Machine Learning solution for AirBnB in New Orleans

1. Summary

According to [IPX 1031](https://www.ipx1031.com/1031-exchange-services/?gclid=CjwKCAjw5vz2BRAtEiwAbcVILwk9Dhfhn3VsT3ElUOW4k0QXN9cZv9wAkfQsC4ZHfS2Su3tknkKefxoCrI4QAvD_BwE) recent survey Airbnb hosts and guests, 47% of hosts don't feel safe renting to guests while 70% of guests are fearful to stay at an Airbnb. The hosts expect a 44% decrease in revenue June-August. They have dropped their daily rates as much as $90 on average. 45% of hosts won't be able to sustain operating costs if the pandemic lasts another 6 months (16% ­have already missed or delayed a mortgage payment on one or more of their properties).

Considering the current situation, the objective of this analysis is to answer the questions below and provide empirical evidence on how COVID-19 has affected the Airbnb market in New Orleans, while including the number of COVID cases by zip code, city, state and country level statistics (which has a major effect on guests’ decisions on renting) as a part of determinants of demand & price such as listing details & amenities, quality/performance indicators of the hosts and also the number customer reviews for each listing.­­­­­­

***The questions to be answered****:*

1. *How does the COVID 19 pandemic affect the New Orleans Airbnb market?*
2. *How successful were AirBnB hosts on renting during the COVID 19 pandemic?­­­­­­*
3. *What factors have affected Airbnb hosts’ market exit decisions?*

To answer these questions, we aim to develop reliable prediction models using machine learning techniques like regression & classification to aid both the property owners and AirBnb with overall market evaluation given minimal available information about each property.

For the first question, we will build a regression model which predicts the revenue of each hosts for the next 30days.The revenue will be calculated by multiplying a hotel's average daily price by its occupancy rate for each month. For the second question, we intent to develop a classification model on the rental success for each host and also analyize factors that played major role to their success during the pandemic. The success classes will be derived from their revenue under the condition when it above/below the average revenue of all hosts in a zipcode. Other features of the rentals like owner characteristics, the customer reviews, covid cases will comprise the predictors for both analyses.

Finally, we will conduct customer churn analysis, again using classification technique to see what factors have affected hosts market exist decisions. Having a predictive churn model gives Airbnb awareness and quantifiable metrics to fight against it. The hosts exit( or churn) class will be identified by scanning over 6 months calendar dataset to check whether they still have active listings. The input features for this analysis will be mainly the target variables from the preceding analysis, hosts performance indicators with the covid stats added.

