tuning**

**Grid Search CV**

|  |  |
| --- | --- |
| **Model:** | **Tuning Paramters:** |
| **XGBoost** | 'colsample\_bytree': 1, 'gamma': 0.3, 'learning\_rate': 0.01, 'max\_depth': 7, 'min\_child\_weight': 5, 'n\_estimators': 500, 'subsample': 0.8 |
| **Random Forest** | 'max\_depth': 7, 'min\_samples\_leaf': 7, 'min\_samples\_split': 10, 'n\_estimators': 100 |
| **Gradient Boosting** | 'learning\_rate': 0.01, 'max\_depth': 7, 'min\_samples\_leaf': 7, 'min\_samples\_split': 10, 'n\_estimators': 500 |

**Randomized Search CV**

|  |  |
| --- | --- |
| **Model:** | **Tuning Paramters:** |
| **XGBoost** | 'colsample\_bytree': 1,  'gamma': 0.1,  'learning\_rate': 0.01,  'max\_depth': 7,  'min\_child\_weight': 5,  'n\_estimators': 500,  'subsample': 1, |
| **Random Forest** | 'max\_depth': 7  'min\_samples\_leaf': 7,  'min\_samples\_split': 10,  ' n\_estimators': 300, |
| **Gradient Boosting** | 'learning\_rate': 0.01,  'max\_depth': 7,  'min\_samples\_leaf': 7,  'min\_samples\_split': 10,  'n\_estimators': 500 |

**Optuna**

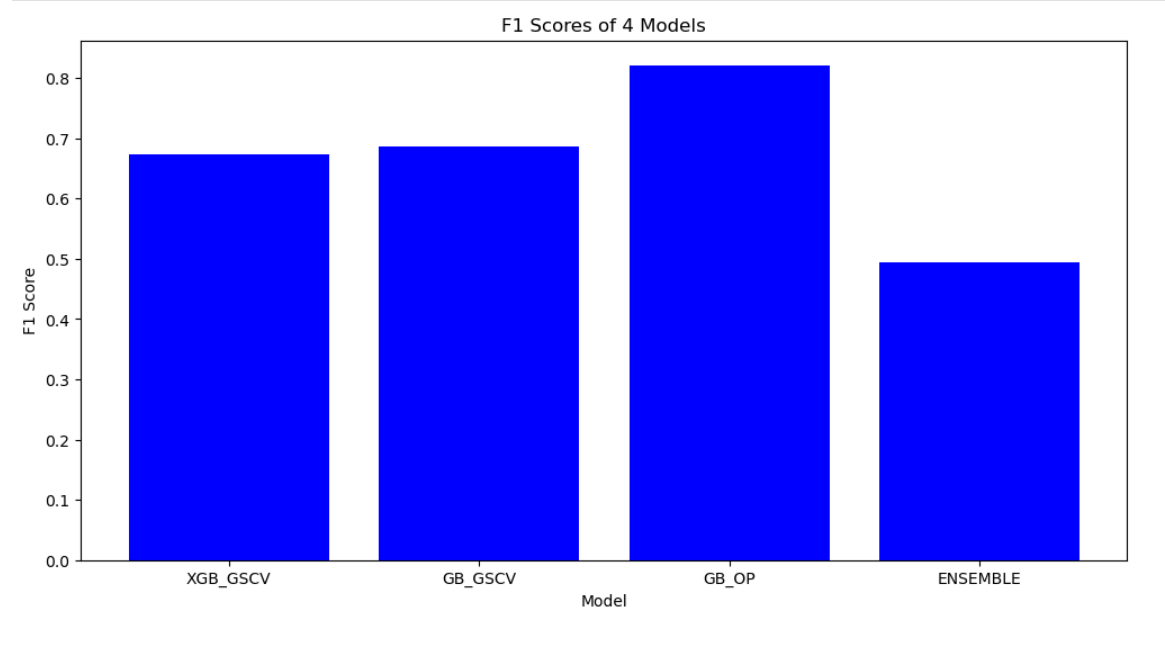
|  |  |
| --- | --- |
| **Model:** | **Tuning Paramters:** |
| **XGBoost** | 'n\_estimators': 503, 'min\_child\_weight': 5, 'learning\_rate': 0.041860223951528844, 'gamma': 0.15715189881593766, 'subsample': 1, 'colsample\_bytree': 1, 'max\_depth': 10 |
| **Random Forest** | 'n\_estimators': 1761, 'min\_samples\_split': 9, 'min\_samples\_leaf': 17, 'max\_depth': 10 |
| **Gradient Boosting** | 'n\_estimators': 1566, 'min\_samples\_split': 13, 'min\_samples\_leaf': 5, 'max\_depth': 10 |

