PHASE 2

PROJECT TITLE: HOTEL BOOKING PREDICTION

Machine Learning and Statistical Analysis

MEMBER 1: Name: Sahithi Bakaram. UBID: sahithib

MEMBER 2: Name: Spoorthy Reddy Avancha UBID: savancha

Overview

Hotel booking cancellations pose a significant challenge for the hotel industry because the number of guests has an impact on the entire operational setup. The goal of this phase is to use machine learning to predict hotel cancellations and analyze which factors have the most influence. Machine learning can be broadly defined as an interdisciplinary science that employs computers to solve a given problem by identifying patterns and learning from existing data. Theory from probability, statistics, optimization, algorithms, and computer science are all used in machine learning. Predicting cancellations is a binary classification problem because there are only two possible outcomes: cancellation or non-cancellation. In statistics, classification is the process of determining what class a given input data belongs to or predicting a qualitative outcome variable.

The data for the thesis was provided by a hotel booking demand dataset where data is from the hotels in Portugal, and the machine learning algorithms used were Random Forest, Decision tree, KNN, Naive Bayes, and Logit. Random forests and Decision trees are tree-based models that generate decision trees to make predictions, which are referred to as classification trees in a classification problem. A classification tree's goal is to determine a qualitative outcome variable through step-by-step binary splits, with the different outcomes denoted as classes. The logit model, also known as logistic regression, is a type of binary regression used as a reference model.

Classification

In statistics, classification is the process of determining what class a given input data belongs to or predicting a qualitative outcome variable. The binary or multiclass qualitative outcome variable can be used. A typical binary classification problem is determining whether or not an individual will be able to repay their loan, i.e., default or not. A multiclass classification problem would be predicting the outcome of a cricket game, where the outcome could be a win, a loss, or a draw.

Here we are working with a binary classification problem in this thesis because a hotel reservation is either canceled or not canceled.

If a model correctly predicts the class of an observation, the predicted class equals the true class; otherwise, a false classification has occurred. This can be stated as follows:

I(yi = ^yi) or I(yi!= ^yi),

where yi represents the correct class for the ith observation.

ˆyi is the predicted class for the ith observation

I(A) is a function that returns one if event A occurs and zero otherwise.

The proportion of misclassifications is referred to as the error rate, and it is calculated as

$\frac{1}{N}\sum_{1}^{N} I$(yi!= ^yi) where N is the sample size.

A prediction model attempting to solve a binary classification problem can make two types of errors: false positives and false negatives.

The false positive ratio is calculated by dividing the number of false positives by the total number of false positives plus true negatives i.e., false positive rate= FP/ (FP + TN)

The false negative ratio is calculated by dividing the number of false negatives by the total number of false negatives plus true positives i.e., false negative rate= FN/ (FN + TP)

where FP denotes the number of observations that were incorrectly classified as positive,

FN the number of observations that were incorrectly classified as negative,

TN the number of observations that were correctly classified as negative, and

TP the number of observations that were correctly classified as positive.

A false positive is also known as a type 1 error, while a false negative is known as a type 2 error.

To calculate the total accuracy of the binary classification model, divide the correct number of classifications by the total number of classifications, which is given as Accuracy=TP+TN/(FN+FP+TP+TN)

Confusion Matrix

A confusion matrix, as shown in the figure below, is a common way to summarize the performance of a model that is attempting to solve a binary classification problem. The figure below shows the correct values along the diagonal from the top left to the lower right, indicating that the predictive values correspond to the true values. The false positives and false negatives are represented by the diagonal from lower left to top right.

Confusion Matrix	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

Confusion Matrix Representation

ROC-curve & AUC

The ROC-curve, or receiver operating characteristic, is a widely used tool for visualizing the potential trade-off between type-1 and type-2 errors, i.e., false positive and false negative, for a classification model. This metric depicts the change in error types with respect to the classifier's threshold T. A binary classification model will output a score or probability for a given observation to belong to class 1, which we call P, and an observation is classified as 1 if P > T, otherwise 0. The threshold, T, is thus the smallest value of P that must be assigned to class 1.

The Area Under the Curve (AUC) is used to summarize the ROC curve and is a measure of a classifier's ability to distinguish between classes. The higher the AUC, the better the model distinguishes between positive and negative classes.

Dataset & Algorithms

The hotel booking demand dataset is used to predict whether a booking is canceled or not based on the input features. The dataset contains 32 columns and 119390 rows in total. After loading the data, we considered the features that are important and even combined some columns now the cleaned dataset has 27 columns and 86639.

After all the data preprocessing and data cleaning is done, the dataset is separated into response and predictor variables, followed by a partition into two parts, where 80% is used to train the model and the rest 20% is used as a test dataset. We fit and transform the training data using StandardScaler() and then transform the test data as well. Finally, the prediction of whether a room is canceled or not is performed based on other independent variables using machine learning algorithms.

