Predicting Cancellations and No-Shows Using R

Capstone Project Public Summary
Author: Karlo Nicolas Jaspe Figuerres
Program: MS in Business Analytics
Supervisor: Prof. Gabor Bekes
Academic Year: 2018 - 2019

Introduction and Background of the Study

The applications of data science are now observed in almost every field imaginable, from the conventional industries of Technology, Finance, and Medicine to Human Resources, Construction,and Government. And this is particularly true in the Hospitality Industry, specifically the hotel chain business.

The applications of data science in this line of business usually involves optimizing staffing, maintaining inventory management, enhancing marketing, and forecasting revenue. But due to the advent of more sophisticated statical methods, the applications of analytics in this line of business are becoming more and more innovative.

Archipelago International is one of the largest operators of hotels and hospitality properties in South East Asia. It currently manages over 130 hotels, with more than 18,000 rooms and apartments across Indonesia, Malaysia, and the Philippines.

As a forerunner in this industry, the client does not only want to maintain its position in the industry but also wants to make the lead larger. With this in mind, it has started exploring the potential of data science in this line of business.

The primary objective of the study is to create a prediction model that will identify which reservations will most likely be cancelled or be no-shows.

The secondary goals are to explore all available data (e.g. lead time, hotel location, hotel star level, time of year, method of booking) and identify which variables are good predictors for cancellations and no-shows, spot questionable information and raise to the client possible discrepancies in the data to properly address these moving forward, create visualizations showing patterns between variables that are relevant in the creation of the prediction model, and validate the performance of the model to see if it met the required specifications of the client (e.g. focus on on predictive power, interpretability).

Data and Methodology

There are over 130 hotels in the database with year's worth of reservation data. However, the client is currently undergoing a data migration and cleansing process and by the start of the project, only 5 hotels have data available. Data consolidation, tech setup, and data familiarization was already proceeding as planned but halfway through the duration of the capstone project, it was identified that there are concerns with the migration of the databases, particularly the quality of data.

With this, the approach changed from starting with 5 hotels to starting with 2 hotels. And by the time the data became available (12th of June, 2019), the timeline of the project has become increasingly challenging as the expected output must be provided in one week (19th of June, 2019).

Even with the strict deadline, the project proceeded as planned, starting off with data exploration. The client specifically prepared two datasets for this study - *reservations data* and *guest data*. The first dataset consisted of details about the booking such as arrival date, number of nights, number of adults or children, room rates and vouchers used. While the second dataset consisted of details about the customer such as address, birthdate, and purpose.

The data at hand comprised of 100,000 observations, each being a unique reservation, which spanned from 2014 to 2019.

As part of the data exploration process, we scanned each variable one at a time, inspected for questionable values, found ways on how to address these concerns, and ultimately looked for patterns that will help us create the predictive model.

At the beginning of the study, there were 68 variables which served as potential predictors of cancellations. Variables were excluded from the analysis for a variety of reasons such as being duplicates of other variables, having too many missing values, or even because of having no variability at all. New variables were also created based on other variables via grouping and categorization. External data sources were also used such as holiday information. By the end of the data exploration, 39 variables remained.

The initial approach was to create two models which prioritized separate things - Logistic Regression for the interpretability of coefficients and relationships, and Random Forest for accuracy and predictive power. But over the course of the analysis, other models were also attempted, specifically Logistic Regression with Regularization (LASSO), Gradient Boosted Machines, and, XGBoost.

However, because of computational limitations, the models were not maximized to their potentials. We encountered problems running multiple variables at a time, or processing factor variables with many levels all due to *memory issues*. To address this, we limited the number of variables taken at a time and excluded the factors variables with too many levels. As a result, we were able to get outputs from all methods except for the LASSO Regularization.

Results and Conclusion

After exploring the data, potential variables were identified and were then used for model development.

For logistic regression, our final model consisted of five predictors for cancellation. The relationships are as follows: reservations wherein customers were availed early booking discounts, used vouchers, or are scheduled on holidays are less likely to be cancelled or be no-shows. While reservations with fixed-rate deals, or with the involvement of travel agencies are more likely to be cancelled. However, the performance of our model is quite poor, as verified by its 40% AUC which means that it is performing worse than an unbiased 50-50 guess.

As for the tree-based models, all Random Forest, GBM and XGBoost had high accuracies which were around 99% and had AUCs near 99%. This is questionable since 99% is too high. We utilized 5-fold cross validations in these analyses but we cannot rule out the possibility of us being victims of overfitting.

As a conclusion, even though the logistic regression performed worse, we will still proceed with its results. With this, we have quantifiable relationships and definite interpretations without the risk of potential overfitting.

As a recommendation for future studies, the variables used in this analysis may be revisited to check whether there were variables that have contributed to possible overfitting. Running the models on more powerful machines (or even the cloud) should also be considered for greater efficiency. And more feature engineering should also be an idea for possible next steps.

As a final note, we may have found possible predictors but it is still difficult to confidently say that these variables are applicable to all of the remaining hotels. Nonetheless, this is still a great step moving forward since the data has now been explored, additional variables (from feature engineering) have been suggested, and the foundation of the analysis has been laid for future scaling and deployments.