Title

MSc Research Project
Data Analytics

Ugwuanyi Arinze J
Student ID:x18139442

School of Computing
National College of Ireland

Supervisor: Prof Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Ugwuanyi Arinze J
Student ID:	x18139442
Programme:	msc Research project
Year:	2020
Module:	MSc Research Project
Supervisor:	Prof Vladimir Milosavljevic
Submission Due Date:	13/8/2020
Project Title:	Title
Word Count:	XXX
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	25th September 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A predictive Analysis of airline passengers demand between (Eco and Business class) during and post economic recession using machine leaning algorithm

Ugwuanyi Arinze J
X18139442

Abstract

Naturally, the airline industry is competitive, a strategic process is needed to make more profit, understanding customers demand will increase the service performance, Economy or Business class airline are the most often use airline, the requirements and price differ, the need to find out the root cause of demand imbalance has been a challenge, this research aims to predict with accuracy business class users and the major factor the effect passenger preference, it is important in this is economic downturn to know the factors that influence the demands of a passenger. this research applied a machine learning approach to give detail analysis of the factors that contribute to the customers' demand.

The application of data science on IBM survey data, help to predictive demand on Business class flight from the attributes, removing some attributes that do not relate or contribute was done using Principal Component Analysis (PCA) it helps improves the accuracy performance of the model. An exploratory data analysis (EDA) performed to identify the significant factors, machine learning algorithm (K-Nearest Neighbour (KNN), Random Forest, Logistic Regression and Decision Tree (C50) was employed for the prediction. The machine learning model performance was evaluated using Recall, Accuracy, Precision, and FI-score parameters, the best result was gotten from KNN when K=1 with an accuracy of 91.3%.

1 Introduction

1.1 Background

The airline is one of the means of transportation, which a great number of people in the world wish and admire to use, it is highly modernized, very comfortable, reliability and very fast way of travelling, there are other means of transportation like, by Car, by Water and Railways. but, the airline is the fastest growing sector of transportation (Sacha; 2010). At this period of world economic crisis, a big sector like aviation must understand the customer demand, the factors that contribute to the choice of a particular airline. The impact of economic recession on business and post effects have to lead to price reduction, competition, downsize of employed staff, it is important to acknowledge the current condition and take necessary action to overcome this business challenges (Igor et al.; 2015)

(Ana et al.; 2010) Annual report on US airline shows that airline successfully flaws 2.2 billion people and a total of 35 million tonnes of cargo bring it to 35% of annual value successfully trade internationally and a whopping of 8% of annual global GDP, a total of net profit of $12.9b with a different of 2.5% margin. From (CBP) The industries net loss was estimated at $16.8b, a drop in passenger carrier and particularly the cargo traffic reduced to negative (-13%) all this was as a result of 15% downturn in the economy. Aviation industries is a big sector of the economy with estimated 32million jobs. Through the direct and indirect job, the 5.5 million jobs come from employees, 4.7 million jobs are from airline and aviation industries, 6.3m employees came from the indirect staff purchase of goods and services,2.9m are from the spending from an induced job through the spending of aviation 800000 jobs from the aircraft manufacturer companies, body frame and engine job through the tourism was also effected as results of recession.

However, oligopoly exists among the united state airline operator where small company dominate the market share of the industry (Rian et al.; 2019) that the enactment of 1978 bills that allow an airline to charge any amount that they deemed fit and flaying to any destination have introduced high competition among the company.improvement in customer services is key to get more customer which accurate prediction of customers traffic demand will help to reduce the Number cancellation and delay in departure which will increase customer satisfaction, inaccurate prediction of factors that influence passengers demand will cause weak or poor service provided to customers. at this time of economic recession which airline industries are one of the most affected section because of restrictions on the local and international flight have imposed limited resources in the industry it important to judicious make use of remaining resources.

1.2 Project Motivation

Recently, the current situation of the airline industries is in a bad sharp as a result of COVID-19 Pandemic, restriction of travelling both locally and internationally flights has affiliated serious financial crisis in the industry, that they depend on government financial support to start operation again, (Joseph; 2020) the global pandemic result from the COVID-19 has created many chaoses as quarantines, hospitalization, social distance policy on land and air travel continue to put the economy in negative shape. This current challenge leads to this research to conduct prediction analysis on passenger demand and major factors that affect passenger preference during and post-economic recession on business class. this model can be also applied to other countries used not only in the United State as the dataset specified.

This analysis was derived from the current and past trends in the airline, appropriate demand forecast will enhance good services, it will increase safety and security measures. From previous research, it shows that tourism and business travel have increased the development in airport (Opera; 2010) it is important to adopt modernisation of technology and data science, (Markus and Florian; 2011) when the airline operation managers are still leaning lesson on the impact of the 2001/2003 economic crisis, boom the double deep economic recession of the 2007/2009 downturn start, it was a result of lack of proper guideline or information of the past trend that leads to the long time recovery. but the aviation industries were quick to recover from the 2007-2009 recession than 2001.(Igor et al.; 2015) as a result of a good strategy in place.

1.3 Research Question

How effective can, application of machine learning algorithm predict with accuracy the Business class flight users and the major factors that contribute to their preferred demand during and post-economic recession?

1.4 Research objectives

(a). This research aimed to apply a machine-learning algorithm to predict with accuracy the business class users and factor that contribute to their demand

(b). This research aimed to analyse the contribution of the variables in the dataset that are significant in the prediction business class users, and use the variables to build a model.

(c). To study the performance of machine learning models predictions and classification and adopt the best-performed model.

(d). To carefully review previous related work on passenger preference and attributes that affects demands on a business class airline and improve the research by adoption new innovates and significantly contribute to the research problem.

(e). To propose the best evaluation criteria for this similar study, within the acceptable arrange of error for the models and compare their results.

This research work focused on the available attribute to predicts airline business class users and major factors that effects passengers preference, this similar work have been researched by many authors like, (Rian et al.; 2019) and (Jian et al.; 2018) tried to find the factor that influences the passenger preference between the low-cost carrier and legacy carrier using the logistic regression model.

1.5 Report Structure

This report is structure accordingly, starting with section 1, Introduction: introduced the research problem and motivations to the project, section 2is the Related works where I discussed the previous research, their limitations and future work and how they were able to contribute. In section 3: Methodology: this is where I detailed the method and process used to carry out this research. section 4: Design Specification: this is the area technique and architecture of the project was discussed. Section 5: Implementation of the models and the process followed. section 6: The Evaluation of the performance of the models implemented in section 4. Section 7: is the conclusion and future work, this is where I discussed my result finding and compare the results with the one on literature review, lastly Reference.

2 Related Work

2.1 Introduction

(Praphula et al.; 2019), continued the research on a study the flight recommendation service using another means to generate data investigate customer recommendation and the likelihood of repurchase from the feedback generated data, the authors stated that online reviews have offered more quality benefits than the simple market strategy. this research was conducted with the data generated from the airlinequality.com using machine learning algorithm, KNN, SVM and Decision Tree. The performance of the model was evaluated using Accuracy, Precision, Recall and F-Measure result. The author finds out that SVM performs better than KNN and DT algorithm with 82.75% accuracy, 82.08% precision, 86.83% Recall and 85.94% F-Measure. interesting result.feature research is using different data set with another text processing will give a good result.

