1. Problem Statement

Over the years, hotel booking cancelation rates have been increasing steadily, from 32.9% in 2014 to 39.6% in 2018 (Hertzfeld, 2019). Booking cancellations have a substantial impact in demand-management decisions, and impromptu cancellations can really affect forecasts and revenue management. To circumvent this problem, hotels often implement rigid cancellation policies such as deposits, and overbooking strategies, but they can also have a negative influence on revenue and reputation.

Our goal is to build a binary classification model to accurately predict booking cancellations. With that, hotel managers can take measures to avoid cancellations for bookings with high cancellation probability, such as offering discounts and perks. Another use case is to adjust the discrimination threshold so that the False Positives and False Negatives are around the same, so that overbooking strategies can accurately use up cancelled rooms without major consequences.

2. Evaluation Metric

We used Area Under ROC (AUROC) for our evaluation metric. It describes the discriminative power of a classifier independent of the target class distribution, and also looks at different discrimination thresholds, making it more robust than F1 score.

Usually, AUROC isn't the best for imbalanced datasets because the False Positive Rate (False Positives / All Negative Samples) does not drop much when the number of negative samples is a lot bigger than the positive samples. Hence, AUROC might provide an overly optimistic view of the performance of your classifiers, and Area Under Precision-Recall Curve might be a better evaluation metric. However in our case, since our target class distribution is only 37%-63%, it’s not considered too unbalanced, so we will stick with AUROC.

3. Machine Learning Techniques Used

3.1 Preprocessing

3.1.1 Dealing with null values

Features	# of null values
children	4
country	488
agent	16340
company	112593

For children, we filled the null values with 0.
For country (polytomous variable), we removed all the rows with null values for country.

Agent and company are polytomous variables. Since there are a significant number of samples with null values for agent and company, we encoded them into binary variables instead (is agent/company or not). This should also improve generalization.

3.1.2 Feature Extraction

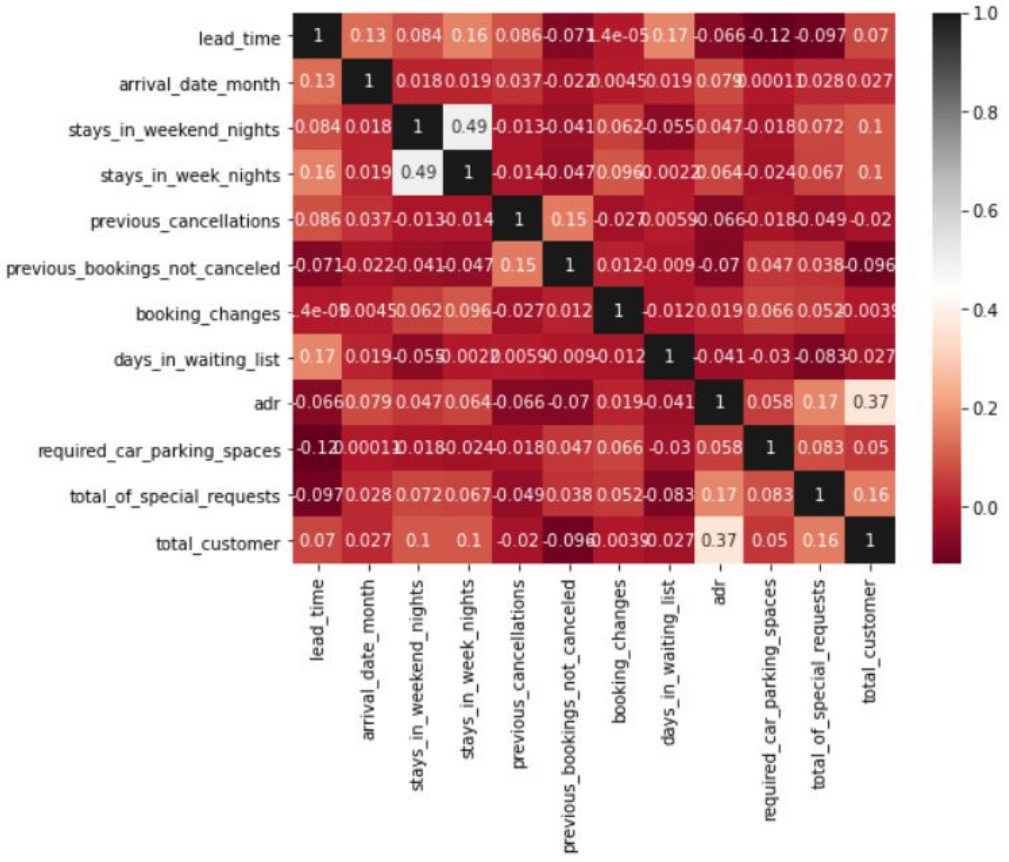
For all the date variables present in the dataset, we decide to drop all of them except the month variable as we think it's likely to be correlated with booking cancellation. For example, seasons, or major events like F1, are key factors to tourists visiting a country at certain times of the year, resulting in different hotel demand (and consequently, booking cancellation) over the year. Dropping unnecessary date variables like year, week, day, should also improve generalization.