Using the nine significant variables, hyper-parameter dictionaries and the XGBoost, Random Forest, and Optuna frameworks we found the above hyper parameters to be optimal for each of the model types using F1 score to evaluate results. We included Now we plan to use these hyper parameters to build the final models for predicting on the testing set, and eventually to build an ensemble learner.

# **Predictive Modeling/Ensemble Method**

We built XGBoost, Random Forest Classifier, and Gradient Boosting models. With these models we predicted on the validation data set and will use these combined likelihoods to create an ensemble. We will then use this ensemble to predict on the testing data set. All results are shown below:

*Figure 7*



We used the top three preforming models to create an ensemble model. Based on the above chart we see that we have F1 scores varying from lower .4 values to upwards of .83 values. This value is seen with the Gradient boosting model done in Optuna, therefore this is our best model overall. We expected a better performance out of the ensemble model. We ran it multiple times and made changes and it still didn’t improve. There are many reasons this could happen, such as overfitting, poor hyperparameter selection and even performance metric. Maybe choosing F1 score wasn’t the best approach. Even the other six models ran had poor performances ranging in the .60’s. We would need to go into deeper studies with the data to test these ideas.

# **Conclusion**

Because of the money lost from hotel reservation cancellations, the possibility of having a machine learning model to predict the likelihood of cancellations is huge. Throughout this project we followed the process of exploratory analysis, variable engineering, hyper-parameter tuning, and creating varying models to determine reoccurring patterns in the cancellations data set, in hopes of creating a model that will successfully flag potential reservation cancellations. Our best model used the Gradient Boosting Classifier and Optuna framework to predict the variable booking status, it had an F1 score of .82. The next top two models had F1 scores of .61 and .68, quite a big difference which allows for improvement! We limited ourselves to one performance metric due to technological and Wi-Fi issues which could also influence our outcomes. We weren’t sure how to evaluate cost since there’s only two possibilities and we struggled getting our computations to compute so we were unable to reach that point.

# **Implications and Further Questions**

One of our first findings was when we were doing feature engineering, we wanted to use the BoxCox transformation on the variable lead\_time but we realized that it contained zeros and the logarithmic function used in the transformation is not defined for these values. So, we decided to engineer a feature from that variable by squaring it to see if it provided any significance for our models and status on booking. Throughout the modeling process we ran into an issue with how our y\_train variable was being stored so we had to use an encoder to change the way it was stored so it could be executed. We would also test out other performance measures to compare to the F1 score. One thing that is interesting is the number of variables that showed up 100% when looking at importance. Neither of us had seen this happen before so we think this would be beneficial to check out. As mentioned above, we aren’t satisfied with overall performance but if we had more time to experiment with the variables and methods, we could try to increase the performance of predicting if an individual will cancel their hotel room.

Works Cited

Hertzfeld, Esther. “Study: Cancellation Rate at 40% as Otas Push Free Change Policy.” *Hotel Management*, 23 Apr. 2019, https://www.hotelmanagement.net/tech/study-cancelation-rate-at-40-as-otas-push-free-change-policy.

Raza, Ahsan. “Hotel Reservations Dataset.” *Kaggle*, 4 Jan. 2023, https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset.