*Aviation Company Dataset*

*Capstone Project  
Sep 2021*

*Manoj P*

# **1). Model building and interpretation**

This data set consists of 2 devices – Mobiles & Laptops with several data points on the users behaviours on digital and social platforms and their demographics. The plan is to build sperate models that will help us decide if the user is going to convert and buy tickets

The entire data set is separated based on Preferred Device - Laptop and Mobile

Mobile 10652

Laptop 1108

Target Variable for Laptop Users - Users who have not taken product from the Laptop dataset are 75% and users who have taken the product are 25%

Target Variable for Mobile Users - Users who have not taken product from the Laptop dataset are 85% and users who have taken the product are 15%

# Build various models

After separating the data set built various models for each dataset

Train/Test spilt was done using the predictor variable and target variable in this case the Taken Product. The split is 70:30 spilt of train and test data respectively.

The X variable will have all the variables except the target variable and Y will be the target variable

As we are looking at predicting how many users will take/buy the product our Interest class is 1 that is Product Taken as 1. Let's look at the performance of all the models on the both Train and Test Data.

Maximizing precision will minimize the number false positives, whereas maximizing the recall will minimize the number of false negatives.

Precision: Appropriate when minimizing false positives is the focus.

Recall: Appropriate when minimizing false negatives is the focus.

We will look at these metrics that we want excellent predictions of the positive class. We want high precision and high recall. We will also see the F1-score it helps to see model's accuracy on a dataset. The F-score is a way of combining the precision and recall of the model

First let’s look at the models built for Mobile Users on train dataset- with Interest Class 1

Table - Mobile Users Train Model

|  |  |  |  |
| --- | --- | --- | --- |
| Mobile Users | Logistic Regression | KNN | Naive Bayes |
| Precision | 69 | 94 | 57 |
| Recall | 19 | 87 | 29 |
| F1 Score | 30 | 90 | 39 |
| AUC | 78 | 99 | 76 |

Now, let’s look at the models built for Laptop Users – with Interest Class 1

Table - Laptop Users Train Model

|  |  |  |  |
| --- | --- | --- | --- |
| Laptop Users | Decision Tree Classifier | KNN | Naive Bayes |
| Precision | 65 | 94 | 66 |
| Recall | 41 | 89 | 55 |
| F1 Score | 51 | 91 | 60 |
| AUC | 79 | 99 | 81 |

From the above tables you can note that in both datasets KNN model performed well on train dataset with high Precision, Recall, F1 Score and AUC.

# Test predictive model against the test set using various appropriate performance metrics

Let’s see some results on the test dataset for Mobile users – with Interest class as 1

Table - Test Model Mobile Users

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Regression | KNN | Naive Bayes |
| Precision | 73 | 88 | 63 |
| Recall | 21 | 70 | 32 |
| F1 Score | 33 | 78 | 43 |
| AUC | 78 | 99 | 76 |

Now, let’s look at the models built for Laptop Users – with Interest Class 1

Table - Test Model Laptop Users

|  |  |  |  |
| --- | --- | --- | --- |
|  | Decision Tree Classifier | KNN | Naive Bayes |
| Precision | 66 | 88 | 66 |
| Recall | 55 | 63 | 55 |
| F1 Score | 60 | 74 | 60 |
| AUC | 82 | 99 | 81 |

The results for test dataset were similar to train with KNN model performing well however there was a reduction in the Precision, Recall and F1 Scores we will perform grid search to find out optimum values for hyper parameter and cross validation method to evaluate the model and check its performance on the unseen data. Also we will build Ensemble models.

# Interpretation of the model

As mentioned above we are building a model to predict if a user will convert and buy a ticket or not, for practical purposes, we will be more interested in correctly classifying 1 (taken the product) than 0(did not buy the ticket).

It makes more sense to look at model performance metric that when we incorrectly predict a user, who actually has bought the ticket as not having bought the ticket which we can get from Precision also it will be good to look at how many are actually negative which we can get from Recall and will also look at F-score which provides a single score that balances both the concerns of precision and recall. The AUC ROC curve will also give us about the overall classifier performance.

There were multiple models built and based on what we see is the k-nearest neighbours performed well than other and with good results as seen below.

Table - KNN Model for Mobile and Laptop Users

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| KNN | Mobile Train | Laptop Train | Mobile Test | Laptop Test |
| Precision | 94 | 94 | 88 | 88 |
| Recall | 87 | 89 | 70 | 63 |
| F1-Score | 90 | 91 | 78 | 74 |
| AUC | 99 | 99 | 99 | 99 |

The model did well in the train data set than the test dataset this might be something with overfitting, we will build other models and look at the performance metrics.