­­­

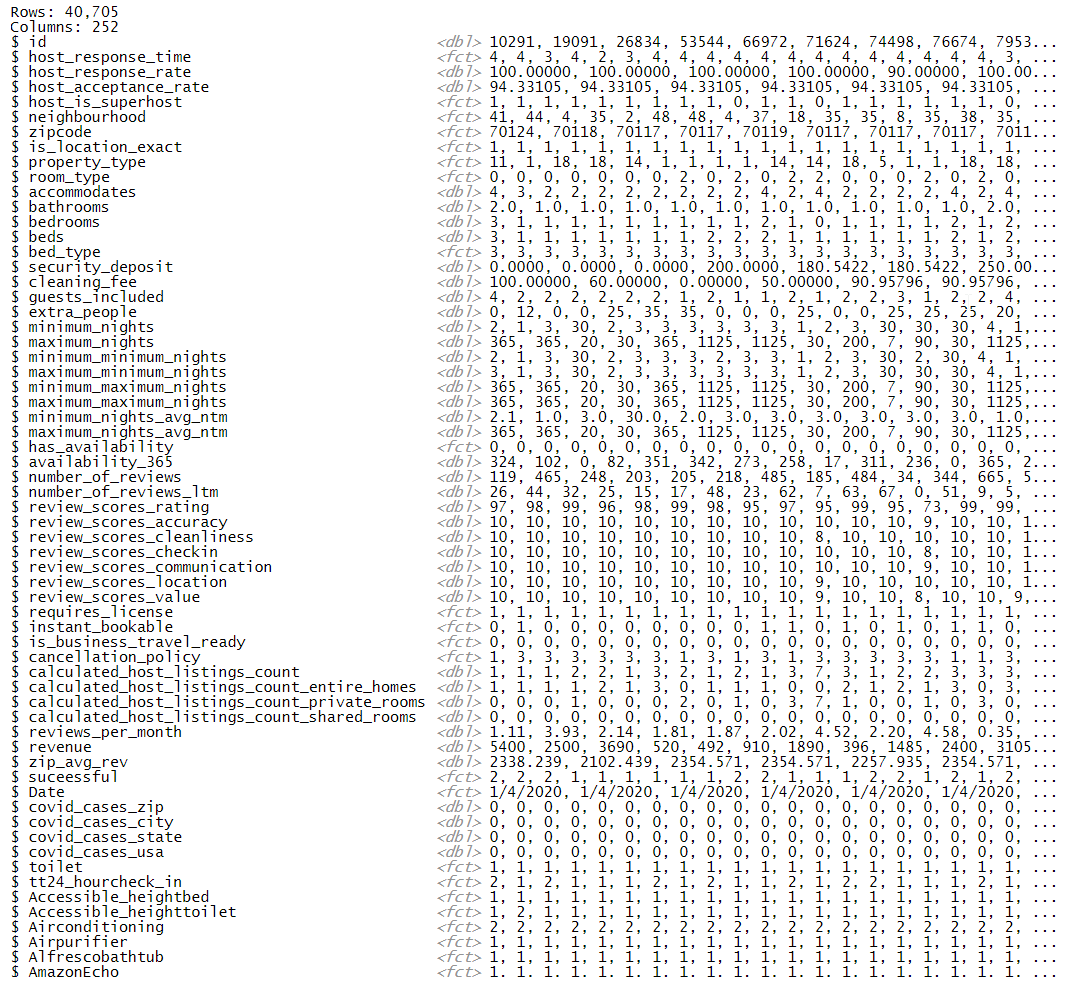
1. Data Preparation

The public Airbnb listings & corvid dataset for New Orleans was used as the main data source for this study. The final dataset included 40,705 entries, each with 252 features. Figure 1 shows the short structure of the listing in this dataset

For the initial prepossessing, we inspected each feature of the dataset to **(i)** remove features with frequent and irreparable missing fields or set the numerical features with missing values to average and categorical missing features were replaced with the most frequent values where appropriate, **(ii)** convert some features into integers (e.g. by removing the dollar sign in prices), **(iii)** change boolean features such as true/false to binaries, **(iv)** remove irrelevant or uninformative features, e.g. host picture url, scrape\_id, summary, listing\_url, fields or duplicate features, and **(v)** convert categorical features in the final set, e.g. ‘neighborhood name’ and ‘cancellation policy,’ into categorical labels. In addition, the features were normalized where the labels were converted into standard range -1 to 1 of the numerical features to mitigate the impact of the outliers in the dataset.

The data was split into two sets; namely, train set (comprising 70% of the original data), test set (comprising 30% of original data). Since the dataset was relatively large, 20% of the data was seemed sufficient for the accumulated testing and validation sets.

Figure 1- The below table shows the description of the cleaned dataset.

**

Covid Data­­­

The covid dataset is sourced from Louisiana Department of health website <http://ldh.la.gov/Coronavirus/> which is publicly available to download. The raw dataset came with Tract code, the USPS geographic location dataset was utilized in order to match zip codes for each tract code. The data was preprocessed and regenerated to match with the AirBnb geographic locations such as zip code, city, state, country and datetime. The excel file CovidData\_Gen.xlsx file includes covid data (March-July)for each state, city and zipcode. The dataset with geograpic indicators was exported and joined to the airbnb data tables based on date/addresses.

**Target Variable construction**

Revenue - To calculate the monthly revenue we simply multiply the occupancy rate by listing Price: the occupancy rate was obtained from the column in the listings dataset is called availability\_30 which stands for how many days the property is available in the next 30 days. So if a property is available 16 days, that means the room is occupied for 14 days (30-16).

