



The  
University  
Of  
Sheffield.

# **MACHINE LEARNING MODEL FOR PREDICTING SATISFACTION**

**CourseWork-2**

**Data Modelling and Machine Intelligence (ACS61013)**

**By**

**Ankur Tiwari**

**Registration Number : 210158775**

## Abstract

The Aim of this course work is to develop a machine learning model to predict/estimate customer satisfaction based on airport features using two different Machine Learning Algorithms . A dataset has been given consisting of 3502 datapoints and 37 features. The first step taken in this regard was Domain Analysis . It helped us in understanding the dataset , identifying its important features which helps us in further taking important decision regarding model selection , feature engineering , data cleaning etc. Next step was to clean the data , perform feature engineering to make the data ready for further processing. The cleaned and processed dataset was then subjected to PCA ( principal component Analysis ) to reduce its dimensions. Then after careful consideration of various factors Linear regression and Random Forest was selected to build machine learning models from the processed data. Once models were implemented cross-validation techniques were implemented on the models to access their performance. Learning curves were plotted for both the Models to evaluate their behaviour towards unseen data. These two models were then also compared with kNN, AdaBoost , and decision Tree . The results were then analysed to identify the best algorithm for given situation.

# TABLE OF CONTENT

|  |           |
|--|-----------|
| .....  | 1         |
| <b>MACHINE LEARNING MODEL FOR PREDICTING SATISFACTION .....</b>  | <b>1</b>  |
| <b>1.Domain Analysis .....</b>   | <b>4</b>  |
| <b>1.1 Exploratory Data Analysis .....</b>   | <b>4</b>  |
| 1.1.1 Understanding the dataset.....   | 4         |
| 1.1.2 Cleaning the Data .....  | 6         |
| 1.1.3 Analysis of Relationship between variables .....   | 7         |
| 1.2 Discuss how what you have found from your domain analysis will support and be carried over to other parts of your coursework. .... | 7         |
| <b>2. Feature Engineering &amp; Data Pre-processing. ....</b>  | <b>7</b>  |
| 2.1.1 Feature Engineering to convert variable of type (Time) to continuous Numerical value .....                                       | 7         |
| 2.1.2 Pre-Processing Data to Impute Missing Values .....   | 7         |
| 2.1.3 Pre-Processing Data to Normalize the parameters .....  | 8         |
| 2.1.4 Feature Engineering to convert variable of TYPE(DATE) in continuous value .....  | 8         |
| 2.1.5 Dimensionality Reduction using Principal Component Analysis .....  | 8         |
| 2.2 Discuss how you used your understanding of the domain from level 1 to support this task. ....                                      | 9         |
| <b>3.1 Steps taken in Dimension Reduction .....</b>  | <b>9</b>  |
| <b>3.2 Steps Taken to prevent Bias in Dataset .....</b>  | <b>9</b>  |
| <b>3.3.Feature catching Most variability in Dataset.....</b>   | <b>9</b>  |
| <b>3.4 Variables highly Correlated to Customer satisfaction .....</b>  | <b>10</b> |
| <b>4.1 Decision of Machine Learning Model.....</b>   | <b>10</b> |
| 4.1.1 Linear Regression.....   | 10        |
| 4.1.2 Random Forest.....   | 10        |
| 4.1.3 Selection of Hyper Parameters .....  | 11        |
| 4.2 Choice of software .....   | 11        |
| <b>5. Application of Cross Validation Techniques in Machine Learning Pipeline .....</b>  | <b>11</b> |
| <b>6. Evaluation of Machine Learning Model .....</b>   | <b>12</b> |
| 6.1 RMSE ( Root Mean Square Error) .....   | 12        |
| 6.2 $R^2$ ( Coefficient of determination ) .....   | 12        |
| 6.3 Learning Curves.....   | 12        |
| <b>7.Discussion .....</b>  | <b>13</b> |
| 7.1 Comparison with two ALGORITHMS not discussed in class .....  | 13        |
| 7.1.1 AdaBoost:.....   | 13        |
| 7.1.2 SVM .....  | 14        |
| 7.2 Mathematical Peculiarities of Linear Regression and Random Forest Classifiers .....  | 14        |
| 7.3 Apply appropriate metrics to compare the algorithm you have chosen with ones discussed in class .....                              | 14        |

## 1.DOMAIN ANALYSIS

The term Domain Analysis in data science refers to the process of acquiring general background information about the field to which data Science is being applied. It is the practice of determining the lineage of the data, identifying its origin, to ask questions such as: what is data trying to tell? What is its purpose? Domain Analysis helps in applying correct methods as well as in judging the performance of applied models. [1].

The revelations of the domain analysis have been mentioned below.

Airport Customer experience can be defined as how an individual felt upon their interactions at various channels (personal, online, service booths, etc.) with the airport community. [2]

According to peak-end theory the customers judge an experience based on the fact that how they felt at its emotional peak and at its end. [3]. However most intense movements differ for each passenger at an airport, and this is key to manage customer satisfaction level. To attain a good level of satisfaction airport needs to map customers journey and define the drives of satisfaction at each touchpoint. [2]

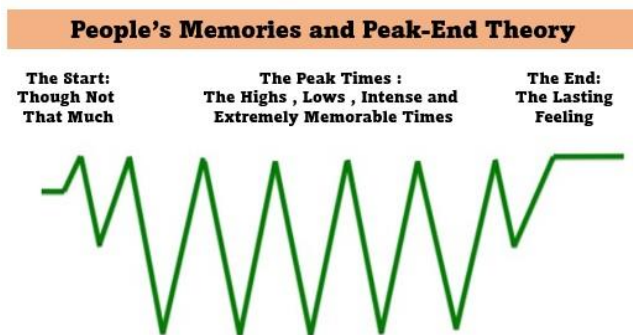


FIGURE 1: PEAK-END THEORY

Based on a report published by **ACI World's Airport Service Quality Program** the key features that are utmost important for customer satisfaction are the airport ambience, the discretionary time which comprises of activities such as visiting retail outlets, enjoying food and beverages, entertainment activities, use of Wi-Fi, security process and Human factor. [2]

Global Passenger Satisfaction Drivers & their World Marginal Impact in %

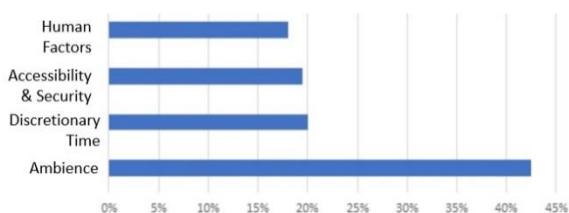


Figure 2 ACI Survey ResULTS [2]

It's Imperative to segment the offerings for passengers into different traveler types. This helps in understanding consumer behavior and their preferences in a better manner. [4]. The segments are:

**Business:** This class of passengers emphasize more on service, speed, and convenience with very little emphasis on prices.

**Luxury:** This segment is prepared to pay a premium for the unsurpassed experience. What they desire is peace and relaxation.

**Family:** This segment is driven with familiarity and Convenience.

**First-time and occasional traveler:** One of the most overlooked segments. They may get intimidated by various aspects of travel as they do not have any prior experience. This group requires extra guidance to make their experience better.

**Los Angeles International airport** implemented **Metis an artificial Intelligence-backed data analytics system** to do the customers sentiment analysis. The following are the findings of this system. [5]

- **Baggage Tracking:** Baggage handling is one of the key components of customer experience. Irrelevant of how good experience is provided if a customer arrives to destination and bag isn't there all the efforts are in vain.
- **Traveler traffic management:** One of the most painful parts of the passenger's journey is long security line (it causes uncertainty as to how long it will take to get through the security checks).

### 1.1 EXPLORATORY DATA ANALYSIS

It is the process of investigating data for discovering patterns in data, spot hidden anomalies, test hypothesis, and check assumptions with the help of statistics and graphical representations. [6]

Objectives of EDA are:

- It helps in cleaning and filtering data from redundancies.
- Help us in understanding relationship between variables which gives us a wider perspective on data.