(Srinidhi; 2009) develop airline traffic forecasting model for an international flight in India, he uses physics equation model and micro-economic theory to test the factors affection airline demand especial factors outside the services, these factors were divided into two sections Geo-economic (GE) and Services Dependent (SD). the passenger traffic model (PTM) was tested and it provides demand estimated for new international route contemplate using the data generated from the existing flight.

(Safak et al.; 2003) investigate the difference in foreign and domestics airline from passenger viewpoint based on four sections, demographic profile, behaviour, understanding of airline services, and satisfaction level. 1014 observers were collected from passengers in Istanbul, it shows that passenger in the foreign airline is more likely to be Male, Older and have higher Education cert unlike the passenger in the domestic's airlines. T-Test and Chi-Square statistical tools were employed to find the relationship (correlation) between the demographic and behavioural characteristics of passenger, but it was significant and there is a great significant difference between the passenger fly foreign airline and domestic airline

2.2 The impacts of economic recession

(Sacha; 2010) study the impacts of the economic downturn on Irish airline this research involve the three major Irish airlines, Ryanair, Aer Arann and Aer Lingus. The aim of the project is to compare the strategic change in the airline during and post-recession period. Thus, this qualitative research focused on strategies implemented to survive the economic recession and to compare the Japanese airline strategies used to survive the 1990 recession, the author finds out that performance of any organisation ultimately depends on the type of strategy and quality of marketing management. The impacts of economic recession in aviation are numerous, (Sacha; 2010)Igor Stimac et al (2005) investigate the effect and changes in low-cost carrier airline during the period of economic recession, the author said that the low-cost carrier (LCC) introduced hybrid carrier(low-cost long haul carrier) as a new strategy and loyalty programs to support their customer and enable their new business marketing strategy to perform well.

(Chandrakumara et al.; 2018), unlike (Sacha; 2010), (Igor et al.; 2015) that study the impact of Pre and post-economic recession on (IT) stock listing in the US stock market with data set collected from NASDAQ. The performance analysis for pre-recession (2005-2007) and post-recession 2009-2011) was analysed using python the result finding show that the price was least in May 2007 and highest December 2007 before the peak of the recession, then drastically down the during the peak of recession 2008.

(Peter; 2011) that financial and operation estimate for airline industries are optimistic in 2009, The research aim is to identify the challenges of airline industry during this economic downturn and to compare the current downturn with the previous ones. The IATAS growth forecast was based on the GDP projector available and find out the 9/11 caused global airline profitability to fell by $22.50 billion. He also finds out that premium cabin traffic has fallen rapidly and this led to the reduction in the number of seats for this class and increase the economy class seat to accommodate demand. That the major challenge of the airline during the economic downturn was the maintaining load factors and yields at reduced cost price.

2.3 Procedure for balancing Imbalance data

(Jose et al.; 2015) study the best method to balance an imbalanced data set in machine learning algorithm, the author proposed that Synthetic Minority Oversampling Techniques (SMOTE) is the best method to uniformly balance the data instead of using traditional Rotation Forest after using (PCA) Principal Component Analysis to select the feature that is significant the prediction. Smote is a process of upsurging the smaller ration of the minority class to the exact or equivalent ratio with the majority class but the original majority class are is untouched. When the compare the model performance of the imbalanced data set with over-sampled data the Accuracy of the over-sample data is always higher both in Recall, Kappa and precision but have a higher tendency of overfitting.

(Fadi et al.; 2019), used a combination over-sampling and downsampling method called Hybrid method. The author used the hybrid method to solve the problem of imbalanced data, this is the process of integrating the majority class in under-sampling and minority class for oversample with a grid search to optimizes the parameter.

(Sushruta et al.; 2020) researched the best way for optimization of skewed data using a sampling-based pre-processing approach, he applied three different sampling method to solve the challenge of skewed data using, Resampling, SMOTE and spread subsampling. The poor performance is a result of the model uneven distribution of dataset is best using any of the three methods depending on the nature of the data. In these three approaches, SMOTE is used as the oversampling method, while spread subsampling was projected as Under-sampling

2.4 Research gap in related work.

Regard to the review I conducted on related works I observed that (Rian et al.; 2019) research work is the most related and recent research like this one but the author/authors used SPSS linear and multiply regression model for the prediction. in this research a higher machine learning approach like K-Nearest Neighbour, Logistic Regression, Decision Tree (C50) and Random Forest was applied to predict business class users and factors that influence the passenger preference, this time of economic recession its important to analyse the effect of the economic downturn on passenger demand. the model will be build using the variables that are significant and the result will be evaluated base on Accuracy, Precious, Recall F1-Measure.

3 Methodology

This research followed a quantitative research method, there are three most common use methods; KDD (Knowledge Discovery Database), CRISP-DM (Cross-Industry Standard

Process for Data Mining), and SEMMA (Sample, Explore Modify, Assess), but in this research, CRISP-DM method was applied because methodology fit perfectly for this type of research that involves business understanding compared to KDD and SEMMA. (Santos and Azevedo; 2008). CRISP-DM is the most structured to solve data mining problem by following 6 cyclical processes explained below.

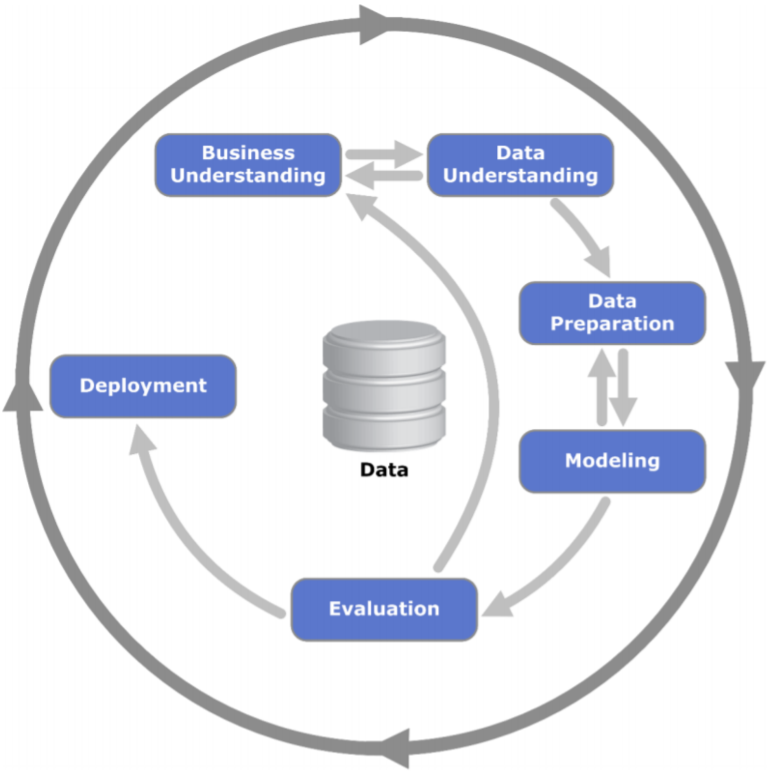


Figure 1: CRISP_DM process

3.1 Business understanding

In this type of projects that required data analysis it is necessary to start with the understanding of background knowledge of the domain, a clear understanding of the business objectives at that particular time is needed, it will enable a perfect conversion of generated data into information using data mining techniques. When using a CRISP-DM method the first step is to read and identify the variables and their function then use it for exploratory analysis to identify the ones that contribute to predicting of the target variable. At this point, it is important to know the passenger's requirement that will enable the management to make it readily available to avoid missing targeted customers. During this time of economic recession, it is important to understand customers requirement to avoid a waste of resource at this critical time in the industries by avoiding misplaced priority.