Successful - Properties that make more than the average monthly revenue of a zipcode will be labeled as 1 (Successful), while properties that earn less will be labeled as 0 (Not Successful).

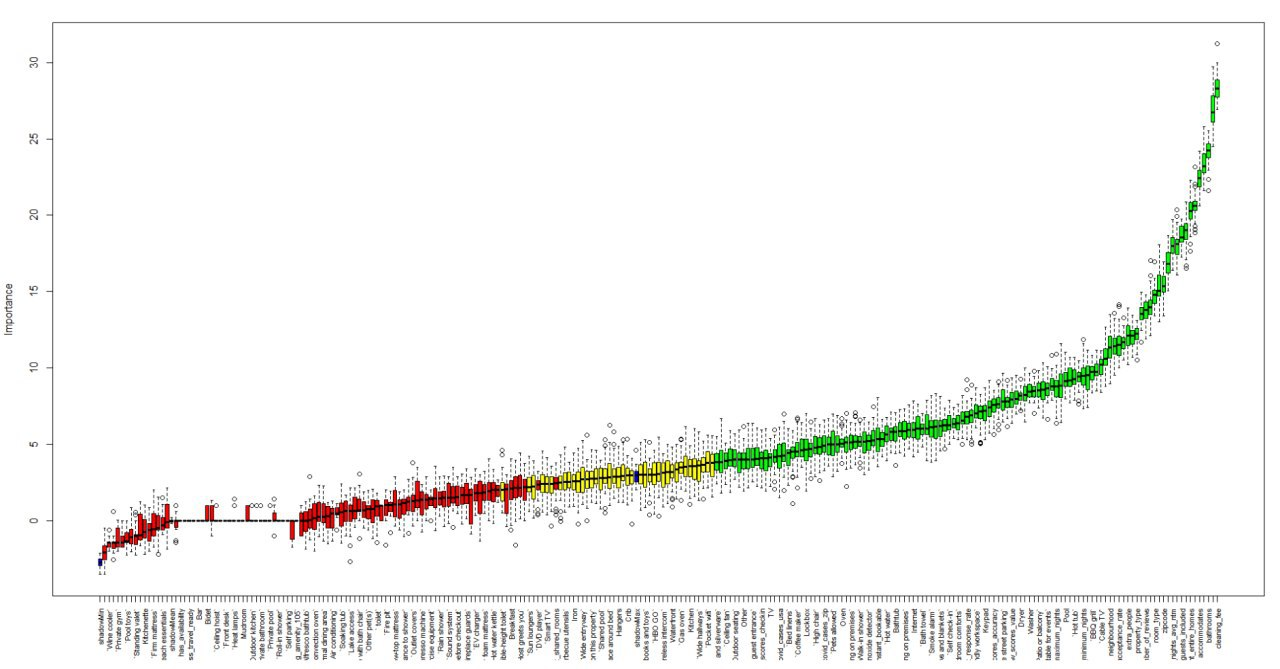
Churn(Exit) - The hosts churn class will be identified by scanning over 6 months calendar dataset by comparing hosts’ unique id’s of the n(1-5) month’s listing to the next (n+1) month’s listings to check whether they still have active listings.

**Feature Selection**

After data preprocessing, the feature vector contained ~252 elements. Feeding this excessive set of features to the models resulted in a high variance of error. Consequently, several feature selection techniques were used to find the features with the most predictive values to both reduce the model variances and reduce the computation time. Based on our knowledge/experience, the first tried method was manual selection of features to create a baseline for evaluating the other feature selection processes.

The second selection method was training on Boruta feature selection algorithm which uses Random Forest algorithm to provide educated sets of important and not so important features, respectively. Not only it save time, but it helped us with a repeatable and automatic way for initial exploratory data analysis. Based on this analysis, the model with the best performance over train with 20% split was selected. The resulting set consisted of 91 confirmed attributes for revenue prediction, 94 confirmed attributes for hosts success probability classification, 16 confirmed attributes for hosts exist(or churn) classification models. Those confirmed important were in green and rejected in red. Unresolved variables were in yellow and classified as tentative which Boruta was not able to conclude their importance.