Steps Involved in EDA are:

- Understand the Data
- Clean the Data
- Analysis of Relationship between variables.

The following steps were taken during EDA for given Dataset.

#### 1.1.1 UNDERSTANDING THE DATASET

The given dataset is made up of features that customers look for in an airport and level of their satisfaction with the airport. The dataset contains 3502 data points and 37 features.

Out of 37 features **34 are of Type Numerical** and 3 features i.e., Quarter (of the year) is Plain Text, Date recorded is of type Date & Time and Departure time is of type Plain Text. All the Numerical type variables are categorical variable (Rating) from 0 to 5 where 0 represents least satisfaction and 5 represent maximum satisfaction level.

**On analyzing the dataset using orange and other tools following observations were made:**

Speed of baggage delivery has 5 outliers at **9.30 AM, 10.40 AM, 12.30 PM, 12.45 PM and 06.05 PM** which indicates that there was a **surge in the number of customers that were moving out around these times** indicated by the higher number of ratings around these times.

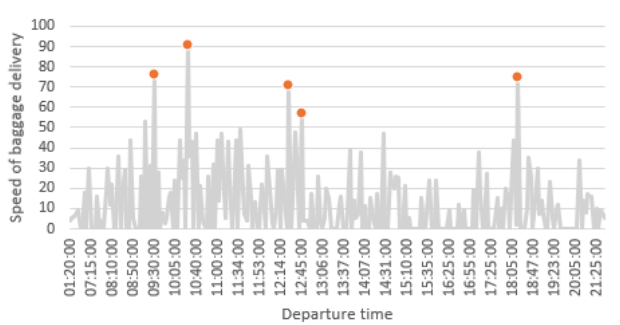


FIGURE 3 SPEED OF BAGGAGE DELIVERY VS DEPARTURE TIME

There was **high dissatisfaction for custom inspection** among the customers of business/executive lounges. Which is natural as these customers are very busy and dislike waiting for any reason.

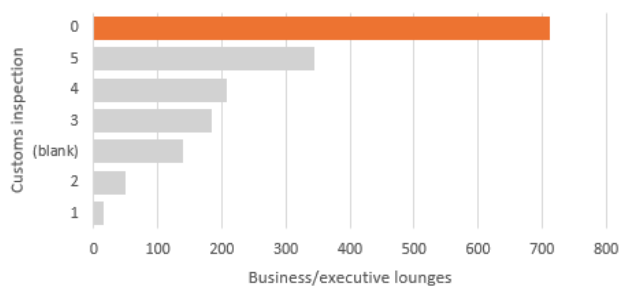


FIGURE 4 SATISFACTION LEVELS FOR CUSTOM INSPECTION FOR BUSINESS AND EXECUTIVE CLASS

A major chunk of customers rated 0 or 1 on a scale of 5 which implies **Overall Satisfaction** level of the customers was **low**.

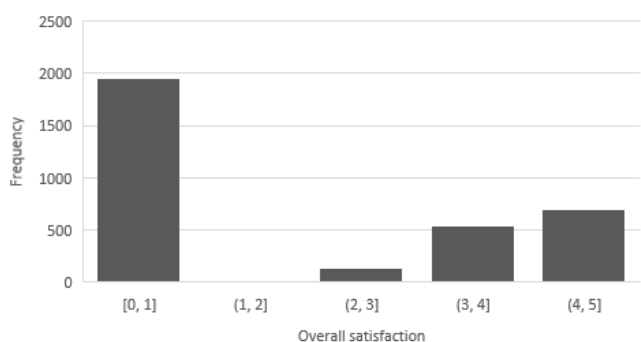


FIGURE 5 FREQUENCY DISTRIBUTION OF OVERALL SATISFACTION

**Most customers** were **happy** with the **comfort provided at the waiting gate area** which is indicated by the frequency distribution of their rating given for this criterion.

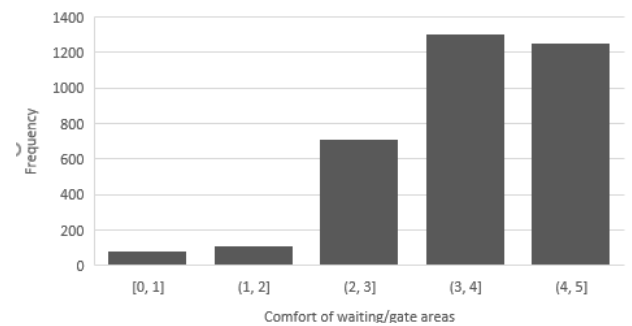


FIGURE 6 FREQUENCY DISTRIBUTION OF COMFORT OF WAITING/ GATE AREAS

'Availability of washrooms' has outliers at 'Departure time': **09:35 AM, 3:00 PM and 04:53 PM**. This is indicative of **high demand** in usage of these facilities at these times.

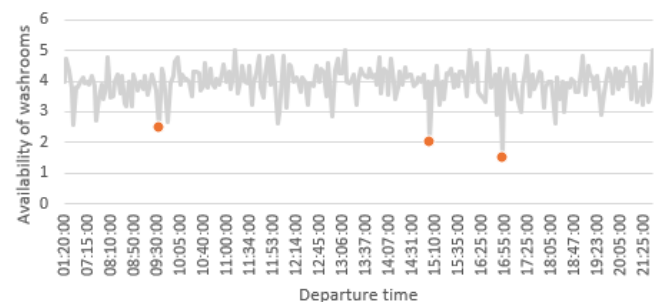


FIGURE 7 AVAILABILITY OF WASHROOMS AT DIFFERENT TIMES OF DAY

**Most customers** were **happy** with the **ambience of the Airport** indicated by the high frequency of ratings between 4 and 5.

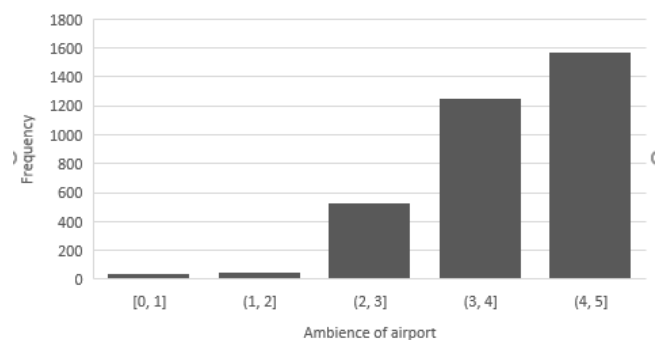


FIGURE 8 FREQUENCY DISTRIBUTION OF RATINGS FOR AIRPORT AMBIENCE

Frequency Distribution of **parking facilities** indicates **high dissatisfaction** among customers for the parking facilities.

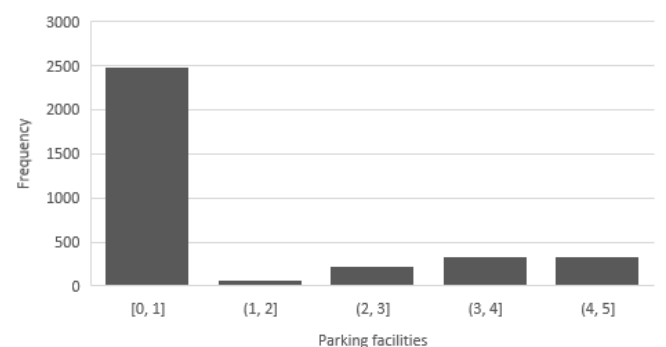


FIGURE 9 FREQUENCY DISTRIBUTIONS FOR RATING OF PARKING FACILITIES

### 1.1.2 CLEANING THE DATA

There are some inconsistencies in the dataset that need addressal. As mentioned in previous section there are some **outliers** in certain categories such as **Speed of Baggage deliveries** and **Availability of Washrooms** at certain times of days. These will be addressed in subsequent section.