Economic recession impose much financial crisis to the airline industries, it makes some section of operation service to close or heavily affected, it is necessary to have a clear understanding of customers desire and identify the weak point in the operation of the industries, analysing customers feedback will enable understanding changes in customers to desire during the economic recession.

3.2 Data understanding.

This section explains the clearly the details of the dataset, the source the dataset was downloaded, the size of the data, the volume of the row and column are discussed in this section. Data understanding help in building a perfect model. Identify the variable and its function in the project is vital, the dataset used in this project was downloaded public data repository IBM community Watson, this dataset was generated from airline passenger in the united state from different airline and state, it consists of 128,863 observation and 28 columns satisfaction, class, Airline status, age, Gender, price sensitivity, year, flight distance, destination, Airline State, shopping, eating and drinks, destination city, flight time, departure time, Airline name, day of the month, and loyalty cards.

3.3 Data preparation

This stage involves the cleaning of the data, Conversion, check for Null and missing values, performing the exploratory analysis on the data to select the variables that are significant to the prediction.

Attribute	Format	Description
Satisfaction	Numerical	This is the quality of service (star level) rating of the of airline operation by the passengers' experience
class	categorical	This is the type of flight book is business or Economy class
Price_Sensitivity	Numerical	This the degree at which the flight ticket affects the passengers.
Eating_and_Drinking	Numerical	This infer the quality of the food and drinking share to the passenger.
Loyalty_cards	Numerical	This infer the rewards or discount for regular use of services it records the shopping habits
Airline_status	categorical	The colour of the flight
Date_of_the_month	Numerical	It infers the day of the month it happens
shopping	Numerical	This is the degree of cost price of the goods and services
Flight destination	Categorical	This is the state destination of the flight
Age	Numerical	The age of the participate
Year	Numerical	The year the flight to place
Gender	Categorical	The sex of the participate
Airline_Name	Categorical	The name of the airline (company name)
Flight_Cancellation	Categorical	This is a record if the flight was cancelled of not
Departure_Time	Numerical	The scheduled time for the departure
Type_of_Travelling	Categorical	This is a type of travel like personal, business or mileage travel

Table 1: Summary of the Attributes in the Data

3.4 Data cleaning

A clean dataset will help in increasing the accuracy of the prediction and also make model prediction easy, because not all model perform with null value, it will increase the quality of data by replacing or removing the missing values. Removing outlier in the dataset will enhance the correlation coefficient.

3.5 Data modelling.

This section of the project is very important it engages all other section, it involves the understanding of various aspects of the project, the knowledge of the literature reviews of the previous researchers and finding the best way to produce an efficient result. After the pre-processing of the dataset and building of model is done than the implementation will take place. In this project principal component analysis (PCA) dimension reduction was employed to remove the non-significant variables in the data, the significant variables were split into two parts the training and Testing. Random Forest, KNN, Logistic Regression and Decision Tree (C50), the model was built for better accuracy prediction and comparison of the result.

3.6 Evaluation

This is the process of estimating the model performance, in this project, Accuracy, Recall, Precision, and F1-Measure was used to evaluate the model performance and comparison of the model predictions. True positive and True negative correct prediction was analysed. The results were analysed to identify the model that best predicts business class users and the effects of the economic recession on passenger preference.

4 Design Specification

The figure 2 describe the process undertaken in this research project, it is a visual representation of steps followed for the execution of this project. The reviews from the previous author on the related research problem, machine learning models were built, CARET (Classification, And Regression Training) package in RStudio was installed, it helps in the classification of models. Caret is well known for perfect data splitting (Testing and Training).

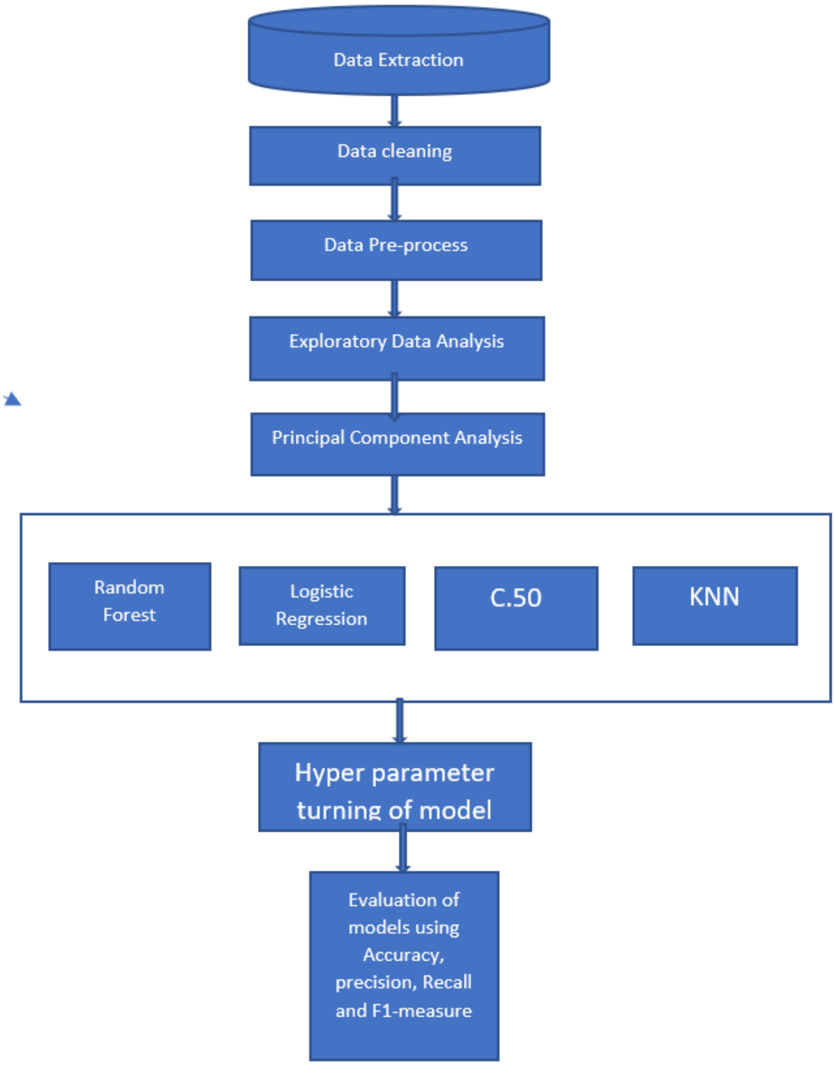

Data Extraction
Data cleaning
Data Pre-process
Exploratory Data Analysis
Principal Component Analysis
Random Forest
Logistic Regression
C.50
KNN
Hyper parameter turning of model
Evaluation of models using Accuracy, precision, Recall and F1-measure