Here classification algorithms such as,Naive Bayes, KNN , Logistic regression, Decision tree, and Random Forest are considered as the dataset comes under the classification with the inference of prediction

Using each algorithm, we fit the model with the training dataset, and then predict cancelation or not for the records in the test dataset

1.Naive Bayes classifier

A Naive Bayes classifier is a probabilistic machine learning model for classification tasks. The gist of this classifier is based on the Bayes theorem.

Using the Bayes theorem stated below, we can calculate the likelihood of H occurring given that E has occurred, where E represents the evidence and H represents the hypothesis.

$P(H|E) = (P(E|H) * P(H)) / P(E)$

Where,

$P(H)$: The probability of hypothesis H being true. This is known as the prior probability.

$P(E)$: The probability of the evidence.

P(H|E): The probability of hypothesis H given the evidence E. This is known as the posterior probability.

P(E|H): The probability of the evidence E given that hypothesis H is true.

In this case, the predictors/features are assumed to be independent. In other words, the presence of one feature has no effect on the presence of another. As a result, it is referred to as naive.

Choosing the hypothesis with the highest probability after calculating the posterior probability for a number of different hypotheses is the most likely hypothesis, also known as the maximum a posteriori (MAP) hypothesis written as:

$$MAP(H) = max(P(H|E) = max((P(E|H) * P(H)) / P(E))= max(P(E|H) * P(H))$$

P(E) is a normalizing term that allows us to compute the probability. We can disregard it when we are only interested in the most likely hypothesis because it is constant and is only used to normalize.

Returning to classification, if each class in our training data has an even number of instances, the probability of each class (e.g., P(h)) will be equal. Again, this would be a constant term in our equation that we could remove to arrive at:

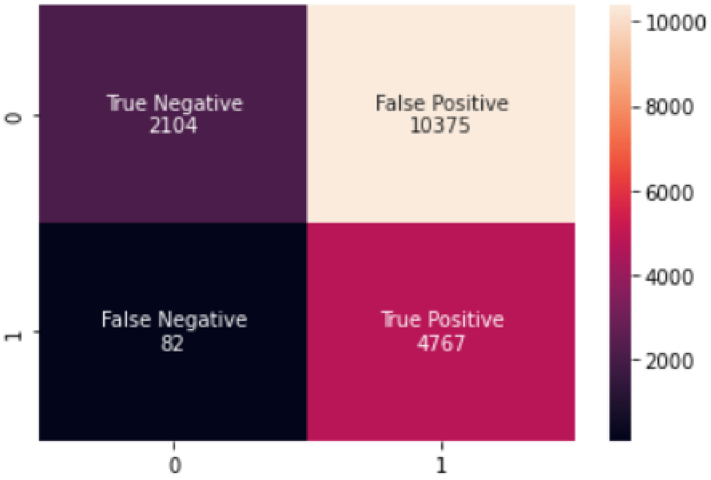
$$MAP(H) = max(P(E|H))$$

Naive Bayes is a classification algorithm for binary classes and multiple-class classification problems. The technique is easier to understand when described using binary or categorical inputs. Assuming a Gaussian distribution, Naive Bayes can be extended to attributes with real values This Naive Bayes extension is called Gaussian Naive Bayes. You can use other functions to estimate the distribution of the data, but the Gaussian (or normal) distribution is the easiest to work with because you only need to estimate the mean and standard deviation of the training data.

Learning a naive Bayes model from training data is fast because you only need to calculate the probability of each class and the probability of each class with different input values (x). It is not necessary to apply coefficients through optimization procedures.

Model Evaluation

In this section results are given when using Naive Bayes classifier as a model for predicting booking cancellations on our hotel data. The confusion matrix displays the result of the Naive Bayes classifier evaluated on test-data. The rows are the actual classes, and the columns are the predicted classes, where 1 represents a cancellation and 0 a non-cancellation.



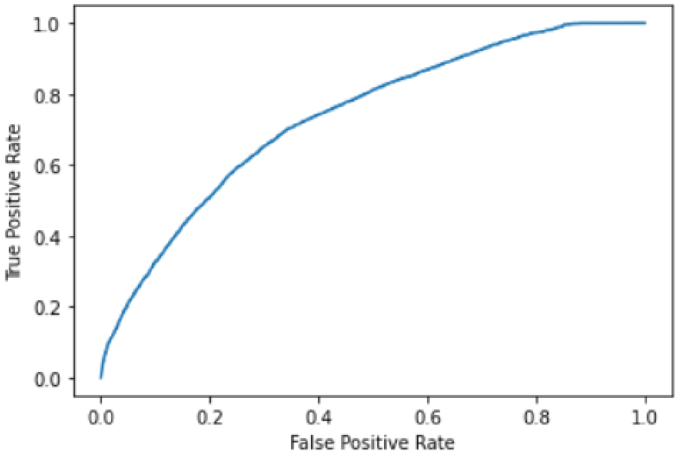
Confusion matrix for Naive Bayes classifier

2104 bookings were correctly predicted as not canceled(is_canceled=0) (true negatives) whereas 10375 were wrongly categorized as canceled when their status is not canceled (false positives). 4767 bookings were correctly predicted as canceled(is_canceled=1) (true positives) whereas 82 were wrongly categorized as not canceled when their status is canceled (false negatives).