#### 1.1.2.1 CLEANING THE INCONSISTENCIES IN DATE RECORDED AND MAKING IT UNIFORM

Some **inconsistencies** were observed in **Date recorded** and **Departure time** column of the dataset. Majority of entries in **Date recorded** are in **MM-DD-YY** format whereas some are in **MM/DD/YY** format

```
In [93]: for i in range(1597,1602):
         print(df1.Date_recorded[i])
```

```
03-02-2016
01/14/2015
03-03-2015
01-04-2016
01/14/2015
```

FIGURE 10 INCONSISTENCY IN DATE

**Python code** has been used to clean the data using **Jupyter Notebook**. The code helps in making the date in uniform format. The code block has been shown in **Figure 11**.

```
In [75]: for i in range(0,3501):
         if "/" in df.Date_recorded[i]:
             x= df.Date_recorded[i].split("/")
             print(x)
             d= x[0]+ "-" + x[1] + "-" + x[2]
             df.Date_recorded[i]=d
         df.to_csv('date_formatted.csv')
```

FIGURE 11 CODE BLOCK FOR CLEANING THE DATE

On application of Code blocks shown in **Figure 11** the dataset obtained is shown in **Figure 12**

```
for i in range(1597,1602):
    print(df1.Date_recorded[i])

df2=pd.read_csv('date_formatted.csv')
for i in range(1597,1602):
    print(df2.Date_recorded[i])
```

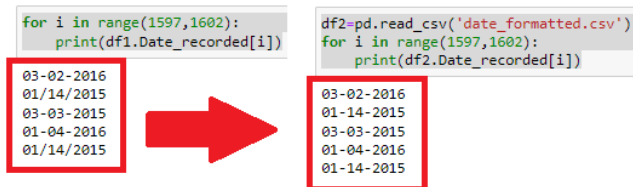


FIGURE 12 DATASET WITH CLEANED DATE\_RECORDED COLUMN

#### 1.1.2.2 CLEANING THE INCONSISTENCIES IN DEPARTURE TIME

Similar to Date recorded column there were **inconsistencies** in **Departure time** column as some of the times were in **24 hours' time format** and some were given in **12 Hours' time format** as shown in **Figure 13**

```
: for i in range(2450,2456):
    print(df1.Departure_time[i])
```

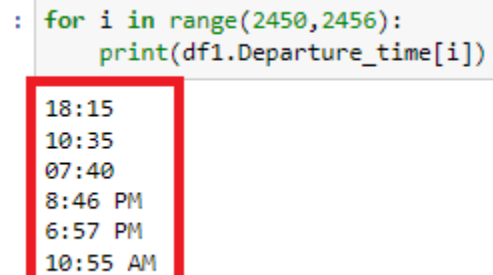


FIGURE 13 INCONSISTENCIES IN DEPARTURE TIME COLUMN

The code block shown in **Figure 14** is used to remove all the inconsistencies in Departure time column and make it uniform in 24 hours format.

```
In [66]: for i in range(0,3501):
         if 'AM' in df.Departure_time[i]:
             y=df.Departure_time[i].split(" ")
             x=y[0].split(":")
             if x[0]!='12':
                 x[0]="00"
             mod_time= x[0]+":"+x[1]

         elif 'PM' in df.Departure_time[i]:
             y=df.Departure_time[i].split(" ")
             x=y[0].split(":")
             if int(x[0])<=11:
                 mod_time= str(int(x[0])+12)+":"+x[1]
             else:
                 mod_time=x[0]+":"+x[1]

         else :
             y=df.Departure_time[i].split(":")
             if int(y[0])<10:
                 y[0]="0"+ str(int(y[0]))
             mod_time= y[0]+":"+y[1]
             df.Departure_time[i]=mod_time
         df.to_csv('time_formatted.csv')
```

FIGURE 14 CODE BLOCK TO REMOVE INCONSISTENCIES FROM DEPARTURE TIME FORMAT

On implementation of the code block shown in Figure 14 the dataset obtained is shown in Figure 15

```
: for i in range(2450,2456):
    print(df1.Departure_time[i])

df4=pd.read_csv('time_formatted.csv')
for i in range(2450,2456):
    print(df4.Departure_time[i])
```

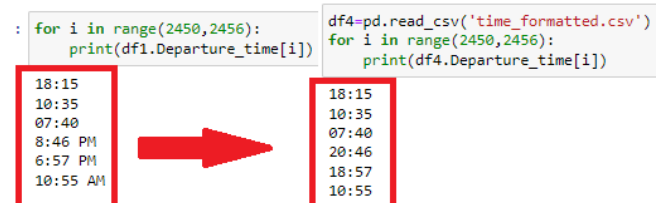


FIGURE 15 DATA SET WITH CLEANED DEPARTURE TIME

#### 1.1.2.3 REMOVING THE FEATURE (QUARTER) FROM THE FINAL DATASET

The feature **Quarter** has been dropped because it is basically segmenting the dates and signifies same information **Date recorded** so we are not using this feature. This is done using **Select column Widget** in orange. Dropping unnecessary variables is an important aspect of Data cleaning.

### 1.1.3 ANALYSIS OF RELATIONSHIP BETWEEN VARIABLES

In order to determine the correlations between various parameters **correlation widget in orange** is used. Correlation between various parameters is shown in **Figure 16**. We have selected Pearson Correlation. And selected (All combinations) to obtain the following table

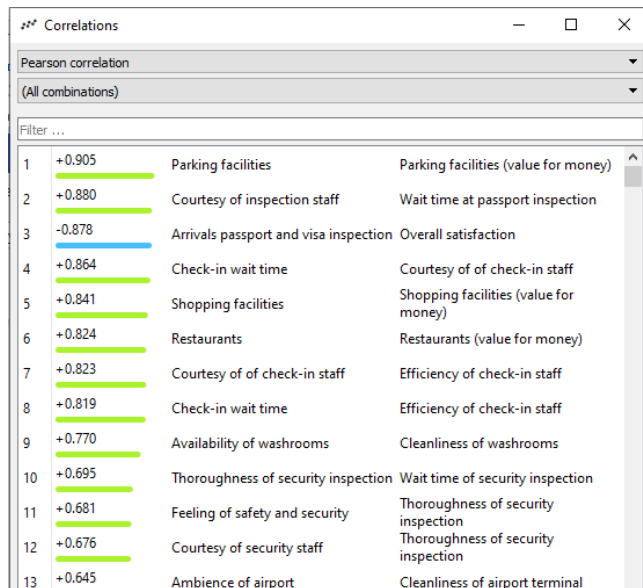


FIGURE 16 CORRELATION BETWEEN VARIOUS PARAMETERS

Positive correlation implies to the fact that both the variables move in same direction. On the contrary Negative correlation implies increase in value of one variable reduces the value on another variable. Neutral or Zero correlation implies variables are unrelated.

High correlation between parameters deteriorates the performance of the algorithm. The problem is called multicollinearity. The problem of correlation is dealt with in next section i.e., Feature Engineering.

### 1.2 DISCUSS HOW WHAT YOU HAVE FOUND FROM YOUR DOMAIN ANALYSIS WILL SUPPORT AND BE CARRIED OVER TO OTHER PARTS OF YOUR COURSEWORK.

During the domain analysis the given dataset was explored using **python and orange pipelines**. Various interesting insights about data were found as mentioned in above sections. **Section 1** helps us in understanding the lineage of data, what are the factors important to satisfy a customer in an airport, customer segmentation, and most important features responsible for customer satisfaction. **Section 1.1.1** help us in **realizing the importance of various parameters** over customer satisfaction. In **section 1.1.2** **irregularities** were found in data and certain **measures were taken to clean the data** and engineer the features. **Section 1.1.3** talks about correlation between various variables. All these sections will collectively contribute by helping us take decisions for feature engineering and Model Selection.