Figure 2: Design flow process

This research focused on the prediction of business-class user and factors affecting passengers preference in the recession period. six steps process are:

1 Data Extraction: this is getting reliable data and downloading it.

2. Pre-processing: the extracted data were cleaned to remove all the Null value and transformation of the categorical variable to numerical value. Exploratory Data Analysis (EDA) was done to get an overview of the data variable significant.

3. Building of model: this is the process of normalising target variable and splitting the data into 70-30 for training and testing respectively. training data is used to train a model while testing is used to test the model.

5. Evaluation: This is a process of estimating the performance of each model, using different measurement depending on the methodology.

5 Implementation

At this stage in the project it describes the transformation done on the dataset and output of the process, the process involves the building of models for evaluation.

5.1 Data processing

The data set used in this project was downloaded from IBM Watson Analytical it was gathered by the passenger in united state in 52 states combined during the IBM survey in excel.xls it contained a period when there is a recession (2007-2009 and post-recession period, it was uploaded into RStudio using library(readxl) it consists of 129,867 row and 28 columns. Subsequently "str(dataset)" used to inspect the data "sum(is.na(dataset))" was used to check for a total number of na-values, "na.omit(Dataset)" was used to remove na-values, "any(is.na(Data))" used to check if any na exist "boxplot(dataset) "was used to check for a duplicated row, missing values and outliers. It is imbalanced data set 90% of Economy class users and 10% business class users. missing values and error values were remover for better accuracy. The selection of the columns was done base on their importance to the prediction.

5.2 Exploratory Data Analysis

After extraction and cleaning, detail knowledge of data like the number of the column, number of the row, the number of numerical columns and categorical column, the number of Boolean and check if any missing or na-values. It was gathered that 56% female and 44% for male responded quantity of the data generated, 61% travel for business, 31% travel for personal travel and 8% are for mileage ticket, 6% of the respondent are between the age of 0-19 years, 12% of the respondent are between the age of 20-29, 20% of respondent are between the age range 30-39, 21% are between the age range of 40-49, 16% are between the age of 50-59, 11% are between the age of 60-69, 7.5% are between the age of 70-79 years while 4.5 are between the age of 80 and above see the figure 4 for more details

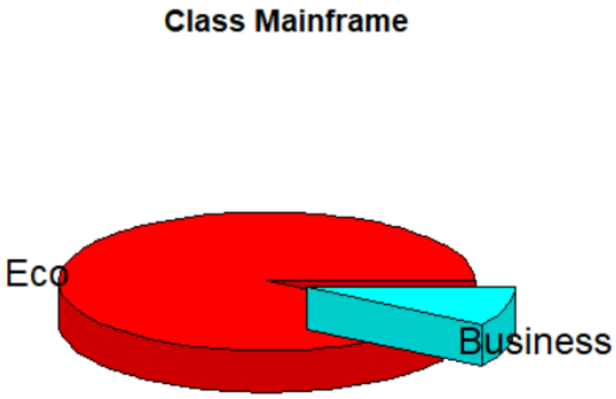


Figure 3: Data sample

However it was observed that majority of the business travel happen at the middle of the month and personal travel at the end of the month this might be the as a result of holidays or other business travels, the majority of the traveller has Texas, California, as their destination state followed by Florida, Georgia, and Illinois.

5.3 Oversampling of data

From the figure 3 it shows the data was not balanced it consists about 91% or Eco class and 9% of business class, (SMOTE) method of balancing imbalance dataset was applied, this a process of upsurging the smaller ration of the minority class to the exact or equivalent ratio with the majority class but the original majority class remains untouched. this class imbalance in data set is best resolved using two major methods: downsampling and Oversampling method (SMOTE). The oversampling method was used in this project because more data was needed for better performance and to enhance accurate prediction. See the figure 4

```
prop.table(table(overbaldata$Class))
barplot(prop.table(table(overbaldata$Class)),
colours=rainbow(2),
ylim = c(0, 1),
main = ”class mainframe”)
```

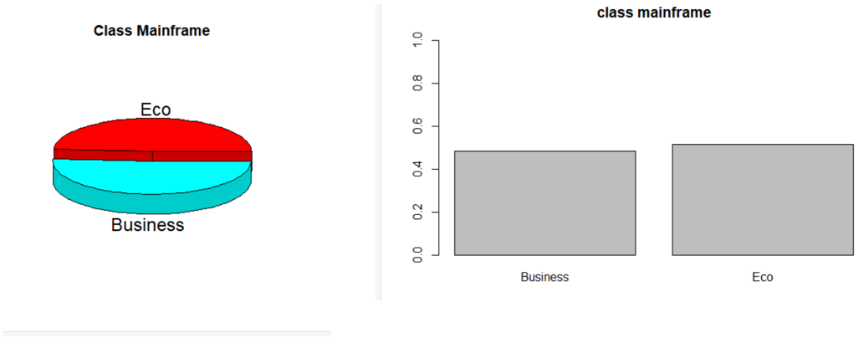


Figure 4: Balanced Data sample

5.4 Dimensionality reduction

The dimensionality reduction is vita in data set that contain several variable, this creating an artificial data from the original data, it minimizes the amount of random variable and reduced noise in data. A data set with high dimensionality will affect the performance of a model. A principal component Analysis (PCA) was employed in this project, it is known as linear dimension reduction is done by selecting variable with significant value and removing the non-significant variable. this to identify the variables that are significant at 95% confident interval at P = 0.5 or below 0.5 figure 4.

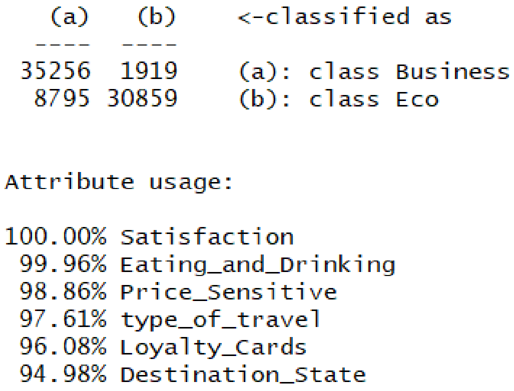
5.5 Model selection

Four models; Decision trees, Random Forest, K-Nearest Neighbour, Logistic Regression was used in this research to prediction of business Class users, the selection of the four models was based on the Research Question, aim and objective of the research project. Secondly from the literature review conducted, the models are known for their outstanding performance in regression, classification, and they are also good learners and gives out interpretable results. (O'Connell and Williams; 2005) (Rian et al.; 2019) recommends the use of a high hierarchy model for more accurate prediction after they predict with SPSS.