For the adult, children, babies variables, we changed them into two different binary variables: is_family and total_customer. Is_family depends if the sample has at least 1 child or baby. Total_customer is the sum of the three variables.

Most important feature to be dropped is the reservation_status and reservation_status date. These are direct substitutes for our target variable is_canceled, and leaving reservation_status in will just lead to a 100% accuracy.

All the remaining categorical variables are encoded using LabelEncoder().

This is the final correlation matrix for the numerical variables at the end of all preprocessing:



As we can see from the matrix, there's little to no collinearity between the final numerical variables. As for the stays in weekend nights and stays in week nights, we initially considered to sum them to one variable total stays in nights, but we felt that some information might be lost there so we kept them as separate variables in the end. Just to note, adr here stands for Average Daily Rate, and is the total lodging transaction divided by total nights stayed.

3.1.3 Train test split

A train_test_split of 20% for test and 80% for training with a fixed random_state of 420.

3.1.4 Scaling and PCA

After scaling the numerical variables, we applied PCA to reduce the number of dimensions, with 11 principal components selected. We ran the models on the different data sets and compared the results.

3.2 Models Used

The problem we are trying to solve here is a binary classification problem. We used 6 different models: Naive Bayes, Logistic Regression, Support Vector Machines, Decision Tree, Random Forest, and XGBoost. We optimised the model using GridSearchCV, which generates models an exhaustive combination with the hyper parameters supplied to it to optimise our model's performance. We then evaluated our models with the AUROC score.

3.2.1 Naive Bayes

The Naive Bayes model uses the naive assumption that all feature probabilities are independent given the class, to estimate the probabilities of each class. Generally, there are 3 commonly-used Naive Bayes classification models - Bernoulli, Multinomial and Gaussian Naive Bayes classifiers. Gaussian Naive Bayes classifier is the most appropriate as a majority of our features take up continuous values.

The following table shows the various metrics and time taken to train the Naive Bayes model:

Data	Accuracy	Precision	Recall	F1-Score	AUROC
Normal Data	0.7297	0.6235	0.6962	0.6579	0.8023
Scaled Data	0.6569	0.5257	0.8226	0.6425	0.8054
PCA Data	0.7273	0.6211	0.6912	0.6543	0.7971

The hyper-parameter for the Gaussian Naive Bayes model was var-smoothing. This parameter was tuned using GridSearchCV and was found to be 1e-06, 1e-05, and 1e-05 for normal data, scaled data and PCA data respectively. This hyper-parameter artificially adds a user-defined value to the distribution's variance for calculation stability, and widens the or smooths the curve.

In general, the model has a very short training time. Despite the short training time, the evaluation metric scores are not very ideal, with the highest AUROC score at 0.8054.