### 2. FEATURE ENGINEERING & DATA PRE-PROCESSING.

The steps involved in Engineering the Data features and pre-processing them have been mentioned in this section.

#### 2.1.1 FEATURE ENGINEERING TO CONVERT VARIABLE OF TYPE (TIME) TO CONTINUOUS NUMERICAL VALUE

The **Departure time** column was in **time format** which was difficult to use for principal component analysis so to **convert it into a numerical value** the cold block is shown in Figure 17.

```
df=pd.read_csv('date_numeric.csv')
for i in range(0,3501):
    x=df.Departure_time[i].split(":")
    time_min=(int(x[0])*60)+int(x[1])
    df.Departure_time[i]=time_min
df.to_csv('cleaned_data.csv')
```

FIGURE 17 CODE BLOCK TO CONVERT TIME TO CONTINUOUS VALUE

The code shown above converts the Recorded time value to minutes according to time of day. This is shown in **Figure 18**

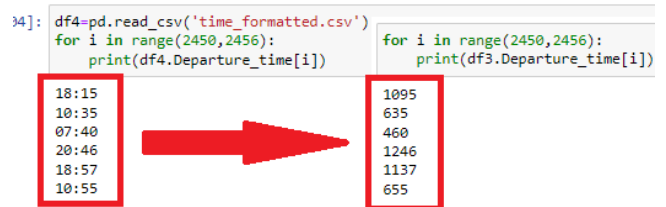


FIGURE 18 CONVERSION OF TIME FROM 24 HOURS FORMAT TO NUMERICAL VALUE

#### 2.1.2 PRE-PROCESSING DATA TO IMPUTE MISSING VALUES

Initially there were some **missing values** in the data as shown in **Figure 19** to analyse the missing values the **feature statistics widget in orange** was used. The **missing values were imputed using Preprocess widget in orange**.

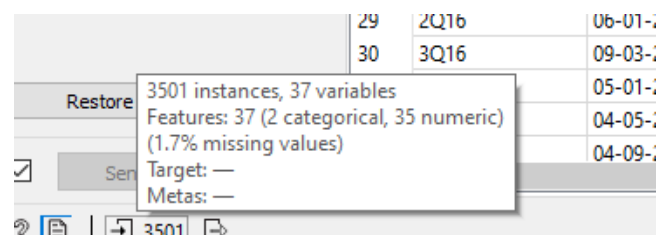


FIGURE 19 FIGURE SHOWING TOTAL PERCENTAGE OF MISSING VALUES IN DATASET

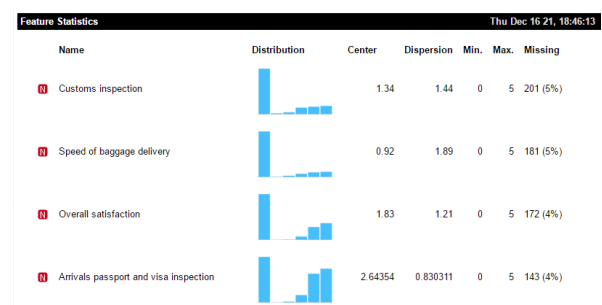


FIGURE 20 FIGURE SHOWING PERCENTAGE OF MISSING VALUES FOR EACH PARAMETERS.





FIGURE 21 STEPS TAKEN IN PRE-PROCESS WIDGET

### 2.1.3 PRE-PROCESSING DATA TO NORMALIZE THE PARAMETERS

In Preprocessing widget all the **parameters are normalized to interval [0,1]** this is one of the most important steps in **feature engineering**.

Machine learning algorithms like **linear regression**, **logistic regression**, **neural network** uses gradient descent algorithm for optimization that needs data to be scaled. Moreover, **Distance based algorithms** like **KNN**, **K-means** and **SVM** are most impacted by feature range because they use distance between data points for determining similarity. However, Tree based algorithm remains insensitive to scale of the feature. [7]

Here we need to normalize our features because although other features lie in the same range i.e. [0,5] but **Departure\_time** is in minutes and its value is quite large as compared to the value of other variables so it will create a bias in final prediction. If we normalize only **Departure\_time** to interval [0,1] then it becomes insignificant as compared to other features and this will result in loss of data. so, we normalize all the features collectively in the interval [0,1]. Another reason for normalizing the features is that this will also be useful for performing PCA (Principal Component Analysis).

Normalization of features can be done using Preprocess widget as shown in **Figure 22**.

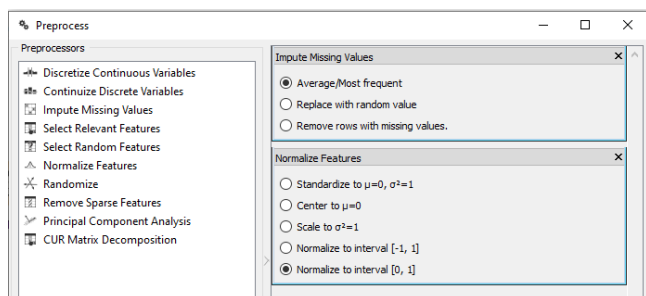


FIGURE 22 PREPROCESSING WIDGET USED TO NORMALIZE FEATURES

### 2.1.4 FEATURE ENGINEERING TO CONVERT VARIABLE OF TYPE (DATE) IN CONTINUOUS VALUE

The **Date recorded** column was in **date format** which was difficult to use for principal component analysis so to **convert it into a numerical value** all the dates were converted in days where start of 2015 is taken as 0 value and all the dates after that is converted into day. For example, 22-06-2016 can be converted to day using the following relation.  $365 + (06-1) * 30 + 22 = 537$  days. This is shown in **Figure 23**

| Quarter | Date_recorded | Dept       | Quarter | Date_recorded | Depart |
|---------|---------------|------------|---------|---------------|--------|
| 0       | 3Q16          | 09-04-2016 | 0       | 3Q16          | 464    |
| 1       | 2Q16          | 05-01-2016 | 1       | 2Q16          | 370    |
| 2       | 2Q16          | 04-07-2016 | 2       | 2Q16          | 549    |
| 3       | 3Q16          | 09-02-2016 | 3       | 3Q16          | 404    |
| 4       | 3Q16          | 08-04-2016 | 4       | 3Q16          | 463    |
| ...     | ...           | ...        | ...     | ...           | ...    |
| ...     | ...           | ...        | ...     | ...           | ...    |

FIGURE 23 CONVERSION OF DATE TO CONTINUOUS VALUE

This is done using python code the code-block used for this is shown below

```
In [26]: for i in range(0,3501):
          x=df.Date_recorded[i].split("-")
          if x[2]=='2015':
              y=0
          elif x[2]=='2016':
              y=365
          elif x[2]=='2017':
              y=730
          days=y+((int(x[1])-1)*30)+int(x[0])
          df.Date_recorded[i]=days
          df.to_csv('cleaned_dataset_final.csv')
```

FIGURE 24 CODE BLOCK FOR CONVERSION OF DATE TO CONTINUOUS NUMERICAL VALUE

### 2.1.5 DIMENSIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS

This is a dimensionality reduction technique that is used to reduce the dimensions of large datasets. The idea is to transform large set of variables into smaller set that still contains the most of information that is contained by the smaller set.

Reducing the dimensions causes reduced accuracy but the idea behind PCA (Principal Component Analysis) is to trade little accuracy for simplicity. It's easier to explore and visualize smaller dataset. Moreover, Different Models works faster on a smaller dataset. [8]

In PCA input variables are combined in a specific way and the least important variables are dropped retaining the important part of all variables. The advantage of newly obtained variables is that all are independent of each other, so we don't need to worry about correlation as it is automatically managed by PCA.

In order to perform PCA an orange pipeline is being used.

