Machine Learning Engineer Nanodegree

Florian Sckade

April 10th, 2020

Capstone Proposal: Prediction of Hotel Booking Cancellations

Domain Background

Predicting churn is a very important part of every subscription based business, which is often addressed by modern machine learning solutions. Prominent examples are often found in telecommunications (Huang, Kechadi, and Buckley 2012). The general problem formulation can be transferred to other domains, for example hospitality management. Here, churn translates to cancellations of bookings. Since this translates very directly to a loss of revenue, the accurate prediction of cancellations is of very high importance for such businesses.

Along with accurate predictions, it can be of major importance to identify possible drivers of cancellations, so that possible opportunities for actions can be derived from a machine learning solution. As an example, a prediction model could generate next best offers for customers at the point of booking, possibly reducing the probability of a later cancellation. The dataset used for this project will be the Hotel booking demand dataset (Mostipak 2020), hosted on kaggle.com, originally published in Antonio, Almeida, and Nunes (2019).

Problem Statement

For hospitality management, the accurate prediction of a cancellations prior to the anticipated check-in date is very important. In this context, the problem to be solved is the prediction of a cancellation probability $P(Y_i|X_i)$ of a customer i, given a set of features X_i. Thus, the problem is a two-class classification problem, allowing the performance of solutions to be quantitatively evaluated by evaluation metrics like accuracy, precision or recall. An implementation should be able to accurately and repeatedly classify customers based on the given features.

Datasets and Inputs

The dataset used for this project is taken from Mostipak (2020), which itself is a republishing of Antonio, Almeida, and Nunes (2019). It holds records of over 130,000 hotel bookings for two different hotels - one city hotel and one resort hotel - from Portugal. Each row represents a booking that is either cancelled or not and described by 31 additional variables. The bookings in the dataset are due to arrive between the 1st of July 2015 and the 31st of August 2017. As such, the data contains multiple holiday seasons and plenty observations.

The booking status being cancelled or arrived is the target variable of interest. The other variables in the dataset will be used as inputs for a classification algorithm. Since they consist of a mixture of numeric and categorical variables, they allow for further feature engineering, possibly allowing for incremental improvements of an utilized classification model.

Solution Statement

The solution considered in this project for the prediction of booking cancellations is a classification model for individual customer's hotel bookings. The goal of such a solution is the accurate prediction of a cancellation given the customer's features. Thus, the goal is easily quantifiable by classification evaluation metrics and replicable for different customers, given the considered features as present.

Benchmark Model

There are multiple possible benchmarks that a solution could be evaluated again. The simplest benchmark would be assigning labels randomly over a uniform probability distribution for both classes. Another simple benchmark could be assigning the label "no-cancellation" per default. Additionally, one could estimate a simple model on a reduced set of features to set a fitted benchmark. In the project, the starting benchmark will be randomly sampled classes from equal probabilities for the different classes.

Evaluation Metrics

The project's solution will be evaluated against the area-under-the-ROC-curve (AUC) score, as well as the accuracy. The ROC-curve is the model's recall against the false positive rate, which is equal to $(1 - Specificity)$ at various threshold settings. The AUC is calculated for the ROC-curve to give a total score for the model. Additionally, different classification measures and their importance and implication in the context of the underlying business case will be discussed.

Following Fawcett (2006), the accuracy is defined as $ACC = \frac{\text{True Positives + True Negatives}}{\text{Positive + Negatives}}$, where the nominators are equivalent to correctly classifying positive and negative observations respectively, while the denominator's variables are the numbers of real positive and negative observations in the data. Furthermore, the recall is defined as $Recall = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$ and the specificity is defined as $Specificity = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$.

Project Design

After motivating the project as a relevant case study in the introduction, the available data will be explored. For this purpose, the data will be analyzed in regards to missing values and possible wrong observations. Afterwards, the different features of the observations will be explored. This includes general summary statistics, as well as a closer look at the distributions dependent on the target variable. After examining the base features available, possible extensions of the feature space via feature engineering will be discussed. Afterwards, every categorical feature will be made numerical, so that the resulting table object can be used by any modelling algorithm.

Given the feature table and data split into training and test sets, a starter model will be trained on the training data, while cross validating. Results will be saved and evaluated against a model further tuned in regards to its hyperparameters, before evaluating the final models on the test data. Since the data is already in a tabular format, we will be using `xgboost` as modelling class.

After evaluating the results, the resulting model will be examined closer in regards to its feature importance, so that prospects for possible business results can be discussed, before concluding with the project.

References

Antonio, Nuno, Ana de Almeida, and Luis Nunes. 2019. "Hotel Booking Demand Datasets." *Data in Brief* 22: 41–49.

Fawcett, Tom. 2006. "An Introduction to Roc Analysis." *Pattern Recognition Letters* 27 (8): 861–74.

Huang, Bingquan, Mohand Tahar Kechadi, and Brian Buckley. 2012. "Customer Churn Prediction in Telecommunications." *Expert Systems with Applications* 39 (1): 1414–25.

Mostipak, Jesse. 2020. "Hotel Booking Demand." 2020. https://www.kaggle.com/jessemostipak/hotel-booking-demand.