- Total accuracy of the model is 39.65%.

The optimal value of the threshold varies depending on the model and the data used. Therefore, AUC is a better metric to evaluate and compare models.

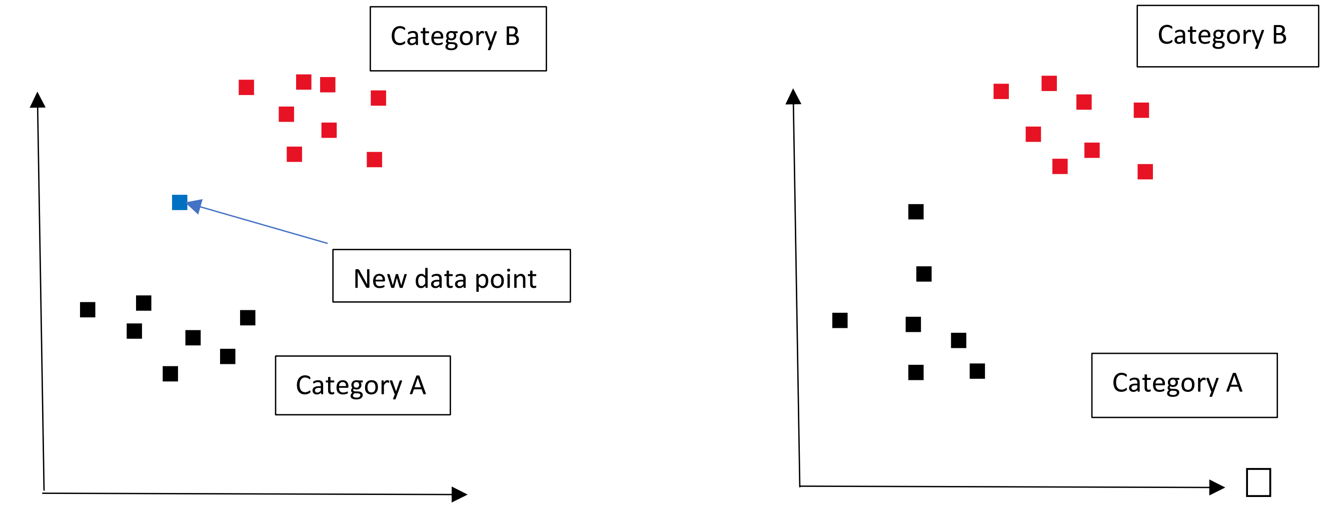
The figure below displays the ROC-curve for the Naive Bayes classifier model applied to test-data from the hotel data. The AUC for Naive Bayes classifier model is 0.73



ROC-curve for the Naive Bayes classifier

2.K-Nearest Neighbour(KNN)

KNN is a supervised machine learning algorithm that deals with the classification of two regions. So, this model will be used in the prediction of hotel cancellations. We use the KNN algorithm to classify the new data into a category that is similar to the new data that is inserted. Consider the two categories A & B.



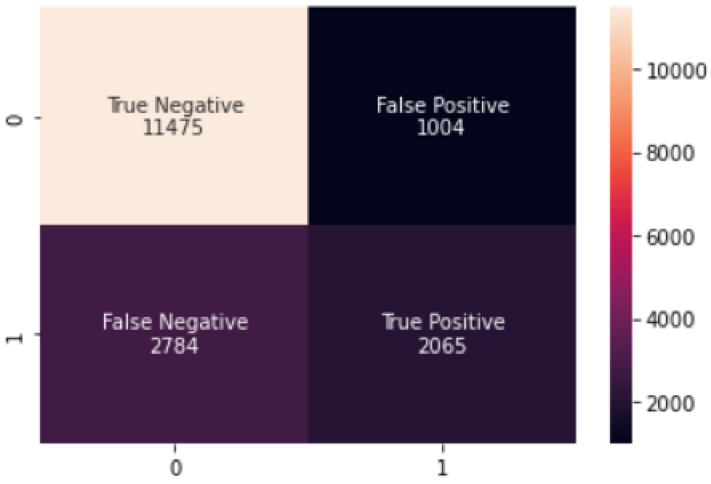
We have the new data point that should be inserted either in one of the categories A or B. The new data point is assigned to the category A.

The K-NN algorithm works as follows:

- First, we chose then number of neighbors which is the value K
- Then, calculate the Euclidean distance from the new data point to the K neighbors of data points
- After calculation of Euclidean distance, we get the nearest neighbors. Among all these, we count the data points in each category
- Finally, we assign the new data point to that category that has the highest number of neighbors.

Model Evaluation

The results of using the K-Nearest Neighbor classifier as a model for predicting booking cancellations on our hotel data are presented in this section. The confusion matrix displays the performance of the K-Nearest Neighbor classifier on test data. The rows represent the actual classes, and the columns represent the predicted classes, with 1 representing a cancellation and 0 representing a non-cancellation.