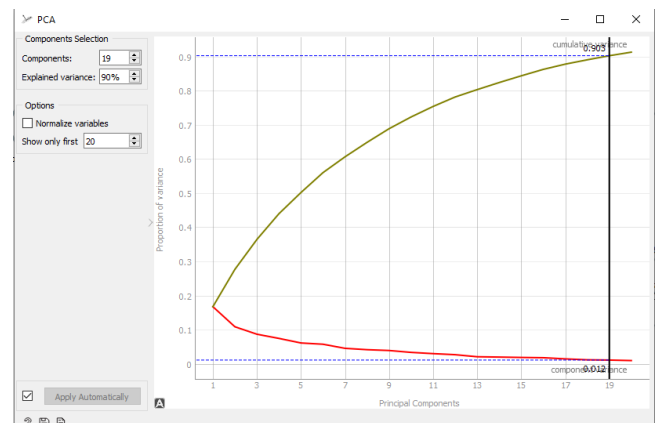


FIGURE 25 CURVE SHOWING 90 PERCENT VARIANCE WITH 19 COMPONENTS



We have obtained 90 Percent explained Variance with 19 Principal Components. **Explained Variance refers to the variance explained by each of the principal components(eigenvectors).**

## 2.2 DISCUSS HOW YOU USED YOUR UNDERSTANDING OF THE DOMAIN FROM LEVEL 1 TO SUPPORT THIS TASK.

**Data cleaning** is one of the most important aspects in any machine learning algorithm. The rule “Garbage in Garbage out” applies here. Using an uncleaned dataset may yield models that may break down or sometimes give very optimistic performance. It is therefore very Important to Clean the data before feeding it to a machine learning pipeline. Moreover, **Feature Engineering and Pre-processing** helps us prepare our data as per the requirement of the model. For example, **normalizing the data for PCA and Linear Regression.**

The **understanding of Domain** is very useful for successfully completing this task as it helps us in recognizing which variables are to be kept for further analysis and which variables to be dropped. Moreover, to give accurate prediction, one needs to know what algorithm to be used and what kind of data is required for that algorithm. This is the reason we **converted variables of date and time as datatypes to continuous values.**

## 3.1 STEPS TAKEN IN DIMENSION REDUCTION

Dimension Reduction has been obtained using **PCA (Principal Component Analysis)**. PCA has been Carried out with the help of **Orange Pipeline** using **PCA widget**. However, the steps involved in PCA are mentioned below.

- Normalization of the dataset.
- Calculation of covariance matrix for dataset features
- Calculation of eigenvalues and eigenvectors for the covariance matrix.

- Sorting eigenvalues and their corresponding eigenvectors
- Picking k eigenvalues and forming a matrix of eigenvectors.
- Transforming the original matrix.

The above steps are carried out for PCA when using Python or MATLAB but are done by the Orange automatically.

## 3.2 STEPS TAKEN TO PREVENT BIAS IN DATASET

The potential type of biases that can be generated in our dataset and steps taken to avoid these are mentioned below:

**Exclusion bias:** This type of bias occurs due to exclusion of data. However, we have removed only one parameter i.e., **Quarter** and that too after careful consideration and correlation analysis. Moreover, we are using PCA so PCA automatically retains important information from all variables hence exclusion bias can be avoided in our model.

In normal practice we combine variables with high correlation but in this scenario because of using PCA all the variables are independent.

In order to avoid bias we performed domain analysis to see any irregularities in our data.

When some parameter has very large value as compared to other parameters it can result in bias so in order to avoid bias, we have normalized all the parameters to same scale.

## 3.3.FEATURE CATCHING MOST VARIABILITY IN DATASET

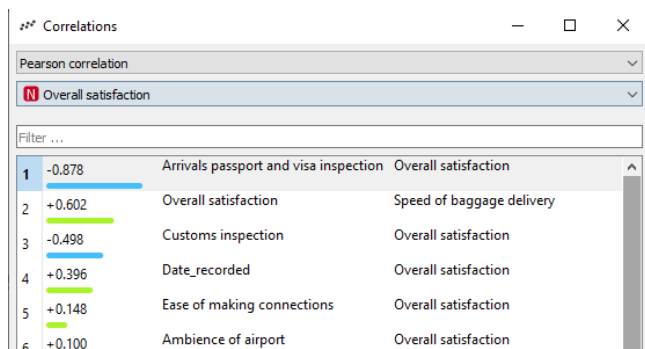
The principal components showing highest variability and the features contributing to these Principal components are shown in **Table 1**.

TABLE 1 TABLE SHOWING THE PRINCIPAL COMPONENTS WITH HIGHEST VARIABILITY AND THE FEATURES CONTRIBUTING TO THIS PRINCIPAL COMPONENT

| Principal Components | Variability | Parameters contributing Highest to the Principal Component |                              |                                      |                                       |
|----------------------|-------------|--|------------------------------|--------------------------------------|---------------------------------------|
| PC1                  | 0.167       | Shopping facilities  | Efficiency of check-in staff | Check in wait time                   | Courtesy of check-in staff            |
|                      |             | -0.257   | -0.2522                      | -0.2586                              | -0.2733                               |
| PC2                  | 0.1088      | Waiting time at passport inspection                        | Courtesy of inspection staff | Shopping facilities                  | Shopping facilities (value for money) |
|                      |             | -0.284   | -0.2783                      | 0.31                                 | 0.295                                 |
| PC3                  | 0.086       | Arrival's passport and visa inspection                     | Speed of Baggage Delivery    | Customs inspection                   | Date recorded                         |
|                      |             | -0.696   | 0.4150                       | -0.458                               | 0.2587                                |
| PC4                  | 0.075       | Ground transportation to/from airport                      | Parking Facilities           | Parking Facilities (value for Money) | Cleanliness of Washrooms              |
|                      |             | 0.49   | 0.534                        | 0.489                                | -0.140                                |
| PC5                  | 0.061       | Wait time at passport inspection                           | Courtesy of inspection staff | Efficiency of check-in staff         | Check-in wait time                    |
|                      |             | 0.5322   | 0.5067                       | -0.3232                              | -0.3313                               |

### 3.4 VARIABLES HIGHLY CORRELATED TO CUSTOMER SATISFACTION

The variables that are highly correlated to customer satisfaction are shown in **Figure 26**



**FIGURE 26 FEATURES HIGHLY CORRELATED TO CUSTOMER SATISFACTION**

The Correlation of following parameters with overall satisfaction of the customer can be explained with following arguments:

**Arrival passport and visa inspection:** This parameter has high **negative correlation** with overall satisfaction which indicates the fact the **customers are not happy** when they are checked for their passport on arrival. This is because **inspection** of passport and other travel document **takes time** and when a person is going to a foreign country, they are excited to move out as soon as possible.

**Speed of baggage delivery:** This parameter has high positive correlation with overall satisfaction that means customers are happy when their baggage gets delivered quickly as it saves waiting time and anticipation which frustrates the customer. To support this fact **Los Angeles International airport** study mentioned in domain analysis can be referred. [5]

**Custom Inspection:** It has a high negative correlation indicative of the fact it makes customers unhappy. It obvious that when someone lands, they want to move out as soon as possible and spending time in customs inspection increases frustration level of the customer.

**Ease of Making connections:** Ease of making connections refers to the ease with which one can transfer to connecting flight. Although transferring from one flight to another is always stressful but if it can be done hassle free it can impart some happiness to customers. The correlation coefficient is low because it makes the customers happy but not to that extent as changing flight will always remain a stressful process.

**Ambience of Airport:** A good ambience always makes a customer happy that is indicated by positive correlation coefficient with overall customer satisfaction.

## 4.1 DECISION OF MACHINE LEARNING MODEL

This section explains the choice of machine learning model to be applied to the problem. The methodology followed for the selection of model is shown in **Figure 27**

### 4.1.1 LINEAR REGRESSION

Linear regression endeavours to model a relationship between target variables and independent variables by fitting a linear equation to observed data.

Independent variables are considered as exploratory variables and target variable is the treated as dependent variable. Linear Regression tries to find out statistical relationship instead of deterministic relationship. The idea is to obtain a line that best fits the data . The best fit line can be defined as the line with least total prediction error. Error is the distance between the datapoints and regression line. Linear regression uses linear predictor function to assign a set of coefficients to independent variable.