**Figure- 2** shows feature importance plot of Boruta output

**

1. Results

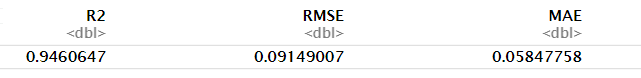
To choose the final model, we have used Azure AutoML to determine the best performing model for our cleaned Airbnb dataset. Our resulting best model was Random Forest which was trained/tested as a final model in our R notebook. Due to Studio’s performance issues, the models were only trained on the subset of the datasets.

* 1. Revenue Prediction – Regression

The Airbnb Hosts Revenue was predicted by RandomForest regression model for each month before and after covid. Model prediction accuracy was determined by the mean of the corrected coefficient of determination (R2), Mean absolute error (MAE), root mean squared error (RMSE) from regression of predicted revenue.

The following table shows the performance of the predicted model.

The model produced an R2 score of 94%, a RMSE of 0.091, MAE of 0.058 (defined on scale (revenue)) on the test set. This level of accuracy is a promising outcome given the heterogeneity of the dataset and the involved hidden external factors and interactive terms, including extra property characteristics/quality of the hosts, which were impossible to consider.



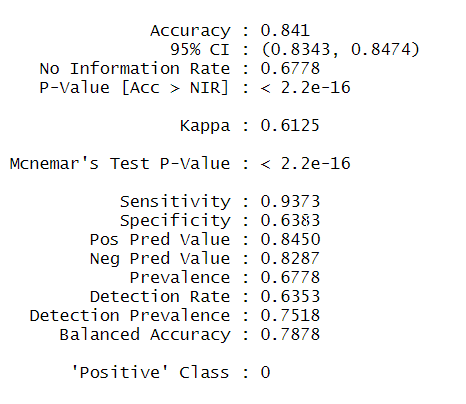
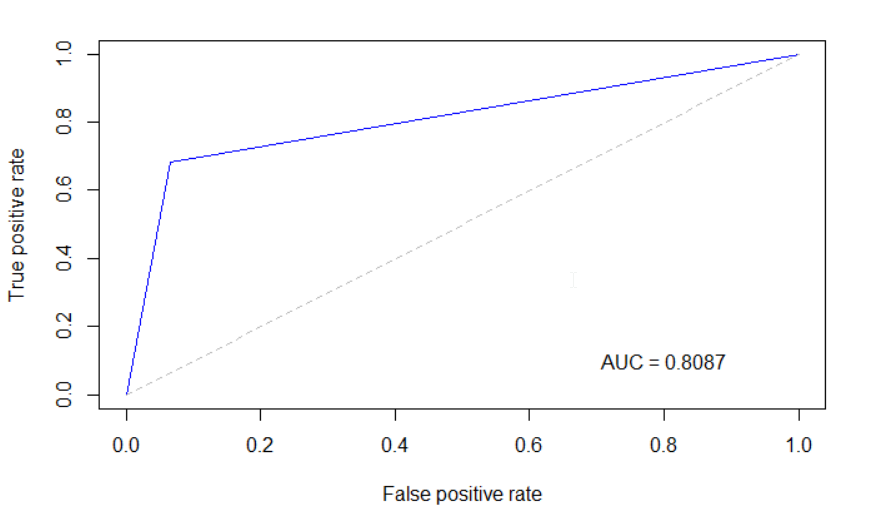
We determined ntree 20 as the optimal choice for the number of trees grown in RF, providing sufficient model complexity at a moderate calculation time. Further increase of ntree is expected to improve the model performance. Relative features importance was used to define the final parameter set in the RF model: The parameters with the smallest parameter importance were stepwise discarded and the highest importances were accepted as crucial factor for revenue prediction. As shown in the table below, COVID at USA level, seem to be one of the top highest importance with 5.4 scaled level. However, the hosts’ historical availability, zipcode level revenues, cleaning fee and number of reviews played major on determining the hosts revenue.