3.2.2 Logistic Regression

Logistic Regression is a predictive analysis algorithm based on the concept of probability. Logistic Regression uses a cost function that can be defined as the 'Sigmoid function' which limits the cost function between 0 and 1.

The following table shows the various metrics to train the Logistic Regression model:

Data	Accuracy	Precision	Recall	F1 Score	AUROC Score
Normal Data	0.7944	0.8136	0.5828	0.6792	0.8650
Scaled Data	0.3834	0.3769	0.9980	0.5472	0.8652
PCA Data	0.7323	0.6872	0.5193	0.5916	0.7779

Based on the AUROC score, the Logistic Regression on the scaled data set was the best performing model with an AUROC score of 0.8655. The best parameters used to acquire this score was a 'max_iter' of 1000, 'solver' set to 'newton-cg' and a 'C' value of 1.

3.2.3 Support Vector Machine (SVM)

SVM is a linear model that attempts to construct a hyperplane which divides the data into the two classes while maximising the margin. The model uses the default hyper parameters and is not optimised using GridSearchCV due to the slow time quadratic complexity $O(m * n^3)$, with m as number of features and n as the number of samples. With a training sample size of 95,512 rows. The team has opted to not optimise SVM due to it's long computation time and as it is less promising results compared to the other models.

The table below shows the performance metrics of the SVM model on the datasets:

Data	Accuracy	Precision	Recall	F1-Score	AUROC
Normal Data	0.7944	0.7797	0.6260	0.6945	0.7944
Scaled Data	0.7857	0.9355	0.4575	0.6145	0.7857
PCA Data	0.7734	0.7409	0.6044	0.6657	0.7734

3.2.4 Decision Tree

Decision tree is a supervised learning model which partitions the data into subsets to classify the data. The decision tree parameter tuning was done using GridSearchCV in which both gini and entropy splitting are considered. The max depths and leaves are run from 2 to 20, 2 to 100 respectively. The resulting best parameters are gini criterion, max depth of 20, max leaf nodes at 99, and minimum sample split is 2.

Data	Accuracy	Precision	Recall	F1-Score	AUROC
Normal Data	0.8432	0.8260	0.7348	0.7778	0.9130
Scaled Data	0.8405	0.7783	0.8009	0.7895	0.8344
PCA Data	0.8146	0.7466	0.7619	0.7541	0.8052

The best performance for this model is on the scaled data with AUROC score of 83.25%.

3.2.5 Random Forest

Random Forest is an ensemble method which combines multiple decision trees to predict the class of a sample by the votes made by the decision trees. The optimal parameters found with GridSearchCV were: criterion='entropy', max_depth=None, min_samples_leaf=1, min_samples_split=2, n_estimators=200. The following table shows the results:

Data	Accuracy	Precision	Recall	F1-Score	AUROC
Normal Data	0.8873	0.8765	0.8124	0.8433	0.9549
Scaled Data	0.8873	0.8764	0.8128	0.8434	0.8471
PCA Data	0.8690	0.8679	0.7656	0.8136	???

3.2.6 XGBoost

XGBoost stands for extreme gradient boosting. It uses a boosted gradient tree to predict labels for the samples. A boosted decision tree trains models with multiple iterations, adding models trained on classification errors of the previous models to improve performance until convergence.

For the optimisation of the model, we utilised grid search to tune the following parameters: min_child_weight which is the minimum weight for a child node in the decision tree, gamma which determines minimum loss for a partition of a node to be considered, subsample which is the subsample ratio of the training instance, colsample_bytree which controls the ratio of features to choose for splitting in the tree, and max_depth terminates the splitting of nodes once it is reached to prevent a tree from overfitting.

The optimal parameters found were: colsample_bytree=1.0, gamma=2,max_depth=5, min_child_weight=1,subsample=0.8 and had the following results in the table below.