# **2). Model Tuning and business implication**

# Ensemble modelling

Now, let’s look at the scores for the Ensemble modelling method - performed for both Mobile and Laptop on train and test dataset.

Ensemble modelling on Mobile dataset

Table - Ensemble Model Results on Mobile Users

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Mobile | Bagging Train | Ada Boost Train | Gradient Boosting Train | Bagging Test | Ada Boost Test | Gradient Boosting Test |
| Precision | 100 | 72 | 91 | 99 | 70 | 88 |
| Recall | 100 | 34 | 45 | 91 | 35 | 40 |
| F1-score | 100 | 46 | 61 | 95 | 47 | 55 |
| AUC | 100 | 86 | 94 | 100 | 87 | 94 |

From the above you can note that Bagging really performed well with highest scores for both on train and test dataset of Mobile users

Ensemble modelling on Laptop dataset

Table - Ensemble Model Results on Laptop Users

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Laptop | Bagging Train | Ada Boost Train | Gradient Boosting Train | Bagging Test | Ada Boost Test | Gradient Boosting Test |
| Precision | 100 | 92 | 100 | 100 | 80 | 98 |
| Recall | 100 | 81 | 93 | 98 | 71 | 87 |
| F1-score | 100 | 86 | 96 | 99 | 75 | 92 |
| AUC | 100 | 99 | 100 | 100 | 99 | 100 |

Form the above metric we can note that all the models performed well with laptop user dataset but among these modelling techniques Bagging had best numbers.

# Model tuning measures

Grid-searching was also performed on the dataset for Laptop and Mobile dataset to find and configure the optimum values for hyper parameters however did not see great performance on the model. The grid-search was done for Mobile user dataset on Logistic Regression and for Laptop users dataset on Decision Tree Classifier. Please see below result from the exercise

Table - Grid Search

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LR Mobile Train | LR Mobile Test | DTC Laptop Train | DTC Laptop Test |
| Precision | 69 | 73 | 65 | 66 |
| Recall | 19 | 21 | 41 | 55 |
| F1-Score | 30 | 33 | 51 | 60 |

Trying to minority oversampling technique with SMOTE and resampling with Cross-Validation on both Mobile and Laptop Dataset let look at the results with KNN Model.

Table - SMOTE on Mobile and Laptop Users

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SMOTE | Mobile Train | Laptop Train | Mobile Test | Laptop Test |
| Precision | 98 | 99 | 90 | 100 |
| Recall | 100 | 100 | 100 | 100 |
| F1-Score | 99 | 100 | 95 | 100 |
| AOC | 100 | 100 | 100 | 100 |

After performing SMOTE on dataset we see that it has performed very well on both train and test dataset. Hence this is the best model to accurately classify the classes and predict if a user will convert and buy a ticket or if the user will not buy the ticket.

Let’s see the stability of the model with Cross-validation and how it performs on unseen data

Mobile User

|  |  |
| --- | --- |
| Train | 0.96, 0.97, 0.97, 0.98, 0.97, 0.99, 0.99, 0.99, 0.99, 0.99 |
| Test | 0.87, 0.88, 0.89, 0.89, 0.87, 0.90, 0.87, 0.89, 0.89, 0.87 |

Laptop Users

|  |  |
| --- | --- |
| Train | 0.97, 1.0, 0.99, 0.978, 0.99, 0.99, 1.0, 0.99, 0.99, 0.99 |
| Test | 0.82, 0.76, 0.82, 0.82, 0.82, 0.76, 0.73, 0.76, 0.73, 0.85 |

After 10 fold cross validation, scores both on train and test data set respectively for all 10 folds are almost same. Hence the model is valid.

# Interpretation of the most optimum model and its implication on the business

SMOTE utilizing KNN was the most optimum model with the below results which means that this model performed very well with best with minimizing the False Positives and False Negatives which means that the model has accurately predicted the Classes where users have bought the ticket or not bought the ticket.

Table - SMOTE Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SMOTE | Mobile Train | Laptop Train | Mobile Test | Laptop Test |
| Precision | 98 | 99 | 90 | 100 |
| Recall | 100 | 100 | 100 | 100 |
| F1-Score | 99 | 100 | 95 | 100 |
| AOC | 100 | 100 | 100 | 100 |

Some of the implication on the business are as follows. It can take multiple measures after identifying the set of target customers

* Reduce operational cost of tele-calling since digital/social media platforms are relatively cheaper and faster way to target and personalize for the users
* Right pricing strategy in terms of pricing offer and discounts.

Users that are likely to buy won’t need high discounts however users who are not buying will need more aggressive offers there by bringing in incremental revenue.

* Cross sell and upsell opportunities among the top customers
* More customized products for certain customers.