|  |  |  |
| --- | --- | --- |
| **Table 1** - Features that have impact on revenue of the AIRBNB hosts | | |
| **Rank** | **Features** | **Importance** |
| 1 | availability\_365 | 38.109332 |
| 2 | zip\_avg\_rev | 12.278989 |
| 3 | cleaning\_fee | 10.456995 |
| 4 | reviews\_per\_month | 8.936404 |
| 5 | accommodates | 8.648692 |
| 6 | minimum\_maximum\_nights | 7.859383 |
| 7 | maximum\_nights | 6.991816 |
| 8 | bathrooms | 6.88273 |
| 9 | room\_type2 | 6.246888 |
| 10 | security\_deposit | 6.222913 |
| 11 | number\_of\_reviews | 6.011325 |
| 12 | maximum\_minimum\_nights | 5.882756 |
| 13 | guests\_included | 5.822435 |
| **14** | **covid\_cases\_usa** | **5.408074** |
| 15 | bedrooms | 5.287816 |
| 16 | beds | 4.898985 |
| 17 | host\_response\_rate | 4.811532 |
| 18 | host\_acceptance\_rate | 4.732467 |
| 19 | number\_of\_reviews\_ltm | 4.650819 |
| 20 | Kitchen2 | 4.636069 |

* 1. Probability of Hosts Success – Classification

The Airbnb Hosts probability of success during the pandemic was predicted by Random Forest classification model. The model prediction accuracy was determined by the Confusion matrix, Classification Accuracy, and AUC. The following Figures show the performance of the predicted model. We trained the RF with only ntree 20, considering the calculation time. Thus, we expect that with the more number of trees the model will improve in performance. Nonetheless, our model did well performing with 84% accuracy and with 80% AUC values. In addition, the confusion matrix confirms that the number of True positives and True negatives are exceeding the false predictions.





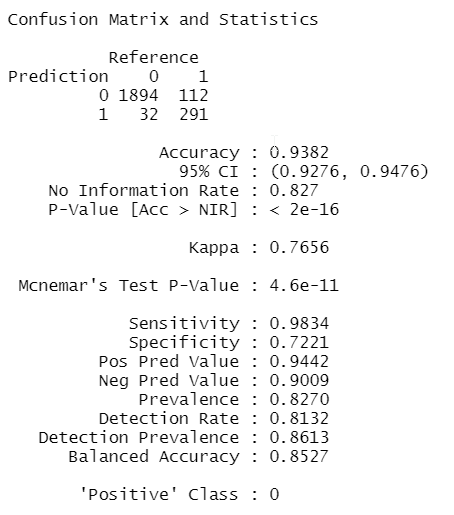
As shown in the importance table, the covid cases by zipcode and usa variables have some impact on the hosts success. Other important variables such as availability, bathrooms cleaningfees have the most impact on their success level too. In other words, the highly available properties with great bathrooms and less cleaning fees were successful during the pandemic.

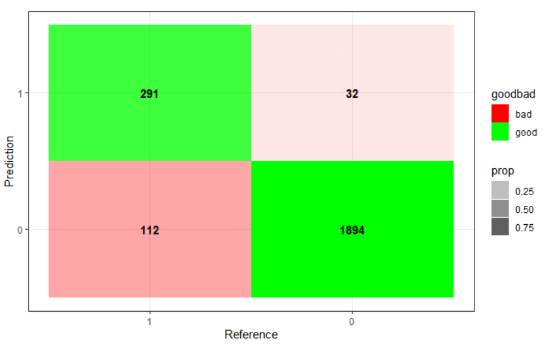
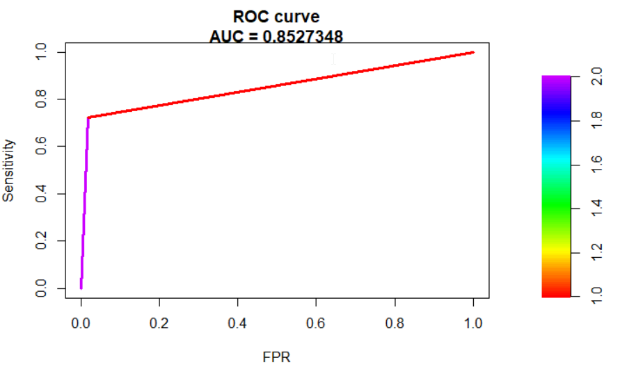
|  |  |  |
| --- | --- | --- |
| **Table 2** - Features that have impact on Success of the AIRBNB hosts | | |
| **Rank** | **Features** | **Importance** |
| 1 | availability\_365 | 36.146249 |
| 2 | bathrooms | 6.788041 |
| 3 | cleaning\_fee | 6.356757 |
| 4 | accommodates | 5.709668 |
| 5 | zip\_avg\_rev | 5.470186 |
| 6 | number\_of\_reviews\_ltm | 5.439871 |
| 7 | covid\_cases\_zip | 4.532877 |
| 8 | calculated\_host\_listings\_count\_entire\_homes | 4.171016 |
| 9 | bedrooms | 3.613736 |
| 10 | covid\_cases\_usa | 3.449166 |
| 11 | PacknPlay\_travelcrib2 | 3.290735 |
| 12 | host\_acceptance\_rate | 3.171576 |
| 13 | maximum\_nights | 3.151125 |
| 14 | number\_of\_reviews | 3.142117 |
| 15 | minimum\_nights\_avg\_ntm | 3.14194 |
| 16 | room\_type1 | 3.101045 |
| 17 | review\_scores\_rating | 3.0923 |
| 18 | covid\_cases\_city | 3.06926 |
| 19 | calculated\_host\_listings\_count\_private\_rooms | 3.037784 |
| 20 | beds | 3.020887 |
|  |  |  |