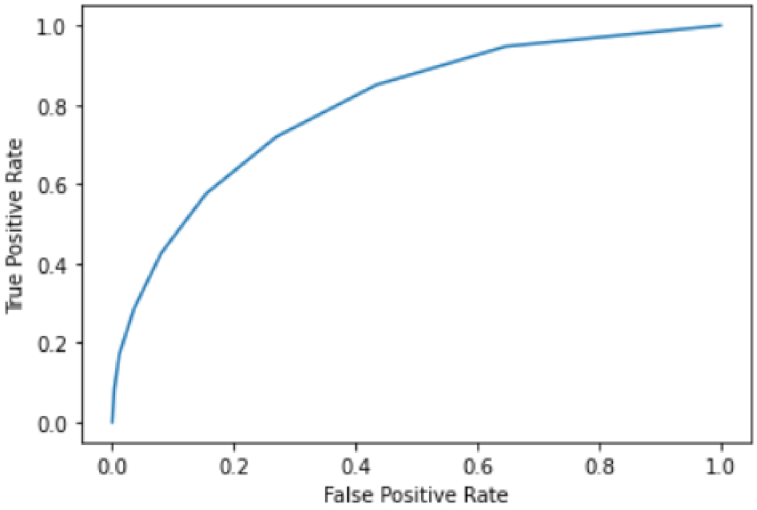
Confusion matrix for K-Nearest Neighbor

11475 bookings were correctly predicted as not canceled(is_canceled=0) (true negatives) whereas 1004 were wrongly categorized as canceled when their status is not canceled (false positives). 2065 bookings were correctly predicted as canceled(is_canceled=1) (true positives) whereas 2784were wrongly categorized as not canceled when their status is canceled (false negatives).

- Total accuracy of the model is 78.13%.

The optimal threshold value varies depending on the model and data used. As a result, AUC is a better metric for evaluating and comparing models.

The ROC-curve for the K-Nearest Neighbour model applied to hotel test data is shown in the figure below. The K-Nearest Neighbour classifier model has an AUC of 0.79.



ROC-curve for the K-Nearest Neighbor classifier

3.Logistic Regression

Logistic regression is a statistical approach for calculating the probabilities of an event vs alternative scenarios. It performs better than Naïve Bayes classification. In this, we calculate the conditional probability of each observation to determine as to which category it belongs. This model is based on the assumption that the dependent variable must be categorized. In this, we have S-shaped curved which describes the values 0-1.

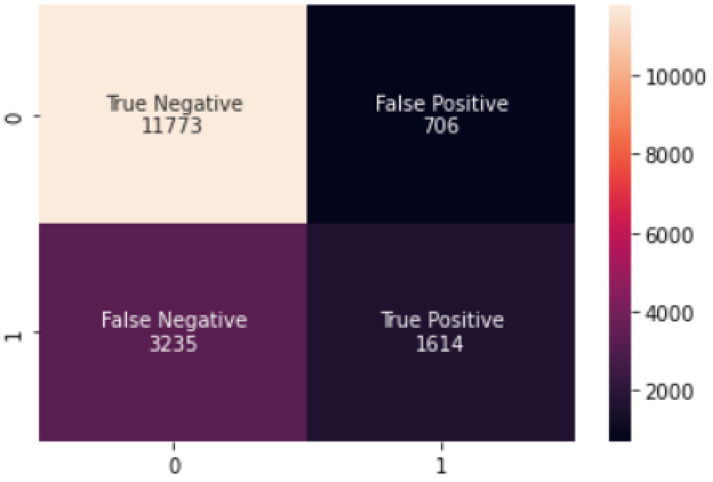
Logistic regression is based on the idea that, for a binary outcome, assign each observation a conditional probability as to whether the observation belongs to a specific category, implying that the probability is dependent on the observation's value for the explanatory variables. Because the result of logistic regression is a probability, the defined range of values is [0, 1]. As a result, the form has been modified to meet the criterion of being defined only for the set of values [0, 1]. (James et al. 2013).

Let Y be a random variable that typically represents the dependent variable, p(X) = P[Y = 1|X], and X = (X1, X2, Xi) be the independent variables. The main idea in logistic regression is to model p(X) as follows:

$$p(X) = \frac{e\hat{}\beta0+\beta1X1\cdots+\beta iXi}{1 + e\hat{}\beta0+\beta1X1\cdots+\beta iXi}$$

Model Evaluation

The results of using the Logistic regression as a model for predicting booking cancellations on our hotel data are presented in this section. The confusion matrix displays the performance of the Logistic regression on test data. The rows represent the actual classes, and the columns represent the predicted classes, with 1 representing a cancellation and 0 representing a non-cancellation.



