Predictive Analysis of Hotel Booking Data

Deepak Rao Nadipelly—Sriram Hariharan Neelakantan—Suteja Bhimashankar Patil—Pareekshit Reddy Gaddam

1 Summary

Data analytics in the hotel industry is key to marketing strategy, building customer loyalty and enhancing productivity. It enables hotels to personalize experiences for their guests, introduce better hotel pricing strategies and expand their customer base.Also, it helps customers in analysing which is the best season to travel to a particular location,what are the services included in booking,is the hotel ready to customize their experience based on their special requests and when and where can they find a room with the lowest rate. The dataset we are using contains booking information for a city hotel and a resort hotel over the period 2015-2017. The dataset consists of over 100k records and contains fields such as date of booking, lead time, duration of stay, country, the status of the booking, etc.

Our project goals are :

1. To analyze the best time of year to book a hotel room and the optimal length of stay to get the best daily rate.
2. To classify customers based on who is more likely to cancel their booking.
3. To predict the number of future guests for both hotel types.
4. Predicting the likelihood of a disproportionately high number of special requests.

We leveraged the relationship between the features that had a major impact on the overall detailed analysis and prediction. Our findings will help the hotel industry understand factors affecting customer decisions. We have used various linear and gradient boosting models for our prediction tasks and we compared these models using different metrics.

2 Methods

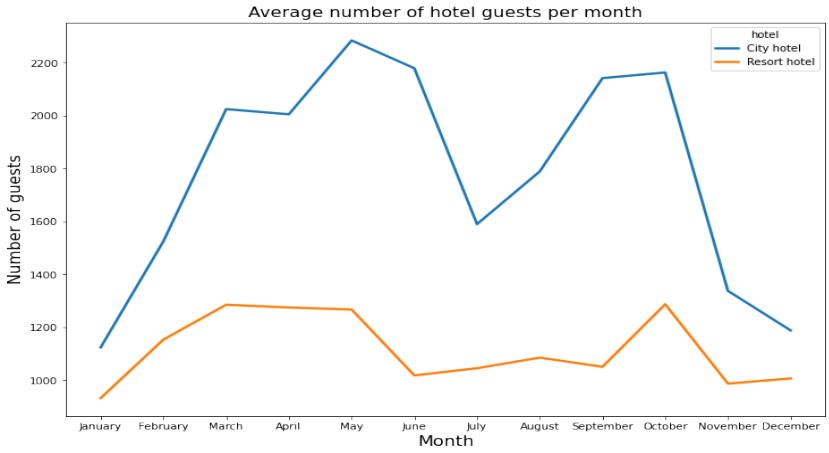
2.1 Data Pre-processing

We took our dataset from kaggle and loaded it into a dataframe pandas library. The dataset consisted of 32 features which included both numerical and categorical columns. 4 out of these 32 columns had null values present. Those columns were 'company','agent','country' and 'children'. The column 'company' consisted of around 94% of null values and 'agent' had around 13 percent of null values. We decided to drop these columns as they would not be useful for our further analysis. The null values in the 'country' column were dropped and the null values in 'children' were imputed with zeroes.Some of the columns had incorrect data types. So,we converted those columns to the correct data type. We also created new columns from the existing columns. Total number of guests was a variable created by adding the number of children and adults and total length of stay was computed by adding values from stays in the weekend and stays during the week. Arrival day, month,year were combined to create a new feature arrival date. We also performed label encoding on categorical columns to convert them to numeric.

2.2 Exploratory Data Analysis

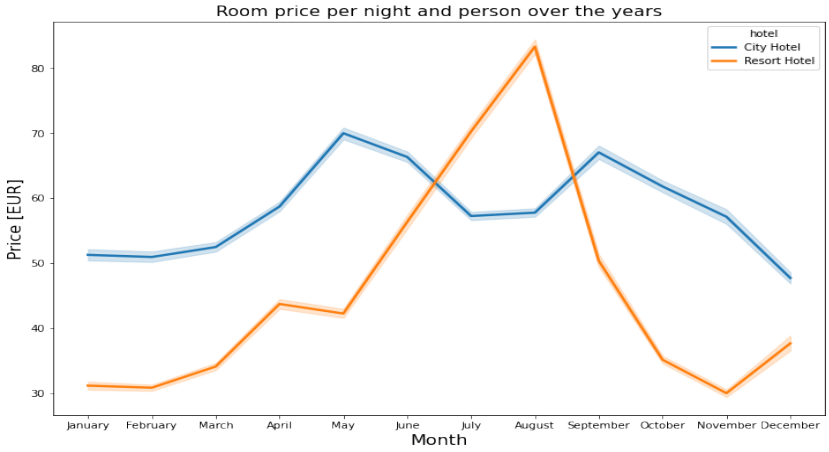
We performed data analysis to maximize insights and understand our data better. It exposed us to different trends and relationships among the variables.