5.6 Implementation of Decision Tree (C50)

similar to another supervised learning algorithm, it is an advance of C4.5 models, however, it was adopted in this research because of its powerful performance in the classification prediction, its architecture it consist of nodes that form the root trees, it has a structure like a tree that consists of rooted trees, edges, and leaves.works by information entropy, it creates subgroups of data, and later split the subgroup data into several close related clusters (Rokach and Maimon; 2005). C50 version of the decision tree work is a robust version it works fine with numeric and categorical data. figure 5 show the number of best 3 performing variable, and the number of 35,256 and 33,0859 correctly classified and 1,919 + 8,795 wrongly classified in 1.4 seconds.

. Lior Rokach and Oded moaimon (2005)



```
  (a)   (b)    <-classified as
 ----  ----
35256  1919    (a): class Business
 8795 30859    (b): class Eco


Attribute usage:

100.00% Satisfaction
 99.96% Eating_and_Drinking
 98.86% Price_Sensitive
 97.61% type_of_travel
 96.08% Loyalty_Cards
 94.98% Destination_State
```

Figure 5: C50 model performance

5.7 Implementation of Random Forest

This model architecture is like the decision trees architecture except for the fact that the RF is an ensemble of decision trees, it uses the majority voting technique to aggregate the trees when performing the classification and regression. It performs well in the regression and multiclass classification; it trains and predicts fast. RF also utilizes the bootstrap aggregating known as bagging to reduce the overfitting and increase accurate performance. The advantage of RF is that it gives a more accurate estimation of the error rate for classification and Root mean Square for Regression.(Adele et al.; 2011), after the implementation of random forest it performs excellently well in 79.2% accuracy prediction.

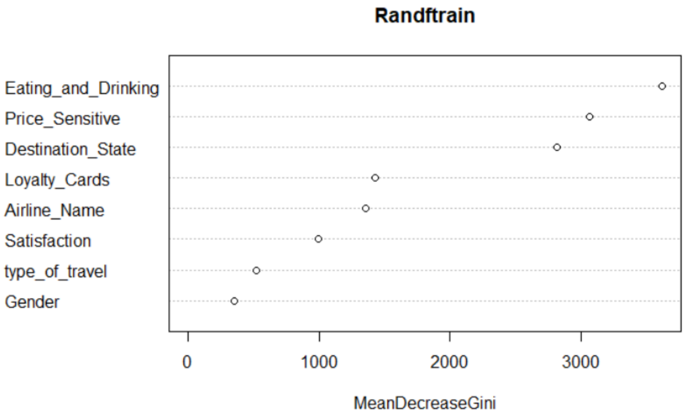
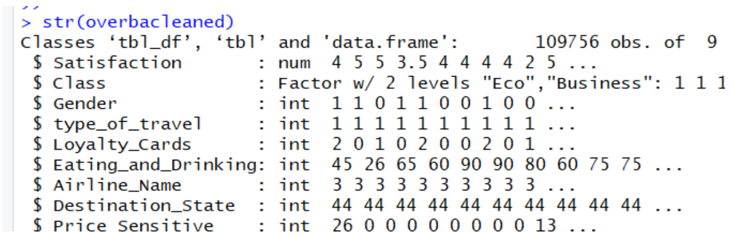


Figure 6: Variable significant

5.8 Implementation of Logistic Regression

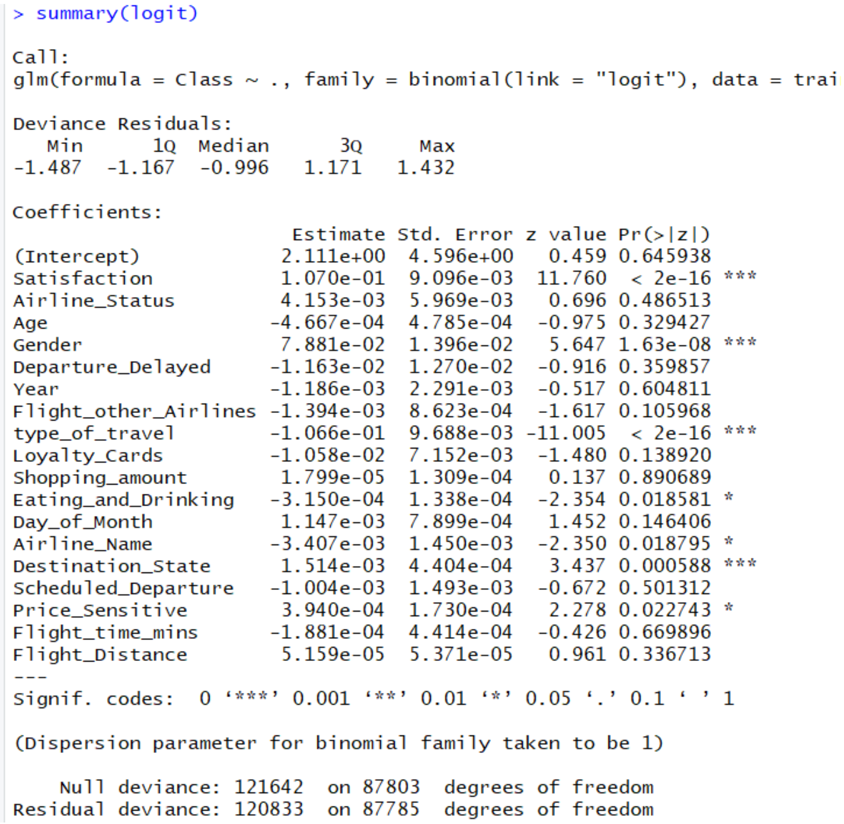
This type of model is appropriate for regression analysis to conduct when the dependent variable is dichotomous. It can also use to describes the data and explain the relationship with the dependent variable (binary) and one or more nominals. It does not perform well in high multicollinearity or outlier among the variables. Logistic Regression can be applied to predict the probability of getting lung cancer (yes or No) or the probability of staff attrition or the probability of flying business class or Economy. figure 7 show that there is 109,756 observation, 9 variable in the dataset.



```
> str(overbacleaned)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       109756 obs. of  9
 $ Satisfaction       : num  4 5 5 3.5 4 4 4 4 2 5 ...
 $ Class              : Factor w/ 2 levels "Eco","Business": 1 1 1
 $ Gender             : int  1 1 0 1 1 0 0 1 0 0 ...
 $ type_of_travel     : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Loyalty_Cards      : int  2 0 1 0 2 0 0 2 0 1 ...
 $ Eating_and_Drinking: int  45 26 65 60 90 90 80 60 75 75 ...
 $ Airline_Name       : int  3 3 3 3 3 3 3 3 3 3 ...
 $ Destination_State  : int  44 44 44 44 44 44 44 44 44 44 ...
 $ Price Sensitive    : int  26 0 0 0 0 0 0 0 0 13 ...
```