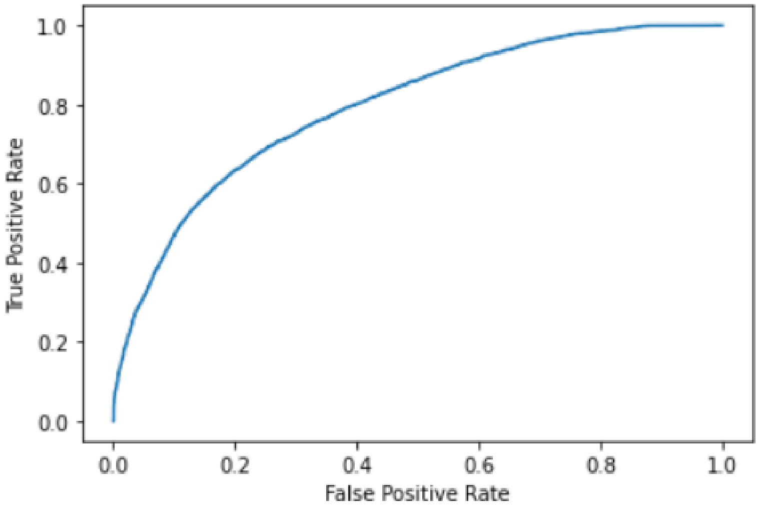
Confusion matrix for Logistic regression model

11773 bookings were correctly predicted as not canceled(is_canceled=0) (true negatives) whereas 706 were wrongly categorized as canceled when their status is not canceled (false positives). 1614 bookings were correctly predicted as canceled(is_canceled=1) (true positives) whereas 3235 were wrongly categorized as not canceled when their status is canceled (false negatives).

- Total accuracy of the model is 77.25%.

The optimal threshold value varies depending on the model and data used. As a result, AUC is a better metric for evaluating and comparing models.

The ROC-curve for the Logistic regression model applied to hotel test data is shown in the figure below. The Logistic regression model has an AUC of 0.794.



ROC-curve for the Logistic regression model

We here remark that ROC curve for Logistic Regression are created via the condition ˆp(X) > T, where the threshold T ranges between 0 and 1 and ˆp(X) is constructed where X is a vector representing the independent variables.

4.Decision tree:

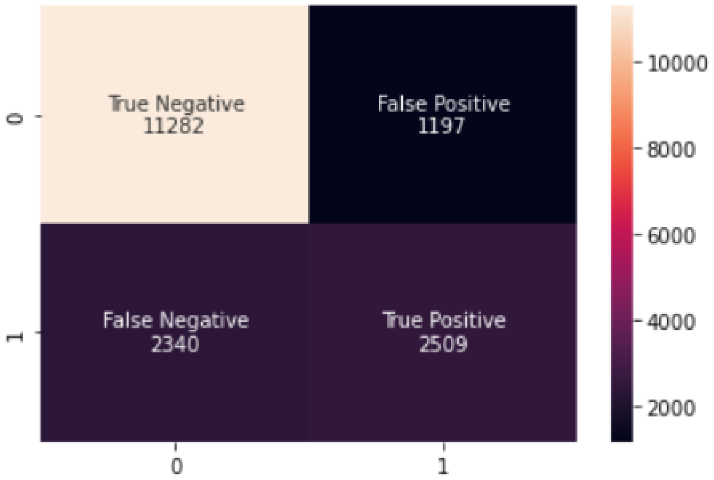
Decision tree is supervised modelling technique used in classification and prediction. Decision tree has a tree-like structure which is important to find the logic behind the decision tree. Decision tree algorithm starts by considering a reference which is titled to be the root node. This root node is now compared with the dataset and proceeds further until the end of the branch. It then jumps to the sub- nodes where the process repeats till it reaches the leaf node. The root node in the decision tree is split into sub trees based on the answer by asking a question (Yes or No).

The tree majorly depends on the following entities i.e., decision nodes, leaf nodes and root node. In the decision tree, the data is continuously split at the decision nodes based on the specific parameters. Final outcomes can be leaf nodes. Decision trees can be of various types:

- Classification Trees – categorical
- Regression Tress – continuous

Model Evaluation

The results of using the Decision tree as a model for predicting booking cancellations on our hotel data are presented in this section. The confusion matrix displays the performance of the Decision tree on test data. The rows represent the actual classes, and the columns represent the predicted classes, with 1 representing a cancellation and 0 representing a non-cancellation.



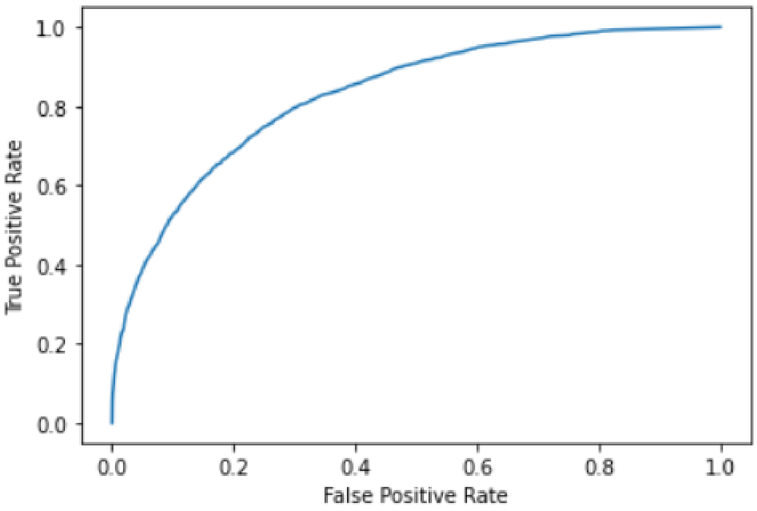
Confusion matrix for Decision tree

11282 bookings were correctly predicted as not canceled(is_canceled=0) (true negatives) whereas 1197 were wrongly categorized as canceled when their status is not canceled (false positives). 2509 bookings were correctly predicted as canceled(is_canceled=1) (true positives) whereas 2340 were wrongly categorized as not canceled when their status is canceled (false negatives).

