Predicting the reservation of a hotel is canceled on not

Abdalla Alnujaidy – Faisal Alasgah

Abstract:

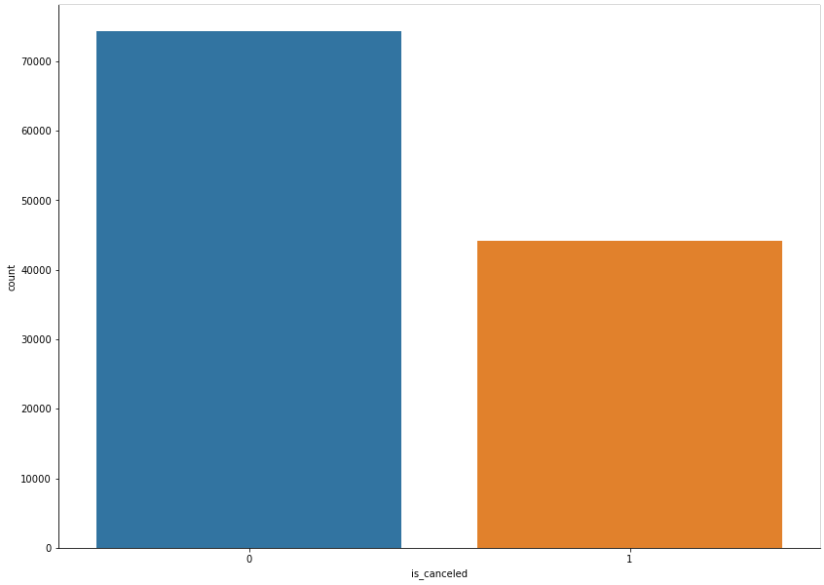
Our goal for this Classification Model is to predict the individuals that are going to cancel their hotel booking, we will take the (canceled or not) as our dependent value and the number of adults, children and babies, Country, repeated customer, number of nights and more as our independent value. This model will help us identify which type of people cancel their booking more, we got our data form Kaggle, we used a lot of libraries to clean, visualize and analyze to come up with the best model possible.

Data:

Hotel booking data contained around 119 thousand rows of data and 32 features, some of our main features were:

- Type of hotels
- Price
- Reservation status
- Number of guests
- Date
- Lead time
- Reservation canceled or not

After cleaning the data there were around 118 thousand rows of data and for the model features it been reduced to 28 features.



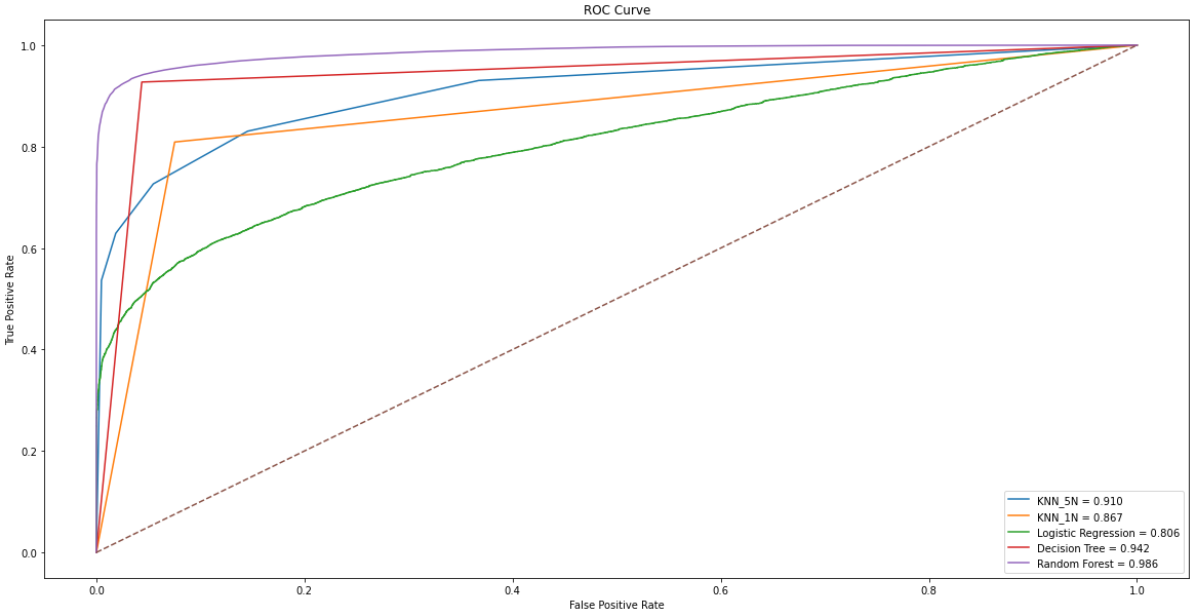
Algorithms and Models:

For feature engineering we convert some columns to binary variables. As long the data is balanced we did use four different algorithms for modeling Knn, Logistic Regression, Decision Tree and Random Forest. Also, we did cross validation to find out the best number of neighbors for Knn which was 1. After compering the result for all models, it looks Random Forest is the best between them.

	KNN_5	KNN_1
Accuracy	86.404	88.178
Precision	88.688	86.393
Recall	72.653	80.906
F1_score	79.873	83.559

	Logistic regression	Decision tree	Random forest
Accuracy	79.140	94.644	95.100
Precision	84.684	92.850	98.684
Recall	53.500	92.717	87.977
F1_score	65.572	92.783	93.023

ROC:



Tools:

- Pandas – Numpy for data cleaning and manipulation
- Matplotlib – Seaborn – Plotly – Folium for visualization
- Sklearn for building our models