Figure 7: Data Information

The figure 8 show the summary or the logistic regression model the Deviance Residuals, it also shows that at 95% confident interval variables with ””*”” are significant of the prediction so we can say that, that at 95% degree of confident that variable with ”*” like Satisfaction, Gender, Type of travel, Eating and drinks, destination State, Price sensitive, contributes to the prediction.



```
> summary(logit)

Call:
glm(formula = Class ~ ., family = binomial(link = "logit"), data = trai

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.487   -1.167   -0.996    1.171    1.432

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.111e+00  4.596e+00   0.459 0.645938
Satisfaction           1.070e-01  9.096e-03  11.760  < 2e-16 ***
Airline_Status         4.153e-03  5.969e-03   0.696 0.486513
Age                   -4.667e-04  4.785e-04  -0.975 0.329427
Gender                 7.881e-02  1.396e-02   5.647 1.63e-08 ***
Departure_Delayed     -1.163e-02  1.270e-02  -0.916 0.359857
Year                  -1.186e-03  2.291e-03  -0.517 0.604811
Flight_other_Airlines -1.394e-03  8.623e-04  -1.617 0.105968
type_of_travel        -1.066e-01  9.688e-03 -11.005  < 2e-16 ***
Loyalty_Cards         -1.058e-02  7.152e-03  -1.480 0.138920
Shopping_amount        1.799e-05  1.309e-04   0.137 0.890689
Eating_and_Drinking   -3.150e-04  1.338e-04  -2.354 0.018581 *
Day_of_Month           1.147e-03  7.899e-04   1.452 0.146406
Airline_Name          -3.407e-03  1.450e-03  -2.350 0.018795 *
Destination_State      1.514e-03  4.404e-04   3.437 0.000588 ***
Scheduled_Departure   -1.004e-03  1.493e-03  -0.672 0.501312
Price_Sensitive        3.940e-04  1.730e-04   2.278 0.022743 *
Flight_time_mins      -1.881e-04  4.414e-04  -0.426 0.669896
Flight_Distance        5.159e-05  5.371e-05   0.961 0.336713
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 121642  on 87803  degrees of freedom
Residual deviance: 120833  on 87785  degrees of freedom
```

Figure 8: Significant Variable

5.9 implementation of K-Nearest Neighbour (KNN)

K-Nearest Neighbours algorithm is one of the simplest supervised machine learning model, its can be used for classification, its working principle when performing classification base on the closest training examples in the feature spaces. It is said to be in its best state of classification when dealing text categorization and has Low tolerance to noise.(Imandoust and Mohammad; 2011)

5.9.1 Application of KNN.

1. Text mining: KNN Algorithm is known for good performance when handling text categorization or text mining.

2. For forecasting stock market: predicting or forecasting the price stock market, it will enhance company performance or forecasting currency exchange rate and identify debit or credit cards usage.

3. It can also use in the medical field to predicts the patient with a heart attack if he/she will suffer the second attack or apply in politics to predict a potential voter as an "I will vote" and "I will not vote" candidates. 4. It can also apply in market basket analysis, and in the recommendation of YouTube music.

6 Evaluation

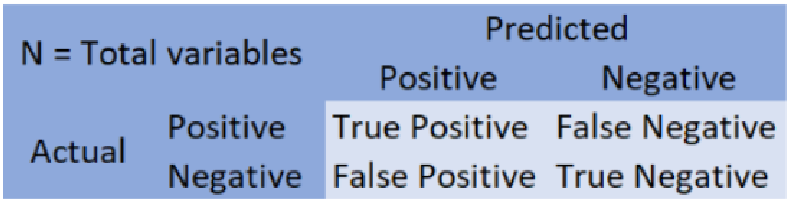
The accuracy of machine learning model is how efficiently a model predicts or classify the data, the performance of machine learning model is estimated base on the accuracy of the prediction, this accuracy of prediction work properly when the data set is balance, the data set is split into two sections: 70% training is used to train a model and 30% testing data is used to test the model, the performance of models is base of the confusion matrix result. The confusion matrix is based on the following parameter.

True Positive(TP): This is a numerical value of instance that are correctly classified as been a business class user, it is measured by TP/TP+FN

False Negative(FN): this a (type 11) error this is a numeric value of instance that are incorrectly classified as not been a member of business-class user it is measure by FN/TP+FN

False Positive(FP): This a (type 1) error, the numerical values of instance that are incorrectly classified as been a member of business class users, it is measured by FP/TN+FP

True Negative(TN): This is a numeric value of instance that are correctly classified as not a member example Eco class user that are classified as Eco class user it is measured by TN/TN+FP



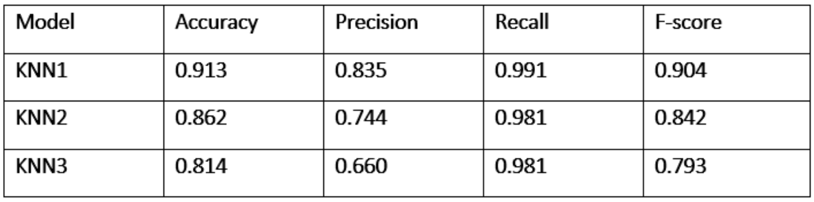
N = Total variables		Predicted	
		Positive	Negative
Actual	Positive	True Positive	False Negative
	Negative	False Positive	True Negative

Figure 9: confusion matrix

6.1 KNN model Evaluation /case Study 1

K-Nearest Neighbour was used in predicting the business class users, the dataset was divided into two sections Training and Testing sets 75% and 25% respectively. Confusion matrix was employed to predict the best K-value, It was trained from when K=1-100 to enable a wide range of probability of predicting a good result. The best result is when

K=1 with 91.3% accuracy, 90.4% of f1-score, 99.1% of recall and 83.5% of precision value figure10

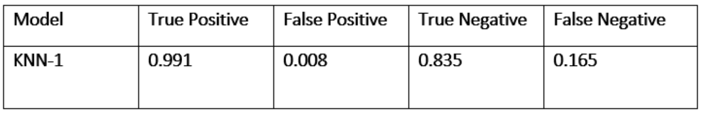


Model	Accuracy	Precision	Recall	F-score
KNN1	0.913	0.835	0.991	0.904
KNN2	0.862	0.744	0.981	0.842
KNN3	0.814	0.660	0.981	0.793

Figure 10: KNN Accuracy

6.1.1 Result explanation

From the figure 11 show the confusion matrix result at 95% confident interval KNN model predictions, when K=1, correct prediction of 99.1% True positive prediction of Business class user, 83.5.0% True negative correct prediction and 0.8% and 16.5% incorrect False-positive and False Negative prediction respectively.