• Total accuracy of the model is 79.58%.

The optimal threshold value varies depending on the model and data used. As a result, AUC is a better metric for evaluating and comparing models.

The ROC-curve for the Decision tree model applied to hotel test data is shown in the figure below. The Decision tree classifier model has an AUC of 0.82.



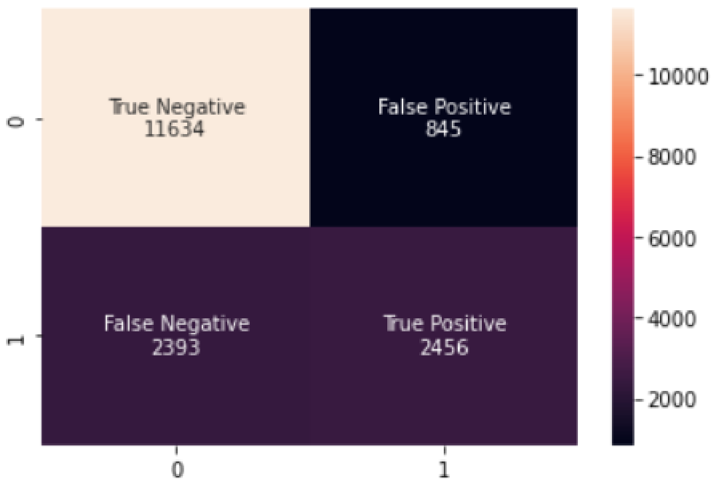
ROC-curve for the Decision tree

5.Random Forest

Random Forest Algorithm is based on the ensemble learning which to solves a problem by combining many classifiers. Random Forest Algorithm is a supervised machine learning technique which is used to solve classification and regression problems. Decision trees are the basic units in the random forest. The random forest contains multiple decision trees and each decision tree contains root node, decision nodes, leaf nodes. In classification, we chose different parameters randomly and split the root node. The leaf nodes of each decision tree represent the output of that tree. The final output is predicted by majority voting system. In Regression, the result of random forest is by predicting the mean of output of each tree using Bagging technique. As the number of trees increases, the precision increases. We use this algorithm as it prevents overfitting and performs better than decision tree algorithm. Thus, it has high precision with high accuracy. This algorithm requires less training time than other algorithms

Model Evaluation

The results of using the Random Forest Algorithm as a model for predicting booking cancellations on our hotel data are presented in this section. The confusion matrix displays the performance of the Random Forest Algorithm on test data. The rows represent the actual classes, and the columns represent the predicted classes, with 1 representing a cancellation and 0 representing a non-cancellation.



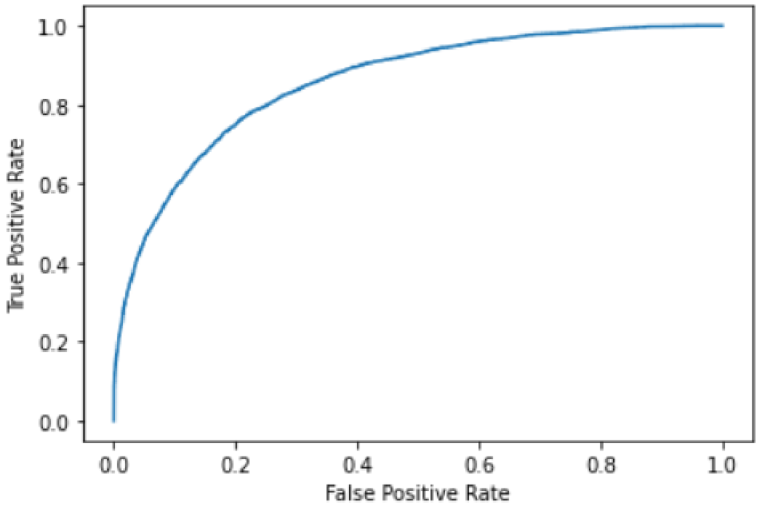
Confusion matrix for Random Forest model

11634 bookings were correctly predicted as not canceled(is_canceled=0) (true negatives) whereas 845 were wrongly categorized as canceled when their status is not canceled (false positives). 2456 bookings were correctly predicted as canceled(is_canceled=1) (true positives) whereas 2393 were wrongly categorized as not canceled when their status is canceled (false negatives).

• Total accuracy of the model is 81.31%.

