Machine Learning Engineer Nanodegree

Florian Sckade

April 17th, 2020

Capstone Project: Prediction of Hotel Booking Cancellations

I. Definition

Project Overview

This project seeks to adress a major point of planning efforts in hospitality management: the prediction of booking cancellations. Within the project, historical data of cancelled and checked-in hotel bookings is explored for interesting relationships and utilized in machine learning models to predict whether a customer will cancel their booking.

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). It contains booking information from two different portugese hotels, one based in a city, the other being a resort hotel. The focus will lie on the target variable - which describes whether a booking was cancelled before the customer arrived or not. Additionally, the dataset includes information about the booking and its associated customers - like the number of adults and children, over which travel agent the booking was made, the date of the booking or the average daily rate for the customer.

A hotel needs to take future cancellations into account when they allow a customer to book a room. Due to this, a hotel tends to be overbooked by default, building upon the assumption that some customers will not actually arrive. If a hotel predicts cancellations inaccurately, by overestimating the actual number of canceled bookings, customers would be need to be turned away upon their arrival. Similarily, if the predicted number is undererstimating the actual number of cancellations, the hotel could operate on too little capacity and lose money. As such, the prediction of cancellations is a forecasting problem with high business impact.

Problem Statement

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.

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. These solutions are two-class classification models. An implementation should be able to accurately and repeatedly classify customers based on the given features. 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.

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

$$\text{Accuracy} = \frac{\text{True Positives } + \text{ True Negatives}}{\text{Positive } + \text{ Negatives}}$$

, 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. Recall and specificity, which are needed for calculating the ROC-AUC score, are defined as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives } + \text{ False Negatives}},$$

and

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

respectively. Again, *True Positives* denote the number of correctly classified positive cases, i.e. correctly predicted booking cancellations, while *True Negatives* denote the number of correctly classified negative cases, i.e. customers who were predicted not to cancel their bookings and did not cancel. *False Negatives* and *False Positives* are wrongly classified negative and positive cases respectively.

For this problem, it is also worth considering precision, defined as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positvies}}$$

. This is important, because it is valid to assume cancellations to occur much less frequently in booking data than actual check-ins. Where accuracy allows for measurement of the model's ability to seperate the two classes, precision tells us how well the model is at predicting the less-often occuring class.

II. Analysis

(approx. 2-4 pages)

Data Exploration

In this section, you will be expected to analyze the data you are using for the problem. This data can either be in the form of a dataset (or datasets), input data (or input files), or even an environment. The type of data should be thoroughly described and, if possible, have basic statistics and information presented (such as discussion of input features or defining characteristics about the input or environment). Any abnormalities or interesting qualities about the data that may need to be addressed have been identified (such as features that need to be transformed or the possibility of outliers). Questions to ask yourself when writing this section:
- *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset?*

Table 1: Overview of the categorical variables present in the dataset

Variable	Type	Description
hotel	categorical	Type of hotel, H1 = Resort Hotel or H2 = City Hotel
arrival_date_month	categorical	Month of arrival date
meal	categorical	Type of meal booked
country	categorical	Country of origin
market_segment	categorical	Market segment designation
distribution_channel	categorical	Booking distribution channel
reserved_room_type	categorical	Code of room type reserved
assigned_room_type	categorical	Code of room type assigned due to hotel operation reasons
deposit_type	categorical	Indication on if the customer made a deposit to guarantee the booking
agent	categorical	ID of the travel agency that made the booking
company	categorical	ID of the company that made the booking
customer_type	categorical	Type of booking, one of: Contract, Group, Transient, Transient-party
reservation_status	categorical	Reservation last status, one of: Canceled, Check-Out, No-Show
reservation_status_date	categorical	Date at which the last status was set

Has a data sample been provided to the reader? - If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed? - If a dataset is ***not*** *present for this problem, has discussion been made about the input space or input data for your problem? - Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

The dataset utilized is a publicly available data set hosted on Kaggle (Mostipak 2020), originally published in Antonio, Almeida, and Nunes (2019). The original data was cleaned and prepared by Mock and Bichat (2020). Since the data is a result of an exercise in data cleaning and manipulation, only minimal preprocessing is needed to enable predictive modelling with this data. There are no real missing values in the dataset that contain actually missing information. Instead, empty values indicate the non-existence of the specific attribute in that case. For example, the variable `children` has some null values. However, that does not mean there are actually missings for that variable. Instead, this simply implies that there are zero children associated with a booking. Overall, this means there is not a lot of data cleaning needed to start working with the presented data. Nonetheless, the data will be explored further in the following section.

To start off, the data includes numerical and categorical data.