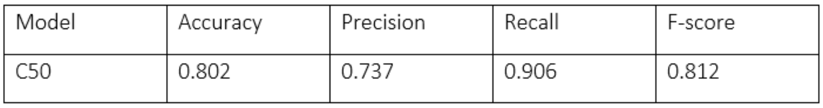


Model	True Positive	False Positive	True Negative	False Negative
KNN-1	0.991	0.008	0.835	0.165

Figure 11: KNN, at k=1 confusion matrix

6.2 C50 model Evaluation/ Case Study 2

This algorithm adopts a tree-like method when doing a classification, it uses entropy for thee measurements of data when making a prediction, the accuracy archived after implementing C50 model is 80.2% , 90.6% Recall, 73.7% of Precision and 81.2% of F1-score as shown in the figure 12 bellow.

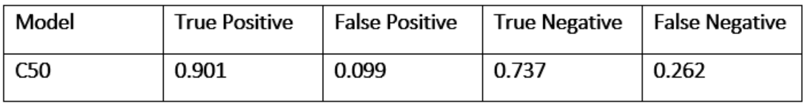


Model	Accuracy	Precision	Recall	F-score
C50	0.802	0.737	0.906	0.812

Figure 12: C50 accuracy

Figure 13 show the confusion matrix explanation of correct and wrong prediction. True Positive and True Negative, the correct prediction for True Positive rate is 90.1%

and True Negative 73.7% correct predictions, and the wrong prediction that is False positive rate and False Negative prediction of 9.9% and 26.2% respectively.

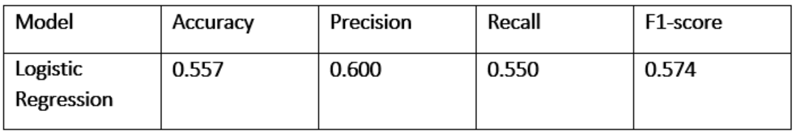


Model	True Positive	False Positive	True Negative	False Negative
C50	0.901	0.099	0.737	0.262

Figure 13: C50 confusion matrix

6.3 Evaluation of logistic regression/ Case Study 3

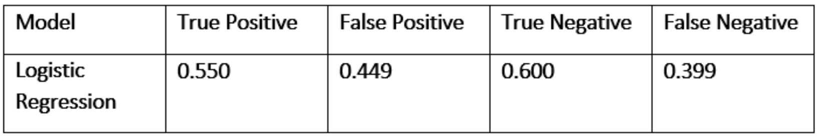
Figure 13 below shows the results for implementing logistic regression after training and adopt 70 /30 splinting of data for training and testing respectively, 55.7% of accuracy and 55.8% of precision and F1-score of 55.4%. The model did not perform well like other models.



Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.557	0.600	0.550	0.574

Figure 14: logistic Regression Accuracy

When implementing logistic regression model, I observed that price-sensitive, loyalty Card variable are not significant that probably contribute to the fair performance of the model with 55.0% of True positive, and 60.0% True negative prediction, 49.9% and 39.9% False Positive and False Negative prediction.

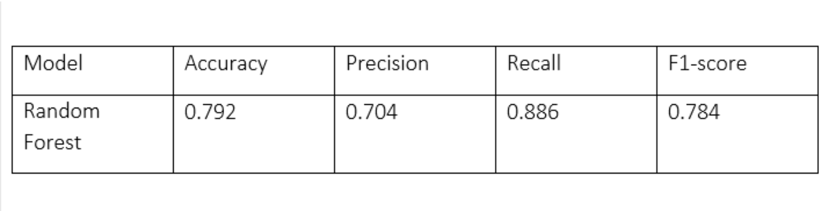


Model	True Positive	False Positive	True Negative	False Negative
Logistic Regression	0.550	0.449	0.600	0.399

Figure 15: logistic confusion matrix

6.4 Evaluation of Random Forest / Case Study N

Random forest is the ensembles of trees for making decisions by applying bagging techniques, it model is easy to implement and it has a tendency of performing good performance even in the presence of outliers, in this model archived 79.2% Accuracy, 70.4%, Precision of 88.6% Recall, 78.4% of f1-score

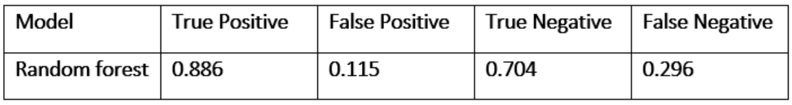


Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.792	0.704	0.886	0.784

Figure 16: Random Forest model Accuracy

6.5 confusion matrix result explanation

From the figure 17, it shows the confusion matrix results, for implementing RF model it archived 88.6% correct True positive prediction and 704% True Negative prediction with very low 11.5% and 29.6% wrong prediction of False-positive and False Negative respectively.



Model	True Positive	False Positive	True Negative	False Negative
Random forest	0.886	0.115	0.704	0.296

Figure 17: Random Forest confusion matrix

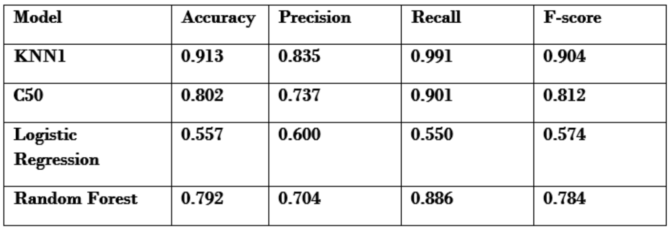
6.6 F1-Score:

F1-measure is known as the weighted average of the precision and Recall. F1-measure work perfects well over the accuracy measure because it has an advantage over the uneven class distribution, it uses positive and negative values. it is measure as;

F1 =2*(precision*Recall) / (precision + Recall).

6.7 Model comparison

All the model have been implemented Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbour with good performance, from the result in figure 18it can be deducted that KNN had the best result with an accurate prediction of 91.2%. The performance of any model depends on the nature of the dataset chosen and the dimensionality of dataset contributes to the performance of the model. KNN is a model that performs better in a dataset with low dimensional spaces. Some other models like SVM perform better in high dimensionality dataset. followed by Decision Tree with 80.2% accuracy. Logistic Regression has the lowest performance with an accuracy of 55.7% I observed that variable like Price Sensitive, loyalty Card and Gender was not significant when training the model but was highly significant when training other models, this might be the reason for the poor performance.



Model	Accuracy	Precision	Recall	F-score
KNN1	0.913	0.835	0.991	0.904
C50	0.802	0.737	0.901	0.812
Logistic Regression	0.557	0.600	0.550	0.574
Random Forest	0.792	0.704	0.886	0.784

Figure 18: Performance table

From the bar chart graph in figure 19 it shows that KNN has the best result in all with 99.1% in Recall value, logistic regression has the lowest performance both Accuracy, Recall, Precision, F1-score except in Run-time. During the process of running Decision Tree and Random Forest, it was time-consuming they produced a similar result in all the measuring parameter.