Data	Accuracy	Precision	Recall	F1-Score	AUROC
Normal Data	0.8659	0.8453	0.7843	0.8137	0.9420
Scaled Data	0.8659	0.8453	0.7843	0.8137	0.9420
PCA Data	0.8429	0.8265	0.7330	0.7770	0.9182

4. Comparison of Different Models

Model	Random Forest	XGBoost	Decision Tree	Logistic Regression	SVM	Naive Bayes
AUROC Score	0.9555	0.9450	0.9130	0.8652	0.8636	0.8150

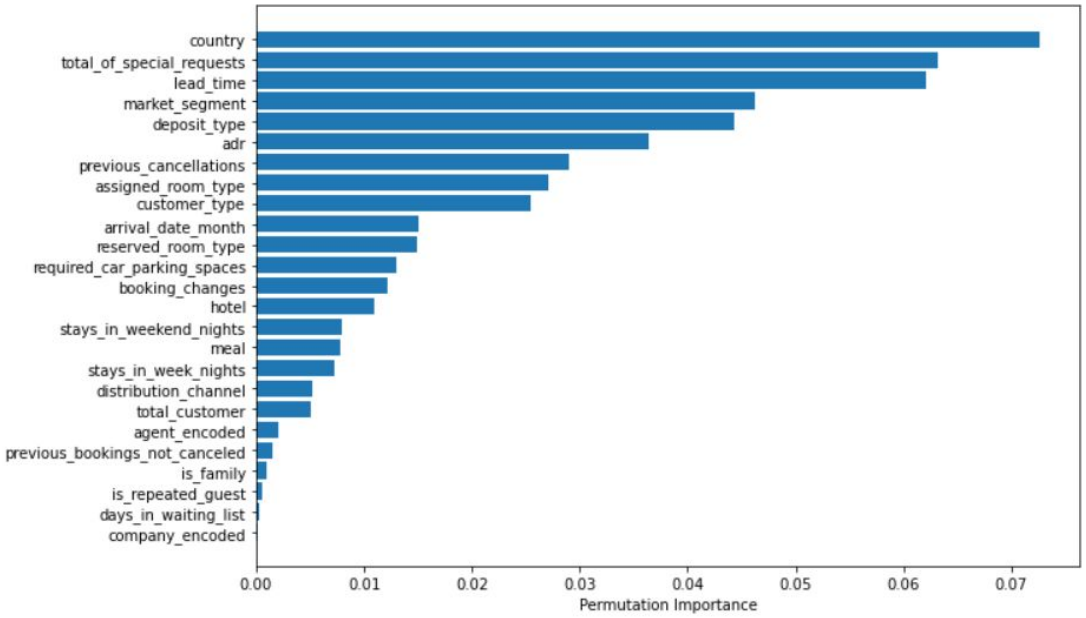
The models that performed the best were models that were based on decision trees like Random Forest, XGboost and Decision Tree. This is followed by Logistic Regression, SVM and Naive Bayes.

Logistic regression and SVM performed worse than the models based on decision trees as there might not have been a single decision hyperplane that clearly split the two classes. Trees, however, use multiple decision rules which could have allowed it to achieve a better performance.

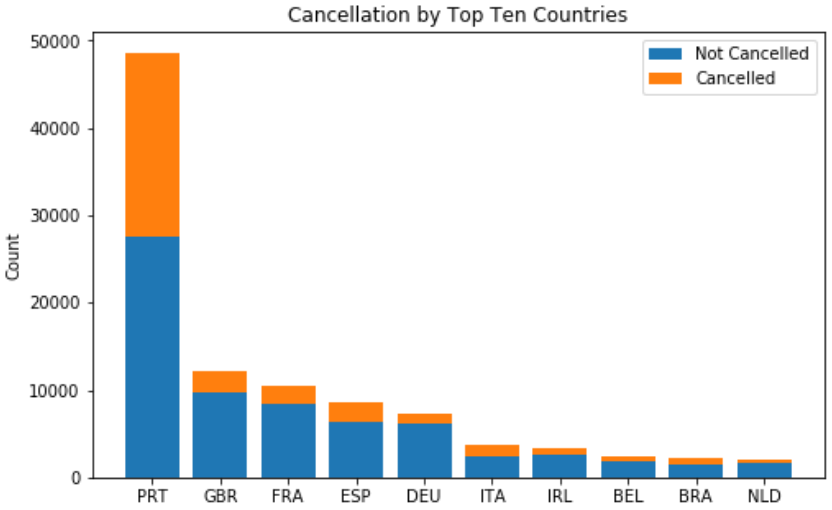
Naive Bayes performed the worst out of all 6 models the team has implemented, it might not have performed well due to the unbalanced nature of the dataset, with only 37% of the samples having a positive label and 63% of the samples having negative labels.