Linear Regression is the first choice , the reasons behind choosing linear regression are :

- Linear regression is very versatile, and it uses static measurements to ascertain the variability in data, explained by our model. Moreover, it also helps us in pinpointing selective features from a large set of features that holds better predictability towards target variables.
- Linear regression is one of the most transparent and simple algorithms. We can easily figure out what's happening with our data and how our model is working.
- When using Linear Regression, it's easier to evaluate our model using  $R^2$  , MSE and RMSE values.
- Linear regression gives output in form of a continuous value.
- Linear regression requires normalized parameters and we have already normalized our parameters for using with PCA

Contrary to these advantages there are some disadvantages as well :

- Linear regression is very sensitive to missing values. However, the given dataset has very smaller number of missing value and that too have been dealt with while data cleaning.
- Linear regression is very sensitive to outliers. Hence normalization is required which as already discussed has already been done while applying PCA.

After careful assessment of pros and cons of linear regression it appears to be a perfect candidate for model building for given dataset when PCA is also being applied.

### 4.1.2 RANDOM FOREST

Random Forest is an ensemble learning method that is used for a variety of problems such as regression, classification, and other problems. The mechanism behind the working of random forest classifier is that it constructs large number of decision trees at the time of training. When we are using it for classification problem, output is the class selected by most trees. However, for Regression task output is the average prediction of individual tree.

The decision to select Random Forest is based on the fact that:

- It can be used for both regression and classification
- It provides higher accuracy through cross validation

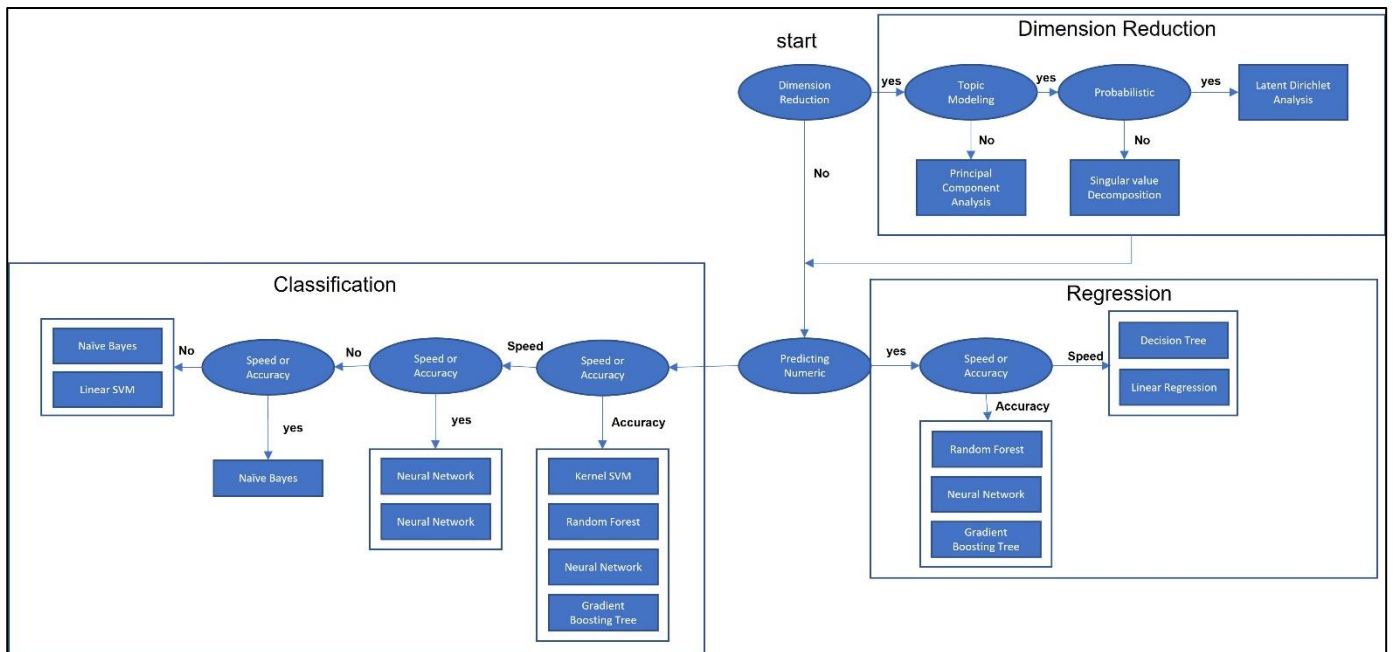


FIGURE 27 METHODOLOGY FOR MODEL SELECTION

- It can handle missing values and still maintain accuracy
- It can handle large dataset with higher dimensionality

Thus because of these advantages we have selected to use Random Forest Classifiers for our Machine learning model.

#### 4.1.3 SELECTION OF HYPER PARAMETERS

Hyper parameters are set of parameters that decides the algorithm's behavior on data. The learning process of the algorithm is impacted by these parameters, and they also affect the final prediction of our model. In order to obtain accurate results proper tuning of these parameter is required. In order to tune these parameters, we use learning curves. Different algorithms possess different set of hyperparameters.

**Linear Regression:** Linear regression doesn't possess any hyperparameters to fine tune the learning process.

**Random Forest:** In case of Random Forest **Number of trees** is the hyper parameter that controls the complexity of random forest. On tinkering around with other parameters, the best result was obtained with the value of parameters as shown in the **Figure 28**.

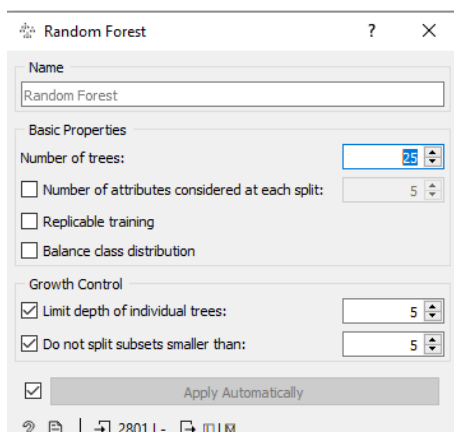


FIGURE 28 RANDOM FOREST WIDGET USED FOR SELECTION OF PARAMETERS

#### 4.2 CHOICE OF SOFTWARE

A combination of python and Orange was use for the successful completion of the task .

- Data cleaning and Feature Engineering was conducted using python as discussed in **section 1.1.2,2.1.1, 2.1.4**.
- Data pre-processing and some of part Data Cleaning was conducted using orange pipelines as discussed in previous sections.
- Data obtained after cleaning and pre-processing has been used in orange pipeline to further conduct PCA on Data .
- Finally, Data obtained after PCA was saved in a file using orange and that file was used to generate learning curves using python .
- Model Building and prediction was done in Orange.

#### 5. APPLICATION OF CROSS VALIDATION TECHNIQUES IN MACHINE LEARNING PIPELINE

When we build a Machine Learning pipeline, we can't just fit the model to the available data and then expect it to work fine with the real-world data. We need to test it before applying it in real life situation. We must ensure that our model gives correct output in every situation. For this we use validation techniques. To avoid trends like overfitting or underfitting shown by our model we use parameters such as R2 and RSME to evaluate our model.

Cross validation techniques are useful in evaluating the accuracy of our model and to check its ability to predict the unseen data correctly. Data is split into Training and Validation set during the implementation of cross validation technique. Then the training set is fed to the model and the predicted values are used to calculate the error with actual values.

To achieve this on orange we use 'Data Sampler' widget to split the data into desired ratio whereas in order to validate

the data we use 'Prediction' widget or 'Test and score' widget. Test and score widget applies the selected validation technique to the machine learning algorithm of your choice and calculates the result. It can also calculate the performance metrics of many models at once and also compare the results.