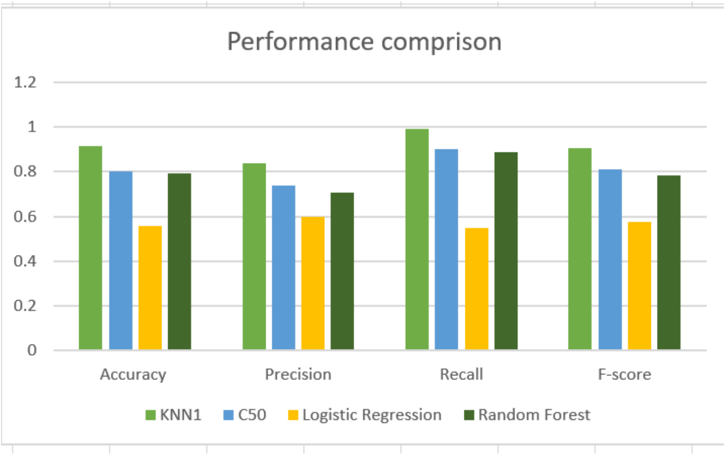


Figure 19: Model comparison

6.8 Discussion

This project was implemented stage by stage, it was challenging to find a suit dataset and the process of extracting suitable dataset, cleaning and transformations. The aim of the project was archived after pre-processing of the dataset, balancing the imbalanced data, using SMOTE method of balancing data, The pre-processing of data was the major challenge as understanding the function of the variables in the dataset and identify the significant variables.

The principal component analysis was (PCA) was applied since the data set contained 17 variables, it is considered as high dimensional, dimension reduction was applied to avoid overfitting in the model accuracy. Exploratory data analysis was conducted in all the model as the experiments involve identify the major factor that influences passenger demand, models were built with significant variables.

Eating and drinking onboard is significant to the factors that affect passenger demand but its part of service to stop because of Covid-19, it will impacts negatively to customers satisfaction.

The case studies have performed exceptionally well more than the previous models in the literature review. This research will contribute immensely to aviation industries by scientifically predicts, the passenger on business and economy route, this will increase their quality of services to passengers, leading to the likelihood of continue purchase

7 Conclusion and Future Work

After successful implementation of the models, I have concluded that the application of machine learning approach to predict the business class users and factor that affect the passenger preference during and the post-economic recession was successfully archived with the application of K-Nearest Neighbour (KNN), Decision Tree (C50), Random Forest, and Logistic Regression Models. The accurate predictions archived have answer the research question stated in the subsection 1.3 and the research objective listed in the section 1.4 of this report.

however, from the list of contributing factors I observed the major factors influencing passenger demand is; Price Sensitive, Satisfaction (quality of service) Type of travel, Eating and Drinking and Gender. most of this variables relate to finance. so, a good strategy on how to make service affordable to passenger and make profits will help in this time of economic recession, running a promo package will help as Loyalty Card was also significant variable in the prediction it will enhance Business class Demand. furthermore, the future research would focused on the geographic location to find out if it is significant to factor that affects the passenger demand, this will help to generalise findings because this research utilises participate only from united states. secondly I would have use and another dataset to predicts customer recommendation for a business class flight extracting data from social media data.

8 Acknowledgment

I would like to use this opportunity to express my gratitude to my supervisor Prof Vladimir Milosavljevic for his immense support, motivation, guardians throughout the project, I would also like to thank (IT) Department National College of Ireland (NCI) for providing us with the reliable network during our online sections. I also owe my parent appreciation for their support and encouragements.

References

Adele, c., David, R. C. and John, R. S. (2011). ensemble machine learning, methods and application, *statistics* **5**: 157–176.

Ana, B., Mirko, T. and Kalanj, Z. (2010). the impact of the global recession on the south east europe airline industrie, *Faculty of transport and traffic Engineering university of Zagreb Croatia* .

Chandrakumara, T. A., Balan, S. and Chakraboty, S. (2018). A time series analysis of the stock market during the 2007- 2009 recession, *Internationl conference on big data (Big Data)WA USA* pp. 5295–529.

Fadi, T., Suhel, H., Firuz, K. and Amanda, H. G. (2019). Data imbalance in classification, experimental evaluation., *information sciences* **513**: 978–988.

Igor, S., Damir, V. and Andrija, V. (2015). Effect of economic, crisis on the change of low-cost carriers business model, *15th international conference on transportation(IMECS 2015)* pp. 978–988.

Imandoust, S. B. and Mohammad, B. (2011). Application of k-nearest neighbor (knn) approach for predicting economic evens theoretical background, *master dissertation* .

Jian, C., xu, Y., Ou, H., Tan, Y. and Oxiao, P. (2018). A personalised flight recommendation service via cross-doman triadic factorization, *international conference on web services(ICWS)* **3**: 249–256.

Jose, F. D., Juna, J. R., Cesar, G. O., Ludmila, I. and Kunche, p. (2015). Random balance ensembles of variable priors classification for imbalance data, *Research Gate(IMECS 2018)* **85**: 96=111.

Joseph, S. B. (2020). Covide-19 and airline employment: Insights from historical uncertainty shock to industry, *transportation Research Interdisciplinary* **5**: 100–123.

Markus, F. and Florian (2011). what come next after recession airline industry scenarios and potential end game, *journal department of transportation* **17**: 19–26.

Opera, G. M. (2010). The effects of global economic crisis on the air transport of passengers in europe and romania, *Geojournal of Tourism and Geosites* **5**(1): 51–61.

O'Connell, J. F. and Williams, G. (2005). passenger's perception of low-cost airline s and services carriers,air asia and malaysia airlines., *Transport managements.* pp. 259–272.

Peter, M. (2011). Current challenge in a distressed industry, *Journal of Air Transport management* **17**: 14–18.

Praphula, K. J., Rajendra, P., Sarfraj, A., Sharma, D. and Lasksmibai, M. (2019). Airline recommendation prediction using customer customer generated feedback data, *international conference on information system and computer Network India IEEE* **4**: 376–379.

Rian, M., Stephen, R., John, D. and Scott, W. R. (2019). Creating a prediction model of passenger preference between low cost and legacy airlines, *Transportation research interdisciplinary* **3**.

Rokach, L. and Maimon, O. (2005). the application of decision trees, *Data Mining and Knowledge Discovery Handbook* **9**.

Sacha, C. (2010). the impacts of the recession on companies strategies in the irish airline industry, *Galway institute of technology* pp. 45–54.

Safak, A., Eda, A. and Serkan (2003). A airline services marketing and by domestic and foreign firms difference from the customers viewpoint, *Journal of Air Transport management* **9**: 343–351.

Santos, M. F. and Azevedo, A. I. L. R. (2008). Kdd, samma and crisp-dm: a parallel overview, *IADS-DM* .

Srinidhi, S. (2009). Development of an airline traffic forecasting model on international sector, *IEEE International conference on Automation Science and Engineering* **5**: 322–327.

Sushruta, M., Pradeep, K. M., Lambodar, J. and Gyoo, S. C. (2020). optimization of skewed data using sampling-based pre-processing approach, *Research gate* .