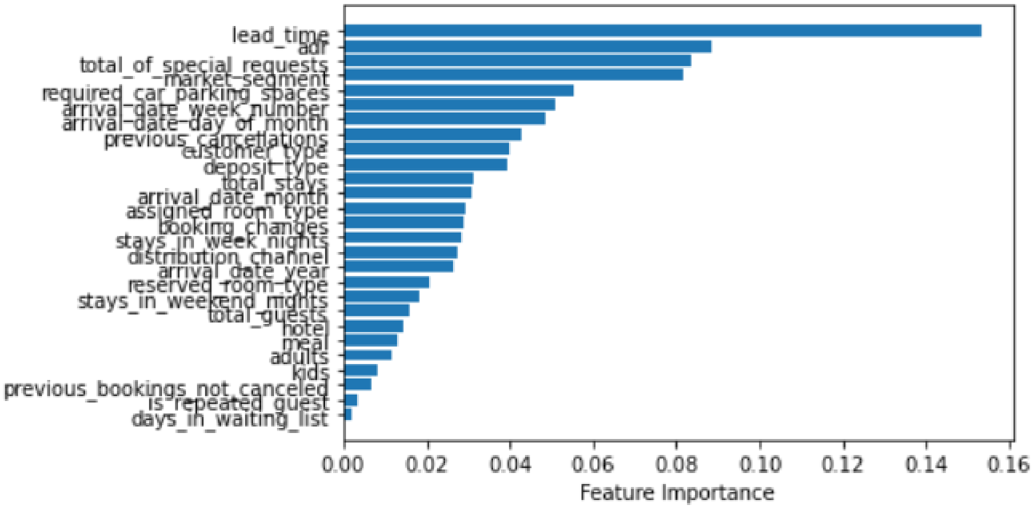
The optimal threshold value varies depending on the model and data used. As a result, AUC is a better metric for evaluating and comparing models.

The ROC-curve for the Random Forest Algorithm applied to hotel test data is shown in the figure below. The Random Forest Algorithm has an AUC of 0.85.



ROC-curve for the Random Forest model

We here remark that ROC-curve for Random Forest are created via the condition Cˆ(x) > T where the threshold T ranges between 0 and 1 and Cˆ(x) is constructed and x is a vector representing the independent variables.



Feature importance through the random forest algorithm

The above figure gives the feature importance for the Random Forest model based on the Gini criteria. Feature importance do not indicate if a feature have a negative or positive impact on the outcome variable, it only show how much information each variable contains in order to determine which class an observation belongs to. In Figure Leadtime is by far the most dominant variable, with the rest of the variables.

Models Comparison

Navie Bayes:

It does not necessitate as much training data. It can work with both continuous and discrete data. The number of predictors and data points is highly scalable. It is quick and can make real-time predictions.So here in this regard, if the goal is to identify cancellations rather than maximize overall accuracy, one could argue that the Naive Bayes model is superior. It should be noted, however, that increasing recall only works up to a point. If recall was 100%, then all bookings could be classified as a cancellation, which provides no insight into the differences between customers who cancel and those who do not.

Logistic Regression:

For binary and linear classification problems, logistic regression is a simpler and more efficient method. It is a classification model that is simple to implement and achieves excellent results with linearly separable classes. So for other algorithms to learn it is a basic algorithm and here the considered dataset has the binary classification predictor.

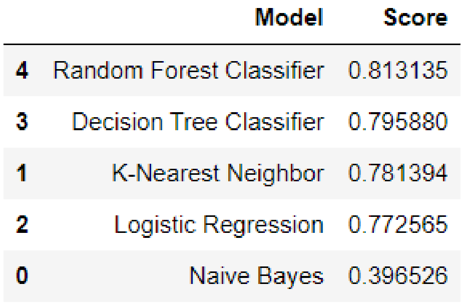
KNN

KNN classifies the new data points based on the similarity measure of the earlier stored data points. So if the data points are given we can classify whether it is canceled or not based on the nearest points given some of the features.