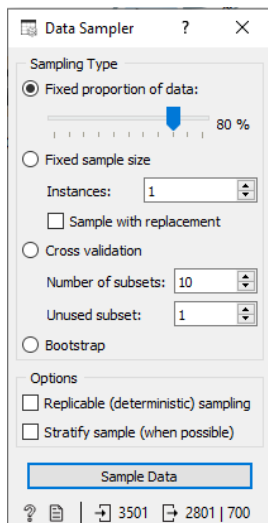


FIGURE 29 DATA SAMPLER WIDGET USED TO SAMPLING DATA

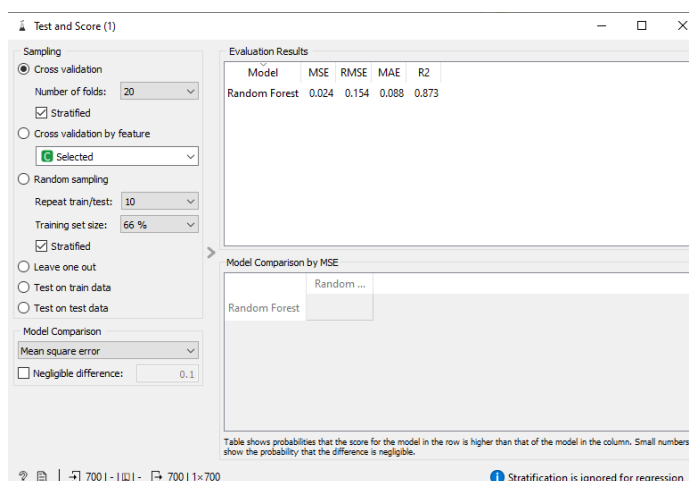


FIGURE 30 TEST SCORE WIDGET USED FOR APPLYING DIFFERENT VALIDATION TECHNIQUES ON DATA

## 6. EVALUATION OF MACHINE LEARNING MODEL

To predict the accuracy of Machine Learning Model Various Evaluation metrics are used. Most used Evaluation matrices are **RMSE ( Root Mean Square Error )** and **R<sup>2</sup> ( Coefficient of Determination )** . These metrics are discussed in detail in subsequent sections .

### 6.1 RMSE ( ROOT MEAN SQUARE ERROR)

The formula used for computation of RMSE is :

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}}$$

RMSE is computed by square root of average of square of errors. The greater the error will be more will be its impact on RSME value since it is a squared error. RSME indicates the standard deviation of the errors. In simple language it indicates the spread around the line of best fit.

RSME value under 0.5 is Ok however a value in range of 0.1 to 0.2 is preferred.

### 6.2 R<sup>2</sup> ( COEFFICIENT OF DETERMINATION )

R<sup>2</sup> tells us about the proportion of variance in the dependent variable that are being explained by independent parameters. It indicates the extent of variance explained by one variable for another variable. To calculate R<sup>2</sup> following relation is used :

$$R^2 = 1 - \frac{\text{Sum of squares of residuals}}{\text{Total sum of squares}}$$

Normally R<sup>2</sup> value in a range of 0.75-1.0 is preferred .

### 6.3 LEARNING CURVES

A learning curve indicates errors of machine learning model , plotted against the number of elements in training set. Learning curves are used to analyse the ability of model to generalize for unseen data. The performance of model can be evaluated from the characteristics of the curve.

Our objective is to minimize the error in our machine learning model . The main source of errors is bias and variance. To build accurate model, we need to minimize these two. But both holds an inverse relationship. So, we need to optimise our model to get minimum possible values of these two . The learning curve obtained for our models are shown below :

Learning curves for a Linear regression model

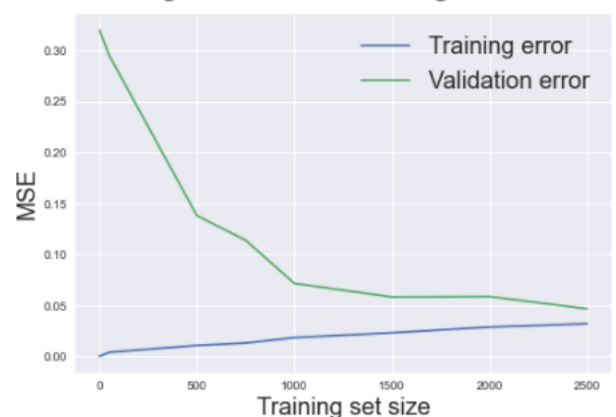


FIGURE 31 LEARNING CURVE FOR LINEAR REGRESSION

Learning curves for a Random Forest regression model

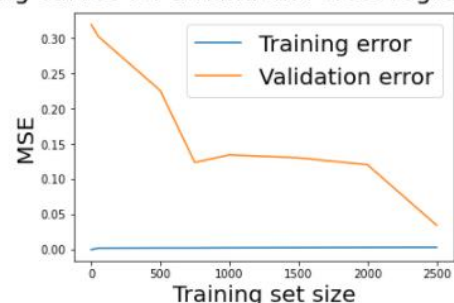


FIGURE 32 TRAINING CURVE FOR RANDOM FOREST REGRESSOR

The interpretations that can be made from the above models are :

- When the training set size is small it is observable that MSE values for both Linear Regression and Random Forest training set are 0 this indicates the model has no problem fitting this data point. However, the validation dataset shows a very high value of MSE because it's most unlikely that our trained model can fit perfectly on a single data point.
- As we increase the training data size the model's error starts increasing for Linear regression but remains constant for Random Forest while the validation error decreases for the algorithm in a similar manner. After a certain set of values, the validation error for Random Forest becomes constant and then starts decreasing again. As we increase the training data set size our model doesn't predict all the training points perfectly however in case of validation dataset the performance of model improves drastically.
- At the end both the curves are still converging towards each other for both the models. This has been explained in subsequent sections.

When the model fits well to almost all the data sets without much change that means it has less variance and more bias. We have tried to oversimplify the model. Such a model is called an underfitting model. However, if the model has less bias it tends to change more quickly for each data set resulting in more variance. Such a model tends to overfit. This is where the role of learning curves comes in. We need to maintain a delicate balance between variance and bias so that our models neither overfit nor underfit.

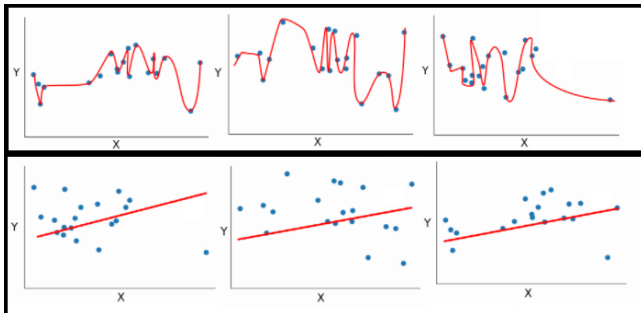


FIGURE 33 TOP ROW SHOWS LOW BIAS DATA (OVERFITTING) WHEREAS BOTTOM ROW SHOWS HIGH BIAS DATA (UNDERFITTING)

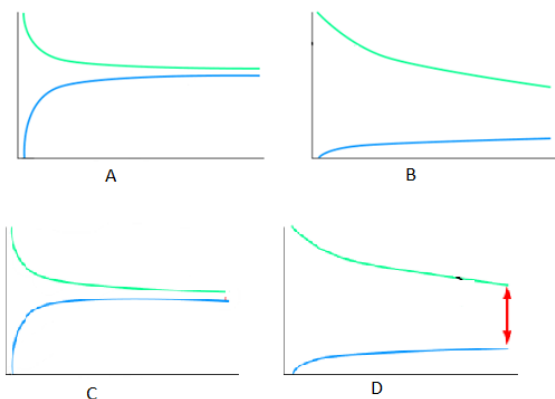


FIGURE 34 (A) HIGH BIAS (B) LOW BIAS (C) LESS VARIANCE (D) HIGH VARIANCE

So, from the above figures we can easily interpret that if the bias is low and variance is high then the model is overfitting on the contrary if the bias is high, but variance is low then the model is underfitting.