* 1. Customer Exit( or Churn) Analysis for AirBnb – Classification

Our predictive Churn model is a straightforward classification: as explained above we looked at the historical data and checked to see who is active after a certain time and here we create a model that probabilistically identifies the steps and stages when a host is leaving from Airbnb.

The model prediction accuracy will be the same as above model since it is a classification task. As shown in the figures, our model was able to predict with 93% Accuracy and 85% AUC. The specificity and sensitivity values seem to have a quite large difference, that’s because our data seem to have an imbalance between classes. For the future implementation, we had to feed more data and use up sampling/down sampling methods to balance the data before training it. Looking at the numbers in the confusion matrix we can clearly see that as well. Overall great performance of the model with just ntree=20 and with default parameters.





Feature imporatances suggest that, the hosts acceptance rate, number of reviews, revenue and covid cases seem to have a major impact on their exist decisions.

|  |  |
| --- | --- |
| Table 2 - Features that have impact on Exit desicions of the AIRBNB hosts | |
| **Features** | **Importance** |
| host\_acceptance\_rate | 24.5204984 |
| number\_of\_reviews | 11.8699628 |
| revenue | 10.0158663 |
| number\_of\_reviews\_ltm | 6.5743205 |
| **covid\_cases\_zip** | **6.4190045** |
| maximum\_minimum\_nights | 6.2294444 |
| reviews\_per\_month | 5.9456923 |
| availability\_365 | 5.9376075 |
| host\_response\_time | 5.9250626 |
| minimum\_nights\_avg\_ntm | 4.606507 |
| cleaning\_fee | 4.4866889 |
| calculated\_host\_listings\_count\_entire\_homes | 4.2579606 |
| calculated\_host\_listings\_count | 4.183783 |
| review\_scores\_rating | 4.1136568 |
| suceessful | 4.0084875 |
| host\_response\_rate | 3.5010675 |
| review\_scores\_checkin | 0.9543868 |

1. Final Remarks

This analysis attempts to come up with the best-performing model for predicting the Airbnb revenue/the probability of success/ and hosts exist decisions with the given the COVID-19 situation and a limited set of features including property specifications, owner information, and customer reviews on the listings. Machine learning techniques including regression & classification and along with feature importance analyses are employed to achieve the best results in terms of Root Mean Squared Error, Mean Absolute Error, R2 score for regression and Accuracy, AUC, confusion matrices for classification.

The future works on this analysis may include (i) studying other external features (ii) further experimentation with ensemble models (such as Stacking, Voting techniques) or neural net architectures, and (iii) getting more historical data samples.(iv) utilize powerful could based solutions to train the models on the full dataset. (v) build an application dashboard with the analysis of external risks for both Airbnb and hosts.