Exploratory Visualization

In this section, you will need to provide some form of visualization that summarizes or extracts a relevant characteristic or feature about the data. The visualization should adequately support the data being used. Discuss why this visualization was chosen and how it is relevant. Questions to ask yourself when writing this section: - *Have you visualized a relevant characteristic or feature about the dataset or input data? - Is the visualization thoroughly analyzed and discussed? - If a plot is provided, are the axes, title, and datum clearly defined?*

Table 2: Overview of the numerical variables present in the dataset

Variable	Type	Description
is_canceled	integer	Value indicating if the booking was canceled (1) or not(0)
lead_time	integer	Number of days between booking and arrival date
arrival_date_year	integer	Year of arrival date
arrival_date_week_number	integer	Week number of year for arrival date
arrival_date_day_of_month	integer	Day of arrival date
stays_in_weekend_nights	integer	Number of weekend nights the guest stayed or booked
stays_in_week_nights	integer	Number of week nights the guest stayed or booked
adults	integer	Number of adults
children	integer	Number of children
babies	integer	Number of babies
is_repeated_guest	integer	Value indicating if the booking name was from a repeated guest (1) or not (0)
previous_cancellations	integer	Number of previous bookings that were cancelled
previous_bookings_not_canceled	integer	Number of previous bookings that were not cancelled
booking_changes	integer	Number of changes made to the booking
days_in_wating_list	integer	Number of days the booking was in the waiting list
adr	float	Average daily rate
required_car_parking_spaces	integer	Number of car parking spaces required by the customer
total_of_special_requests	integer	Number of special requests made by the customer

Algorithms and Techniques

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section: - *Are the algorithms you will use, including any default variables/parameters in the project clearly defined? - Are the techniques to be used thoroughly discussed and justified? - Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

Benchmark

In this section, you will need to provide a clearly defined benchmark result or threshold for comparing across performances obtained by your solution. The reasoning behind the benchmark (in the case where it is not an established result) should be discussed. Questions to ask yourself when writing this section: - *Has some result or value been provided that acts as a benchmark for measuring performance? - Is it clear how this result or value was obtained (whether by data or by hypothesis)?*

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section: - *If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented? - Based on the* ***Data Exploration*** *section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected? - If no preprocessing is needed, has it been made clear why?*

Implementation

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section: - *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data? - Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution? - Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

Refinement

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section: - *Has an initial solution been found and clearly reported? - Is the process of improvement clearly documented, such as what techniques were used? - Are intermediate and final solutions clearly reported as the process is improved?*

IV. Results

(approx. 2-3 pages)

Model Evaluation and Validation

In this section, the final model and any supporting qualities should be evaluated in detail. It should be clear how the final model was derived and why this model was chosen. In addition, some type of analysis should be used to validate the robustness of this model and its solution, such as manipulating the input data or environment to see how the model's solution is affected (this is called sensitivity analysis). Questions to ask yourself when writing this section: - *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate? - Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data? - Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results? - Can results found from the model be trusted?*

Justification

In this section, your model's final solution and its results should be compared to the benchmark you established earlier in the project using some type of statistical analysis. You should also justify whether these results and the solution are significant enough to have solved the problem posed in the project. Questions to ask yourself when writing this section: - *Are the final results found stronger than the benchmark result reported earlier? - Have you thoroughly analyzed and discussed the final solution? - Is the final solution significant enough to have solved the problem?*

V. Conclusion

(approx. 1-2 pages)

Free-Form Visualization

In this section, you will need to provide some form of visualization that emphasizes an important quality about the project. It is much more free-form, but should reasonably support a significant result or characteristic about the problem that you want to discuss. Questions to ask yourself when writing this section: - *Have you visualized a relevant or important quality about the problem, dataset, input data, or results? - Is the visualization thoroughly analyzed and discussed? - If a plot is provided, are the axes, title, and datum clearly defined?*

Reflection

In this section, you will summarize the entire end-to-end problem solution and discuss one or two particular aspects of the project you found interesting or difficult. You are expected to reflect on the project as a whole to show that you have a firm understanding of the entire process employed in your work. Questions to ask yourself when writing this section: - *Have you thoroughly summarized the entire process you used for this project? - Were there any interesting aspects of the project? - Were there any difficult aspects of the project? - Does the final model and solution fit your expectations for the problem, and should it be used in a general setting to solve these types of problems?*

Improvement

In this section, you will need to provide discussion as to how one aspect of the implementation you designed could be improved. As an example, consider ways your implementation can be made more general, and what would need to be modified. You do not need to make this improvement, but the potential solutions resulting from these changes are considered and compared/contrasted to your current solution. Questions to ask yourself when writing this section: - *Are there further improvements that could be made on the algorithms or techniques you used in this project? - Were there algorithms or techniques you researched that you did not know how to implement, but would consider using if you knew how? - If you used your final solution as the new benchmark, do you think an even better solution exists?*

Before submitting, ask yourself. . .

- Does the project report you've written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Analysis** and **Methodology**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your analysis, methods, and results?
- Have you properly proof-read your project report to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?
- Is the code that implements your solution easily readable and properly commented?
- Does the code execute without error and produce results similar to those reported?

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.

Mock, Thomas, and Antoine Bichat. 2020. "Tidytuesday: Hotel Booking Demand Data." 2020. https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-02-11/readme.md.

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