1. What is the average number of guests per month and which season is the most preferred?



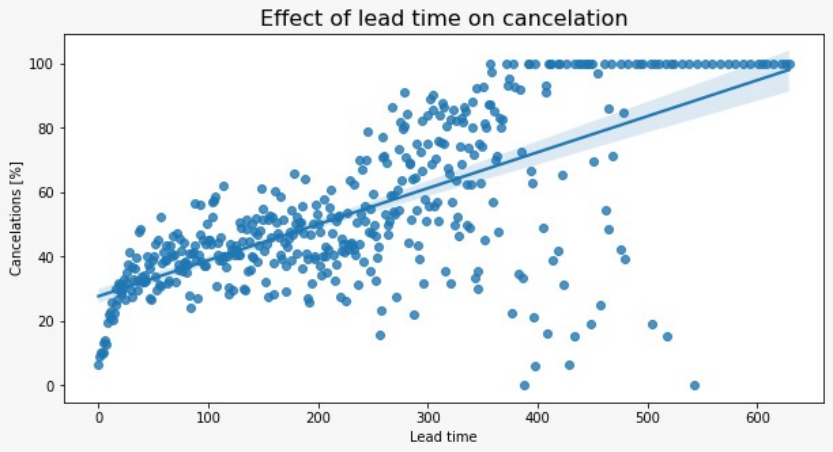
We can interpret from the above graph that April,May, September and October experienced the highest number of guests. So the end of summer and autumn are the preferred seasons by customers for booking.

2. What time of the year has the lowest booking rates?



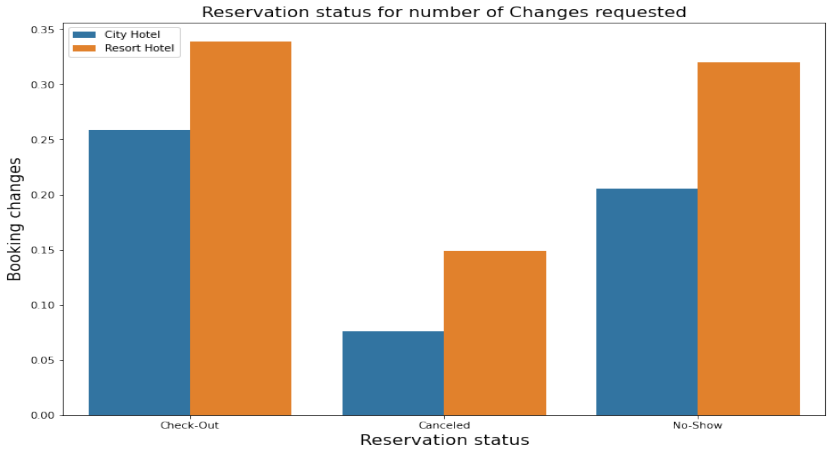
August has the highest booking rate while April,May and from September till November the rate goes on decreasing. So from the above two graphs we can say that the average number of guests are higher during April,May,September and October because of low room charges.

3. Does higher lead time result in cancellations?



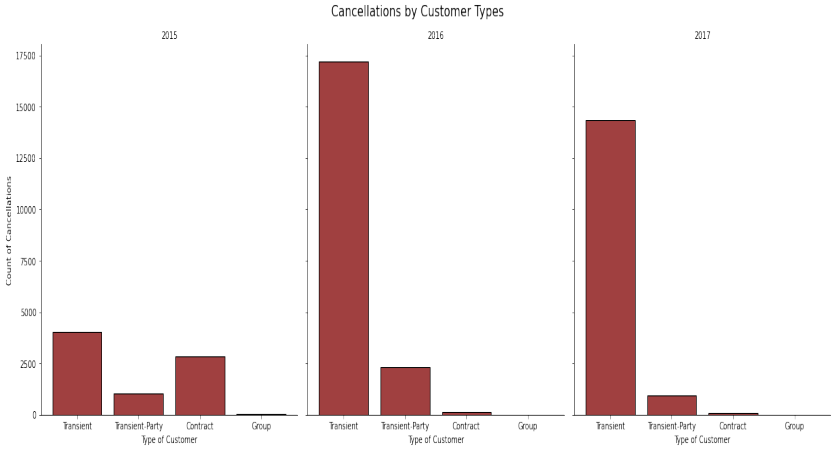
Lead time is the number of days between booking a room and actual check-in date. So we can see that higher the lead time, higher count of cancellations has been observed. This may be because a larger gap can lead to change in plans or better offers from some other hotel.

