PREDICTING BOOKING CANCELLACTION

Rehab Alanazi

Abstract

The purpose of this project is to build a classification model that helps to predict the possibility of canceling a hotel booking, it will help decision makers in the hotel industry to take these actions at the right time with the right customers. I worked with hotel booking demand data provided by Kaggle website.

Design

This project helps to figure out which feature has impact on cancellation booking and predict whether a customer with a specific feature may cancel booking or not, it will be useful for hotel managers to know the probability of cancellation for each customer, this will allow them to make actions to increase the probability of booking confirmation.
The data provided by Kaggle website for first phase, and I started to explore the dataset, then I visualize the data to help identify the features that led the customer to stay at the hotel or cancel the booking, after that I build three models to predict then I evaluated each model.

Data

The dataset contains 119390 data points with 32 features, 14 of which are categorical. A few feature highlights include lead time, previous cancellations, adults, total of special requests, previous bookings not canceled, is repeated guest and is cancel.

Algorithms

Models

- Logistic Regression
- Decision Tree
- Random Forest

Model Evaluation and Selection

The entire dataset of 47506 records was split into 70 train /30 test.

Below is the evaluation of each model

Logistic Regression scores:

- Accuracy 58%
- precision 0.57
- recall 0.65

Decision Trees cores:

- Accuracy 78 %
- precision 0.78
- recall 0.77

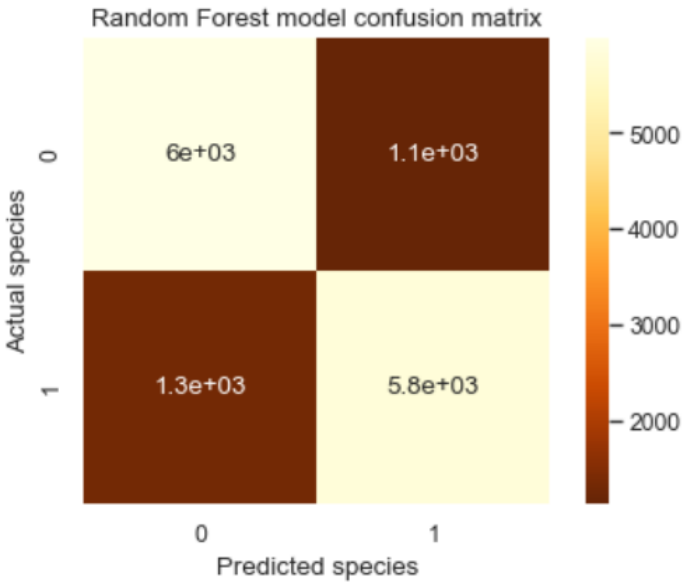
Random Forest scores:

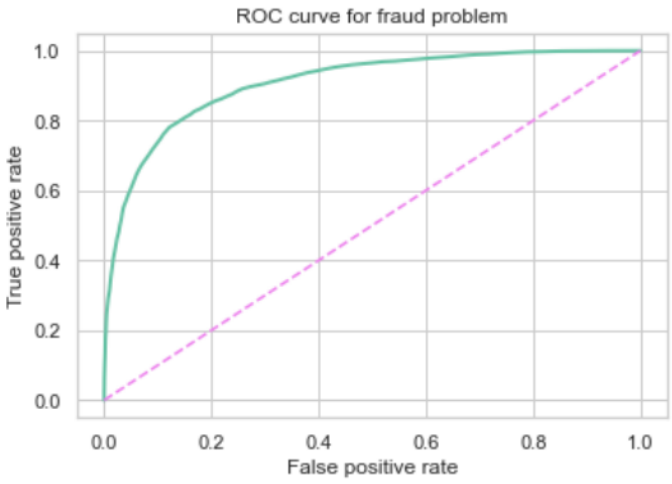
- Accuracy 83 %
- precision 0.83
- recall 0.82

Tools

- Python Language
- Pandas, NumPy for data processing
- Scikit-learn library for building the model
- Matplotlib for visualization.

Communication


Random Forest model confusion matrix
6e+03
1.1e+03
1.3e+03
5.8e+03
Actual species
0
1
0
1
Predicted species
5000
4000
3000
2000


ROC curve for fraud problem
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0