In terms of the performance of the models between the datasets, the models performed better on the normal data as compared to the PCA data. This might have been due to the principal components being selected having a low correlation to the target variable resulting in worse predictions.

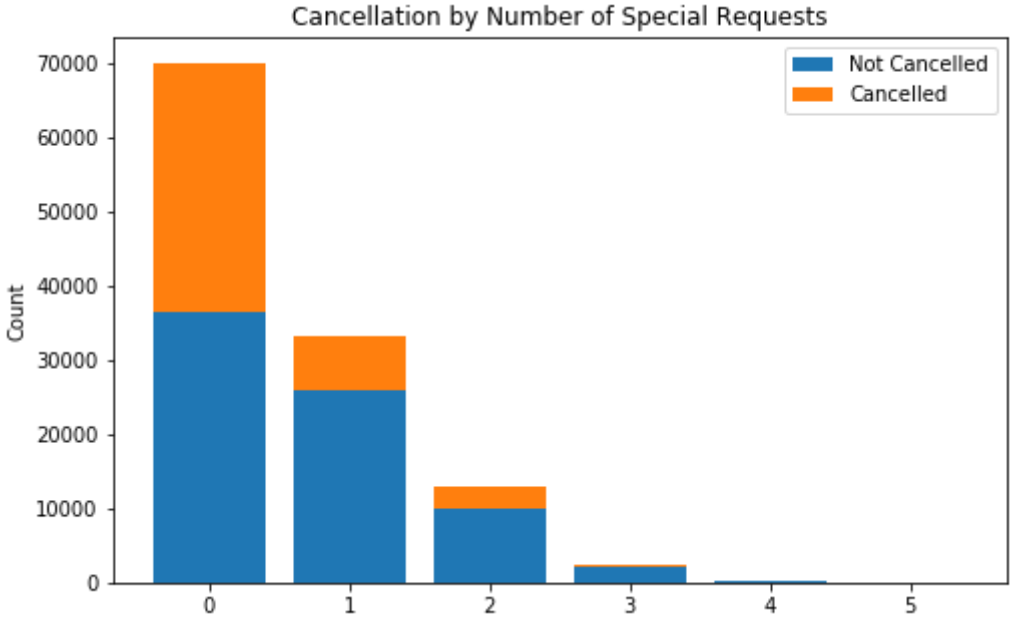
5. Results and Conclusion



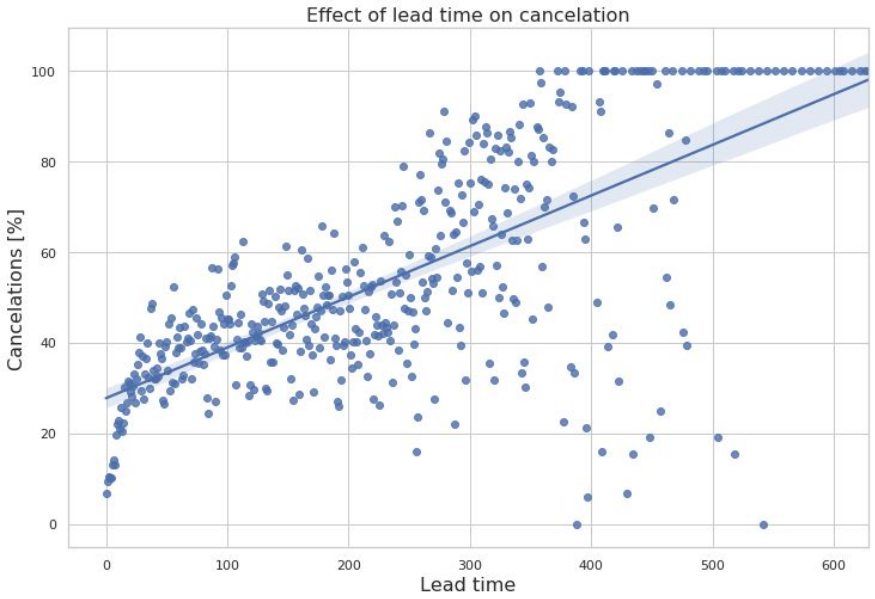
Upon closest examination of the Random Forest model, the three most important features in predicting cancellation were country that the hotel booker is from, total number of special requests made in the booking, and lead time.



When comparing cancellation rates of the top 10 countries with the most bookings, customers from Portugal (PRT) canceled the most, with a much higher cancellation rate as compared to the other countries. With the high number of bookings originating from Portugal, the hotel is likely to be located in Portugal. In this case, the higher number of bookings being canceled might be due to the lower costs lost when locals cancel their bookings and holiday plans without sunk costs such as airplane tickets, etc.



Bookings with more special requests (eg. twin bed, high floor, sea view etc) is less likely to cancel which can be seen in the bar plot above. This might be due to the difficulty in booking another hotel room with the same special requests fulfilled.



Bookings made a few days before the arrival date are usually not cancelled, while bookings made a long time ago before the arrival date have a higher chance of being cancelled. This makes sense because a longer lead time means that a lot of events would have transpired that might cause you to not go for a holiday, thus cancelling the hotel booking.

6. Future Extensions

Extending coverage to other aspects of hotel industry

The current dataset only utilises bookings from resort hotels and city hotels. With Airbnb's market share valued at an estimated $35 billion in 2019 (Statista, 2020), future analysis could include bookings from other players in the market like Airbnb, or even motels.