4. Do all guests who made changes in booking really check-in ?



Even though the count of guests who requested for changes did check-in and check-out as planned, the cumulative percentage of guests who requested for booking changes and cancelled or did not show up is much higher.

5. Which customers are most likely to cancel their bookings?



Transient and Transient-Party type of customers are most likely to cancel their hotel booking.

2.2.1 Feature Selection

- **Correlation matrix**

  Each row and column of the correlation matrix represents a variable, and each value in this matrix is the correlation coefficient between the variables represented by the corresponding row and column.The value of correlation can be positive or negative. Positive value indicates that the two variables are correlated and the negative value indicates that there is a negative correlation. Values closer to 1 show a high correlation.

- **Feature Importance plot**

  Feature Importance is a score assigned to the individual features of a Machine Learning model that defines how important a feature is to the model's prediction. It can help in feature selection and we can get very useful insights about our data. We can calculate feature importances by using tree based models like RandomForest, Light Gradient boosting machines.

2.3 Modelling

We had three different predictions tasks at hand:

1. Predicting whether a booking made by a customer would be cancelled or not. So this would be a classification problem and we have used Machine learning models like Logistic regression, Support vector Machine , XGBoost and LightGBM.

2. Predicting the likelihood of a disproportionately high number of special requests.

3. Predicting the number of future guests for the hotel. This was a time series forecasting model. So we have used ARIMA(Autoregressive Integrated Moving Average).

Brief description of the models used:

1. **Logistic Regression:**

   In a binary logistic regression model, the dependent variable has two levels. The logistic regression model itself simply models probability of output in terms of input and does not perform classification, though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class and the below the cutoff as the other.

2. **Support Vector Machine:**

   The goal of the support vector machine is to find a hyperplane in an N dimensional space that linearly separates the data points.To classify the points from the two classes, there can be many hyperplanes. The goal is to find the hyperplane that has the maximum margin(the maximum distance between the data points of both classes). Maximizing the margin helps us classify new data points more confidently.Support vectors play a crucial role in maximizing the margin. Support vectors are the data points that are close to the hyperplane and influence the positioning of the hyperplane.

3. **Random Forest:**

   Random forest belongs to the class of supervised learning algorithms. It builds a forest which is an ensemble of decision trees, usually trained with the bagging method. Bagging involves bootstrapping the data and learning multiple methods with each randomly sampled subset. When a random forest is used for a classification task and is presented with a new sample, the final prediction is made by taking most of the predictions coming from each individual decision tree.

4. **eXtreme Gradient Boost(XGBoost):**

   XGBoost is a decision-tree based ensemble model that uses a gradient boosting framework Gradient boosting attempts to accurately predict a target variable by combining the estimates of a set of simpler and weaker models.Some of the features of XGBoost are: Parallelized tree building. Tree pruning using a depth-first approach. Regularization to avoid overfitting.

5. **Light Gradient Boosting Machine:**

   LightGBM is a gradient boosting framework based on decision trees to increase the efficiency of the model and also reduces memory usage.LightGBM adopts two novel techniques Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). With GOSS, LightGBM can train each tree with only a small fraction of the entire dataset. With EFB, LightGBM handles high-dimensional sparse features in an efficient manner. It also supports distributed training with a very low communication cost and fast training on GPUs.

6. **ARIMA(AutoRegressive Integrated Moving Average):**

   ARIMA model belongs to a class of models that explains a given time series based on its own past values i.e., its own lags and the lagged forecast errors. ARIMA models are specified by three terms(p,d,q) where,

   - p is the order of the Auto Regression term.
   - q is the order of the Moving Average term.
   - d is the number of differencing required to make the time series stationary.