## 7. DISCUSSION

In this section we will compare our results with some different algorithms, discuss mathematical peculiarities, strength, and weakness of chosen algorithms with the algorithms not taught in class.

### 7.1 COMPARISON WITH TWO ALGORITHMS NOT DISCUSSED IN CLASS

#### 7.1.1 ADABOOST:

To understand Adaboost we need to understand Boosting. In boosting a model is made from training data and then a second model that tries to remove errors from the first model. In this way models are added till a perfect model is achieved. Adaboost is used to enhance the performance of Decision trees and binary classification problems. Adaboost works best with weak learners hence we use it with decision trees with just one level. Such short trees can only produce just one decision and are called decision stumps.

These stumps are left to make decision and their miscalculation rate is then fed to trained model.

Assigning Sample Weight

$$\text{sample weight} = \frac{1}{\# \text{ of samples}}$$

Now, calculating Gini Impurity (GI)

$$GI = (\text{the probability of True})^2 - (\text{the probability of False})^2$$

Total Gini Impurity = Weighted average of two individual impurity.

Lastly, amount of say for the created stump

$$\frac{1}{2} \log \left( \frac{1 - \text{total error}}{\text{total error}} \right)$$

#### Advantages of Adaboost:

- Can be implemented very quickly
- Can be used for tasks like image recognition, text recognition, classification etc.
- It can be combined with any machine learning algorithm.

#### Disadvantages :

- Sensible to noise in data.



### 7.1.2 SVM

SVM (Support Vector Machine) is a supervised learning algorithm which is capable of being used for both classification as well as Regression. SVM attempts to generate a multidimensional hyperplane to divide different sets of data. To predict values, SVM recognizes the class best suitable for provided dataset and makes the predictions accordingly. SVM is commonly used for classification problem but is also capable of solving Regression Problems. SVM is highly effective in dealing with datasets having large dimensions, specifically in cases where sample size is less than the number of features. Although very effective, SVM is prone to error in presence of noisy dataset having outliers. Hyperparameters for SVM are kernel, gamma, and cost C. This is most commonly used in image processing and spam filtering.

#### Advantages of SVM

- SVM operates phenomenally with datasets having clear separation of classes
- Effective for high dimensional spaces
- Effective for cases where the number of dimensions is more than number of samples
- Relatively memory efficient

#### Disadvantages of SVM

- Not suitable for large datasets
- Performs poorly with noisy and overlapping data
- There is no probabilistic explanation for the classification.

### 7.2 MATHEMATICAL PECULIARITIES OF LINEAR REGRESSION AND RANDOM FOREST CLASSIFIERS

#### For Linear Regression:

$$h(x) = \theta_0 + \theta_1 x$$

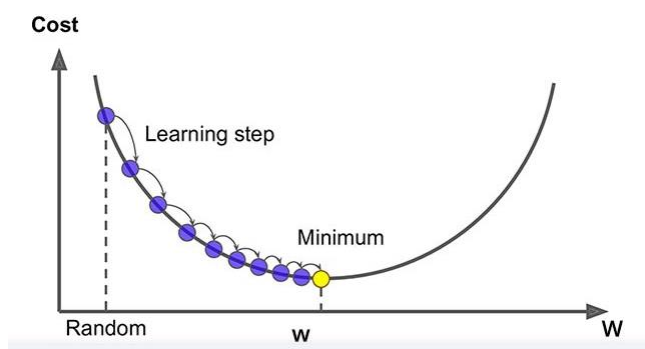
where  $h(x)$  is hypothesis function

Squared error used to determine the fit of hypothesis is given by value

$$\text{Squared error fn} = \frac{1}{2m} \sum_{i=1}^{i=m} (h_0(x^i) - y^i)^2$$

Cost function is given by

$$C(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{i=m} (h_0(x^i) - y^i)^2$$



For Linear regression if learning step is too large the cost function will never reach the bottom of the curve it will keep bouncing from one place to another.

### 7.3 APPLY APPROPRIATE METRICS TO COMPARE THE ALGORITHM YOU HAVE CHOSEN WITH ONES DISCUSSED IN CLASS

A comparison between kNN, Random Forest, Linear Regression, Gradient Boosting and Adaboost has been shown in **Figure 35**. It is apparent from the table that all four except kNN has RSME below 0.2 and R2 value above 0.8. Similar trend was obtained during cross-validation which indicates that the models were stable and generalize to unseen data. Gradient Boosting showed the best accuracy followed by Adaboost then Random Forest and then Linear Regression. However all the model were giving an MSE value in the range 0.032-0.036, RSME value in the range 0.179 to 0.190 kNN was ruled out as it gave the worst results apart from all other algorithms. Random Forest, Gradient Boosting and AdaBoost tends to give almost similar results because they all are classification algorithms.

| Evaluation Results |       |       |       |       |
|--------------------|-------|-------|-------|-------|
| Model              | MSE   | RMSE  | MAE   | R2    |
| kNN                | 0.053 | 0.230 | 0.149 | 0.713 |
| Random Forest      | 0.035 | 0.188 | 0.111 | 0.809 |
| Linear Regression  | 0.036 | 0.190 | 0.133 | 0.804 |
| Gradient Boosting  | 0.032 | 0.179 | 0.104 | 0.828 |
| AdaBoost           | 0.035 | 0.186 | 0.076 | 0.812 |

FIGURE 35 COMPARISON BETWEEN DIFFERENT ALGORITHMS

## REFERENCES

- [1] CFI ( Corporate Finance Institute) , “Domain Knowledge ( Data Science ) - Overview , Subject Areas, Case Study,” 2015. [Online]. Available: [corporatefinanceinstitute.com/resources/knowledge/data-analysis/domain-knowledge-data-science/](https://corporatefinanceinstitute.com/resources/knowledge/data-analysis/domain-knowledge-data-science/).
- [2] ACI Insights, “Defining customer experience: How airports can own the passenger,” ACI, [Online]. Available: <https://blog.aci.aero/defining-customer-experience-how-airports-can-own-the-passenger-journey/>. [Accessed 13 December 2021].
- [3] K. Doll, “What is Peak-End Theory? A Psychologist Explains How our Memory fools Us,” PositivePsychology.com, March 2019. [Online]. Available: [positivepsychology.com/what-is-peak-end-theory/](https://positivepsychology.com/what-is-peak-end-theory/). [Accessed 13 December 2021].
- [4] R. Heuser, “How to understand you travel customers and their needs,” the-future-of-commerce.com, [Online]. Available: <https://www.the-future-of-commerce.com/2018/11/08/travel-customers/>. [Accessed 13 December 2021].
- [5] PhocusWire, “Reimagining customer experience part3: Airport innovations,” phocuswire.com, [Online]. Available: <https://www.phocuswire.com/Customer-experience-part-3-airports>. [Accessed 13 December 2021].
- [6] P. Patil, “What is Exploratory Data Analysis,” towardsdatascience.com, [Online]. Available: <https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15>. [Accessed 15 December 2021].
- [7] A. Bhandari, “Feature Scaling , Standardization Vs Normalization,” analyticsvidhya.com, 3 April 2020. [Online]. Available: [analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/](https://analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/). [Accessed 16 December 2021].
- [8] Z. Jaadi, “A-step-by-step Explanation of Principal Component Analysis (PCA),” builtin.com, [Online]. Available: [builtin.com/data-science/step-step-explanation-principal-component-analysis](https://builtin.com/data-science/step-step-explanation-principal-component-analysis). [Accessed 17 December 2021].
- [9] Analytics Vidhya, “Understanding Principle component Analysis step by step,” Analytics Vidhya, [Online]. Available: [medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9](https://medium.com/analytics-vidhya/understanding-principle-component-analysis-pca-step-by-step-e7a4bb4031d9). [Accessed 17 December 2021].