Classification trees

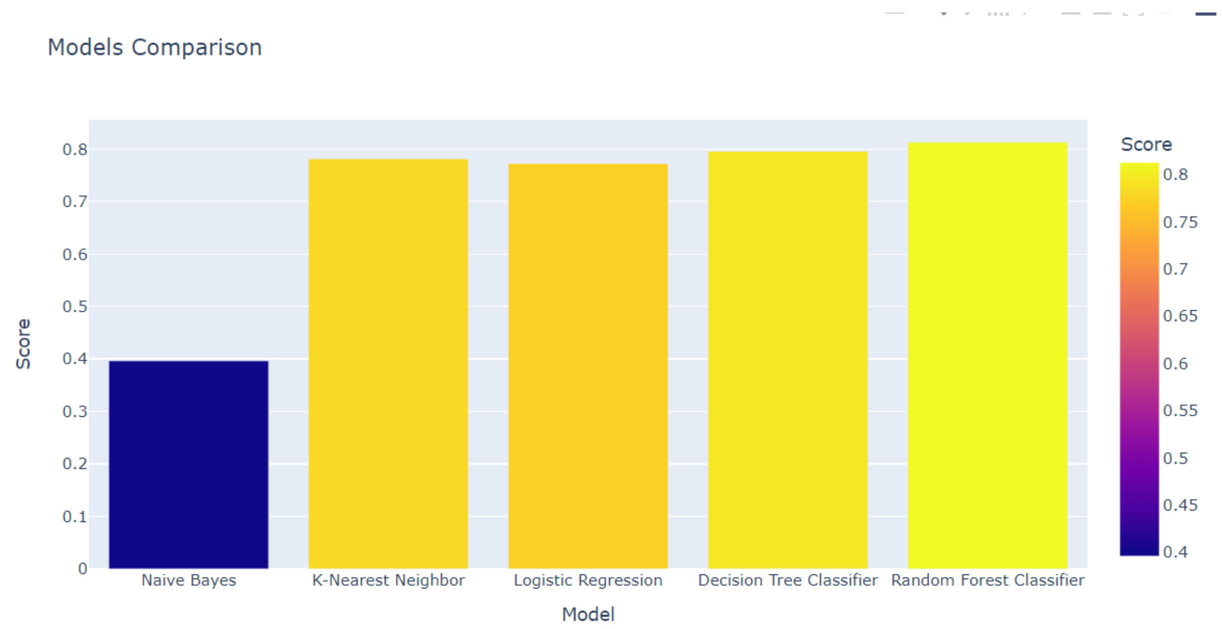
Tree-based methods can be used to solve classification problems with the large number of records and understand easily. The algorithm behind tree-based methods is very intuitive, which makes the tree that is being built easy to understand and interpret.As this dataset has large number of records nearly 1lakh these algorithms are used.

The following table shows a summary of the results for the five models. Random Forest outperforms in regard to accuracy. Furthermore, all models show higher False-negative rates compared to False-positive rates. Note that the FP-rate and FN rate depends on the threshold, which in this case is 0.5. Therefore, a more robust way of comparing the performance between the models is the AUC score. The AUC score shows that the Random Forest model is the best performing model and navie bayes is not suitable for this data.



	Model	Score
4	Random Forest Classifier	0.813135
3	Decision Tree Classifier	0.795880
1	K-Nearest Neighbor	0.781394
2	Logistic Regression	0.772565
0	Naive Bayes	0.396526

Accuracy of all models considered



Analysis

Cancellations for hotel guests be predicted with the use of machine learning, based on the given data

The findings of this study suggest that machine learning can predict hotel cancellations based on a given dataset from a hotel in. Because all models except navie bayes have a higher accuracy, this suggests that using statistical tools and machine-learning algorithms allows hotel guests to predict cancellations. Furthermore, in terms of accuracy and AUC, the two tree-based models Random Forest and decision tree outperform the logistic regression. This lends credence to the notion that machine learning, particularly tree-based models, are not only applicable but also a good choice for predicting cancellations in given data.

According to the results, all models have higher false negative rates than false positive rates, implying that identifying cancellations may be more difficult than identifying non-cancellations. However, changing the prediction threshold can alter the relationship between false positives and false negatives. When evaluating and comparing models, the AUC is a more relevant criterion than the false positive and false negative rate. Based on accuracy and AUC, it is clear that Random Forest is the model with the highest score, followed by Decision Tree, KNN and Logistic regression, and finally Navie Bayes.

As a result, both AUC and accuracy indicate that tree-based models are suitable for predicting cancellations for this data set and these model choices. It is worth repeating that the findings of this study are only applicable to the hotel in question. However, because the explanatory variables used in the models are assumed to be available for most hotels, this lends support to the notion that machine-learning models can be used to predict cancellations in the hotel industry.

Factors that are most influential when predicting cancellations

Random Forest is generated using the Bootstrap and Baggning methods, and the Gini criteria are used to determine feature importance. Determine which factors are critical for each model. The importance of each feature is only indicated, not whether it has a negative or positive effect on the outcome variable. As a result, the following discussion of how the variables affect hotel cancellations should be regarded as speculative. According to the study's findings, leadtime is the variable with the most information to predict hotel cancellations in the Random Forest model. The days between when the hotel reservation is made and the day of arrival are represented by the numeric variable leadtime. It is obvious why leadtime contains so much information.

It is obvious why leadtime contains so much information. If a person makes a hotel reservation with a specific arrival date in mind, the person has many options for canceling the reservation. Another consideration is that it may be easier to cancel an event if it is scheduled for a long time in the future rather than sooner. At the same time, a reservation booked on short notice, i.e. with a low leadtime, may be more likely to be cancelled due to a sudden change of plans. Sundays and different room categories were two other variables that were relatively important in Random Forest.

It is difficult to understand why especially Sundays would contain more information than other weekdays in determining cancellations. However, different kinds of hotel rooms could be correlated with external events and therefore help the model to predict. If an individual books a specific type of hotel room, for example a suite, it is reasonable to assume this person is there for a specific reason, either to celebrate something or attending a specific event. If the assumption is correct, the individual is less likely to cancel the hotel reservation because he/she is also attending an external event

References

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

https://towardsdatascience.com/understanding-random-forest-58381e0602d2

https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052

https://scikit-learn.org/stable/modules/tree.html