3 Results

We have worked on three different prediction tasks. So we divided our results into three parts:

1. Predicting whether a booking made by a customer would be cancelled or not

   The Machine learning algorithms we have used and the hyper parameters we have fit to these models are listed below:

   - **Logistic Regression** -solver='liblinear', penalty = "l2",C=1,max_iter = 100
   - **SVM** -C = 1, kernel="rbf",degree = 3, max_iter = -1
   - **XGBoost** - objective="binary:logistic", random_state=42, eval_metric="auc", n_estimators=100, subsample = 0.8,colsample_bytree= 0.8,learning_rate= 0.09
   - **Light GBM** -learning_rate=0.05,n_estimators=20,num_leaves=31

	Logistic Regression		XGBoost		SVM		Light GBM	
	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	0.814	0.811	0.958	0.957	0.99	0.989	0.855	0.856
F1 Score	0.696	0.694	0.940	0.939	0.89	0.99	0.761	0.765
Recall	0.575	0.575	0.889	0.887	0.99	0.991	0.624	0.630
AUC	0.765	0.763	0.944	0.943	1.0	0.994	0.807	0.810

   From the above table, we can observe that XGBoost performed the best among all other algorithms in predicting whether a booking would be cancelled or not.

2. Predicting the likelihood of a disproportionately high number of special requests.

   Hyperparameters:

   - **Random forest :** n_estimators = 100, max_depth=5
   - **XGBoost :**'subsample': 0.7, 'n_estimators': 100, 'max_depth': 10, 'learning_rate': 0.2, 'colsample_bytree': 0.7, 'colsample_bylevel': 0.4

Metrics	MAE		MSE		RMSE	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
Random Forest Regressor	0.095	0.116	0.018	0.023	0.135	0.152
XGBoost	0.045	0.137	0.0049	0.028	0.07	0.169

3. Predicting the number of future guests for the hotel.

   Hyperparameters:

   ARIMA (order=(2, 1, 0), seasonal_order=(2, 1, 0, 12))

   auto_arima(start_p=1,start_q=1,max_p=4,max_q=4, m=12,start_P=0,seasonal=True,d=1,D=1, trace=True,error_action='ignore',suppress_warnings=True,stepwise=True)

	City Hotels		Resort Hotels	
	Train	Test	Train	Test
MAPE	0.29	0.246	0.35	0.239
RMSE	343.2	433.63	293.3	222.88

4 Discussion

We have divided the impact of the results into two sections:

Impact on Customer:

We have visualized the average price of a room per night in both city and resort hotels.The customer can have a look at the trend of prices for the last 3 years and decide what time of the year would be ideal for booking a hotel room.

Impact on Hotel Industry:

1. We analyzed various factors leading to a cancellation and what type of customers are more likely to cancel. This would help the hotel management in addressing the issues behind these cancellations.
2. We predicted the number of future guests a hotel can expect. This will help the hotel be better prepared for the demand and manage the required resources to ensure customer satisfaction.
3. We predicted the likelihood of a disproportionately higher number of special requests. This will help hotels in better understanding of customer requests and expectations.

5 Future Work

1. In the future, we can implement more algorithms like Catboost,neural networks.
2. We can also use GridSearchCv and Bayesian Optimization for tuning the hyper-parameters of various models.
3. Our findings are based on the data from only three years. With more data in the future, we can predict a better model.
4. We also plan to ensemble complex models like LightGBM and XGBoost.

6 Statement of Contribution

1. **Deepak Rao Nadipelly** - Data Preprocessing, Exploratory Data Analysis, Predicting the number of future guests for the hotel, worked on the report and presentation.
2. **Suteja Bhimashankar Patil** - Exploratory Data Analysis, Analysed the best time of year to book a hotel room and the optimal length of stay to get the best daily rate, worked on the report and presentation.
3. **Sriram Hariharan Neelakantan** - Exploratory Data Analysis, Predicting whether a booking made by a customer would be cancelled or not, worked on the report and presentation.
4. **Pareekshit Reddy Gaddam** - Data preprocessing, Exploratory Data Analysis,Predicting the likelihood of a disproportionately high number of special requests, worked on the report and presentation.

7 References

1. https://www.kaggle.com/jessemostipak/hotel-booking-demand
2. https://pandas.pydata.org/docs/pandas.pdf
3. https://scikit-learn.org/stable/
4. https://numpy.org/doc/1.18/numpy-user.pdf
5. https://seaborn.pydata.org/tutorial.html
6. https://www.overleaf.com/learn/latex/Tutorials

8 Appendix

Our working repository is on GitHub: https://github.com/DeepakNadipelly/DS-5110-