Using Additional Models

In the future we could also try using Neural networks which could produce better results as compared to the models currently implemented with a large data set. Neural networks are complex models that try to mimic the way human brains develop classification rules, in a way that no other algorithms can. They are, however, unintuitive for interpretation and require expertise to tune and run, which is why we did not use them, but there is potential for it to provide better performance. They also require relatively good computation power to train which the team did not possess.

Additional features to improve model

We could include more useful features such as detailed cancellation policies to understand if the severity of the penalties will affect booking cancellations. Another feature would be the direct competitor hotels in the vicinity to understand if the number of competitors would affect booking cancellations.

7. References

Frank, E., & Bouckaert, R. R. (2006). Naive Bayes for Text Classification with Unbalanced Classes. Lecture Notes in Computer Science Knowledge Discovery in Databases: PKDD 2006, 503-510. doi:10.1007/11871637_49

Hertzfeld, E. (2019, April 23). Study: Cancellation rate at 40% as OTAs push free change policy. Retrieved November 20, 2020, from https://www.hotelmanagement.net/tech/study-cancelation-rate-at-40-as-otas-push-free-change-policy

Mahapatra, S. (2019, January 22). Why Deep Learning over Traditional Machine Learning? Retrieved November 21, 2020, from https://towardsdatascience.com/why-deep-learning-is-needed-over-traditional-machine-learning-1b6a99177063

Statista. (2020, April). Company value of Airbnb from 2016 to 2019(in billion U.S. dollars). Retrieved November 20, 2020, from https://www.statista.com/statistics/339845/company-value-and-equity-funding-of-airbnb/