DABI Assignment 2

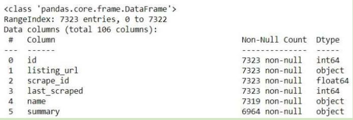
Jonathan Goh, 2081973F

Temasek Polytechnic, School of Informatics & IT – Specialist Diploma in Business Analytics

Introduction & Objectives

Airbnb Inc. is an American company based in San Francisco, California that was founded in 2008 by Brian Chesky, Nathan Blecharczyk and Joe Gebbia. Airbnb is a shortened version of the original name, AirBedandBreakfast. The company operates an online marketplace for vacation rentals which are typically homestays. The company generates revenue by collecting a commission from each booking that is made on its platforms. As such, it does not own any of the properties that are listed for booking.

This dataset contains data 7323 rows and 106 columns of various properties listed in Singapore over the past few years.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7323 entries, 0 to 7322
Data columns (total 106 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   id                                            7323 non-null   int64
 1   listing_url                                   7323 non-null   object
 2   scrape_id                                     7323 non-null   float64
 3   last_scraped                                  7323 non-null   int64
 4   name                                          7319 non-null   object
 5   summary                                       6964 non-null   object
```

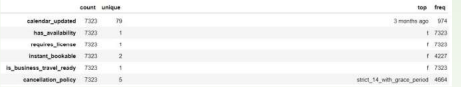
Using the data provided, the study aims to build a classification model to identify the listings that will be popular. The insights gained can hopefully then be used to determine what contribute to the popularity of the more popular listings and help the less popular listings improve.

Data Exploration & Preparation

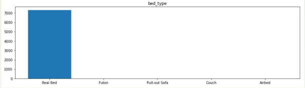
Most of the data exploration and preparation was done in a Jupyter Notebook.

Columns with large number of null values were dropped followed by columns containing ID, name, URL and date.

Categorical columns with too many unique values or only one unique value were dropped. Categorical columns with an overwhelming majority in a single class were also dropped.



	count	unique	top	freq
calendar_updated	7323	79	3 months ago	974
has_availability	7323	1	t	7323
requires_license	7323	1	f	7323
instant_bookable	7323	2	f	4227
is_business_travel_ready	7323	1	f	7323
cancellation_policy	7323	5	strict_14_with_grace_period	4664

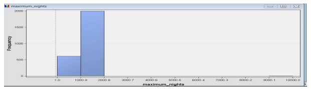


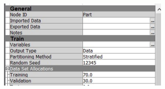
Checking the correlation between each of the numerical columns, those with high collinearity were dropped.

Dummy variables were created for the remaining categorical features.

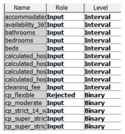
All rows containing null values were dropped.

The final dataset of 2605 rows and 48 columns were exported as a csv and imported into SAS. All the dummy variables were changed to "binary" type. One dummy was also set to "rejected" from each group. (e.g. host is superhost. A "1" on the "false" dummy variable is the same as a "0" on the "true" variable and thus unnecessary) A quick exploration of the imported data reveals an outlier in the maximum nights feature which was removed by a filter node.





Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Stratified
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0

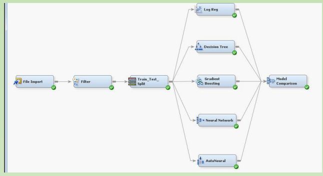


Name	Role	Level
accommodates	Input	Interval
availability_365	Input	Interval
bathrooms	Input	Interval
bedrooms	Input	Interval
beds	Input	Interval
calculated_hos	Input	Interval
calculated_hos	Input	Interval
calculated_hos	Input	Interval
calculated_hos	Input	Interval
cleaning_fee	Input	Interval
cp_flexible	Rejected	Binary
cp_moderate	Input	Binary
cp_strict_14_w	Input	Binary
cp_super_stric	Input	Binary
cp_super_stric	Input	Binary

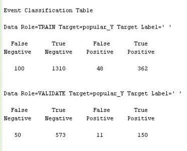
A partition was created to split the dataset between train and validation data in a 70-30 ratio stratified.

Workflow

The diagram used contains the import, filter and partition nodes, as well as various classification models and a comparison node. The configuration for the diagram is below and the models were all run based on default configurations with the results as follows:

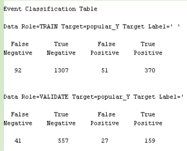


Log Reg



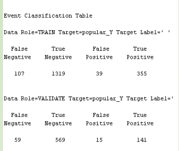
```
Event Classification Table

Data Role=TRAIN Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   100       1310        48         362


Data Role=VALIDATE Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   50         573        11         150
```

Decision Tree



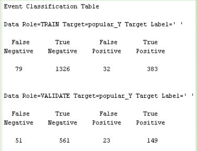
```
Event Classification Table

Data Role=TRAIN Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   92        1307        51         370


Data Role=VALIDATE Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   41         557        27         159
```

Gradient Boosting



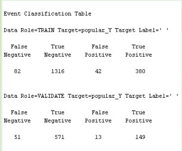
```
Event Classification Table

Data Role=TRAIN Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   107       1319        39         355


Data Role=VALIDATE Target=popular_Y Target Label='

  False       True      False       True
Negative   Negative   Positive   Positive

   59         569        15         141
```

Neural Network



```
Event Classification Table

Data Role=TRAIN Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   79        1326        32         383


Data Role=VALIDATE Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   51         561        23         149
```

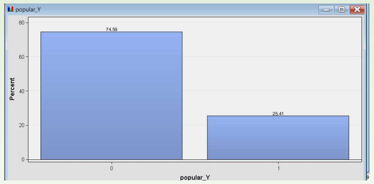
AutoNeural

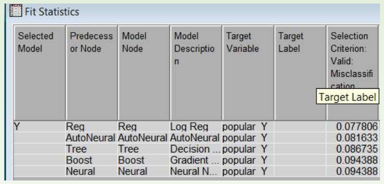


```
Event Classification Table

Data Role=TRAIN Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   82        1316        42         380


Data Role=VALIDATE Target=popular_Y Target Label=' '

  False       True      False       True
Negative   Negative   Positive   Positive

   51         571        13         149
```

Findings and Recommendations

As such, the model comparison module was set to select the best performing model based on misclassification. However, due to a class imbalance, misclassification may not be the best metric.





Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification
Y	Reg	Reg	Log Reg	popular_Y		0.077806
	AutoNeural	AutoNeural	AutoNeural	popular_Y		0.081633
	Tree	Tree	Decision ...	popular_Y		0.086735
	Boost	Boost	Gradient ...	popular_Y		0.094388
	Neural	Neural	Neural N...	popular_Y		0.094388

Instead, F1 score is calculated:

Logistic Regression	Decision Tree	Gradient Boosting	Neural Network	AutoNeural
0.831	0.824	0.792	0.801	0.823

The best performing model is still Logistic Regression. A possible way to improve the model may be to SMOTE after the partition before modelling but that will not be covered within this project.

Examining the feature weight importance of the logistic regression, we are able to identify that the following are the top 10 features that are important in determining popularity:

Latitude, Host Response Rate, Host Acceptance Rate, Number of Bathrooms, Moderate Strictness on Cancellations, Review Scores, Room Type, Bedrooms, whether the host is a SuperHost and Number of Guests included.

Based on the features identified, some cannot be changed such as the Latitude and Number of Bathrooms and Bedrooms. Nonetheless, the opportunities for improvement are available, such as, Responding to and Accepting more Booking Requests, and Trying not to be too Strict on Cancellations which should also increase their Review Scores.

DECLARATION: I declare that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged. I understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement. I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism.