

**ASSIGNMENT 3**

**CT045-3-M-ABAV**

**ADVANCED BUSINESS ANALYTICS AND VISUALIZATION**

**HAND IN DATE: *02 December 2022***

**LECTURER NAME: DR. PREETHI SUBRAMANIAN**

**STUDENT NAME: MOHAMED ZAHID**

**TP NUMBER: TP067803**

**PART C**

**PREDICTIVE ANALYSIS FOR CUSTOMER SATISFACTION IN AIRLINE INDUSTRY: A CASE STUDY**

**TABLE OF CONTENTS**

LIST OF FIGURES.........................................................................................................................1

INTRODUCTION...........................................................................................................................4

BUSINESS GOALS........................................................................................................................4

AIM..................................................................................................................................................4

OBJECTIVES..................................................................................................................................5

SCOPE.............................................................................................................................................5

METHODOLOGY..........................................................................................................................5

DATA VISUALIZATION...............................................................................................................7

EXPLORATORY DATA ANALYSIS (EDA) AND DATA PREPARATION...........................10

MODEL CONSTRUCTION AND INTERPRETATION OF MODEL RESULTS.....................19

MODEL COMPARISON..............................................................................................................31

CRITICAL INTERPRETATIONS OF RESULTS.......................................................................32

DISCUSSIONS..............................................................................................................................37

CONCLUSIONS/PERSONAL REFLECTION............................................................................38

**LIST OF FIGURES**

Figure 1 – Cross Industry Standard Process for Data Mining (CRISP-DM) ..................................5

Figure 2 – Objectives of Data Visualization....................................................................................7

Figure 3 – Visualization of Objective Variables (1) .......................................................................7

Figure 4 – Visualization of Objective Variables (2) .......................................................................8

Figure 5 – Visualization of Subjective Variables (1) ......................................................................8

Figure 6 – Visualization of Subjective Variables (2) ......................................................................9

Figure 7 – Visualization of Subjective Variables (3) ......................................................................9

Figure 8 – Findings and Recommendations through Visualizations.............................................10

Figure 9 – SAS Process Flow........................................................................................................12

Figure 10 – Class Variable Summary Statistics.............................................................................13

Figure 11 – Satisfaction Frequencies.............................................................................................13

Figure 12 – Interval Variable Summary Statistics.........................................................................14

Figure 13 – Variable Worth...........................................................................................................14

Figure 14 – Data Preparation Highlighted.....................................................................................15

Figure 15 – Data Drop (Gender)....................................................................................................16

Figure 16 – Data Imputation..........................................................................................................16

Figure 17 – Data Replacement.......................................................................................................17

Figure 18 – Log Transformation....................................................................................................18

Figure 19 – Log Transformation Output........................................................................................18

Figure 20 – Data Partition..............................................................................................................19

Figure 21 – Tree Models................................................................................................................20

Figure 22 – Tree Nodes..................................................................................................................20

Figure 23 – Subtree Assessment Plot.............................................................................................21

Figure 24 – Event Classification Table (DT).................................................................................21

Figure 25 – Node Rules (DT)........................................................................................................22

Figure 26 – Event Classification Table (HP Tree) ........................................................................22

Figure 27 – Node Rules (HP Tree) ...............................................................................................23

Figure 28 – Iteration Plot (HP Forest) ..........................................................................................23

Figure 29 – Event Classification Table (HP Forest) .....................................................................24

Figure 30 – Iteration History (HP Forest) .....................................................................................24

Figure 31 – Regression Models.....................................................................................................25

Figure 32 – Regression Node Selection.........................................................................................25

Figure 33 – Summary of Forward Selection (Forward Regression) .............................................26

Figure 34 – Maximum Likelihood Estimation (Forward Regression) ..........................................26

Figure 35 – Event Classification Table (Forward Regression) .....................................................27

Figure 36 – Polynomial 2nd Degree Forward Regression..............................................................27

Figure 37 – Summary of Forward Selection (Forward 2nd Regression) .......................................27

Figure 38 – Event Classification Table (Forward 2nd Regression) ...............................................28

Figure 39 – Summary of Selection (HP Regression) ....................................................................28

Figure 40 – Event Classification Table (HP Regression) .............................................................29

Figure 41 – Neural Net Models.....................................................................................................29

Figure 42 – Neural Net Node Settings...........................................................................................30

Figure 43 – Comparison of Neural Nets........................................................................................31

Figure 44 – Event Classification Table (Neural Nets) ..................................................................31

Figure 45 – Fit Statistics of Model Comparison............................................................................32

Figure 46 – Event Classification Table (All Models) ...................................................................32

Figure 28 (repeat) – Iteration Plot (HP Forest) .............................................................................33

Figure 47 – Fit Statistics of HP Forest...........................................................................................33

Figure 48 – Model Information of HP Forest................................................................................33

Figure 29 (repeat) – Event Classification Table (HP Forest) .......................................................34

Figure 49 – Link Graph..................................................................................................................35

Figure 50 – Fit Statistics (HP 1 Layer Neural Net) ......................................................................35

Figure 51 – Surrogate Model (HP 1 Layer Neural Net) ...............................................................35

Figure 52 – Surrogate Model Node Rule (1) ................................................................................36

Figure 53 – Surrogate Model Node Rule (2) ................................................................................36

Figure 54 – Variable Importance of HP Forest..............................................................................37

Figure 55 – Surrogate Model Node Rule (3) ................................................................................38

**1. INTRODUCTION**

Airline businesses represent any organization or company that provides a service for air transportation. Possibly, an oligopoly exists now in the airline business. When a small number of companies dominate a market, an oligopoly results, typically because of high entry barriers that prevent other companies from joining. Oligopolists might be allowed to manipulate customer decision-making given the absence of competition.

However, the whole airline sector was impacted when the COVID-19 issue disrupted air travel. Due to changes in consumer behavior brought on by the COVID-19 problem, travel limitations, and the resulting economic crisis, demand for aviation services has declined significantly. This forced the oligopolists to adapt and focus on a more customer satisfactory model for business growth and retention.

This case study focuses on such a business case wherein, data driven decisions can be implemented to better customer satisfaction in the airline industry. This paper consists of business goals, aim and objectives, scope, and methodology to achieve this.

**2. BUSINESS GOALS**

Customers who are not happy or who are not engaged result in fewer passengers and decreased revenue. A consumer should always have a good experience when they travel. On-time flights, good in-flight entertainment, more (and better) refreshments, and more leg room are some of the apparent elements that may enhance a positive experience and build loyalty.

An airline company's overall objectives for customer satisfaction may be separated into two categories: To draw both leisure and business travelers in order to boost recurring business, and to make sure that customers always receive the best possible service from the airline. Factors affecting customer satisfaction can include baggage theft, delayed flights, uncomfortable seats, hidden charges, poor customer service and a many more such issues.

To combat this, many airlines turn to the use of surveys. To ascertain how typical each customer's ideas and experiences are, surveys are a useful analytical tool. When done correctly, surveys provide accurate information on the attitudes and behaviors of consumers that may be used to guide important business decisions.

**3. AIM**

The main aim of this project is to develop a predictive model for customer satisfaction that can be used for data driven decision making by an airline.

**4. OBJECTIVES**

The specific objectives of this project are:

* To visualize airline passenger satisfaction data and identify patterns how overall satisfaction is distributed.
* To perform exploratory data analysis (EDA) on airline passenger satisfaction data to assess the viability of a predictive model as an analytical tool for business difficulties.
* To build predictive models for customer satisfaction in the airline industry.

**5. SCOPE**

The scope of this project is to understand satisfaction of customers that decide to use airline services using survey data that depicts past experiences of customers. This project uses a publicly available dataset found on Kaggle at the link: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>. Passengers' flying experiences and levels of satisfaction will be utilized as the dataset for this project. Although the dataset already contains a test dataset, the train dataset will receive the majority of attention. It should be noticed that the test dataset makes up about 20% of the entire dataset and that the train dataset and test dataset share the same columns (test and train dataset).

The training dataset has 103904 records, the bulk of which do not appear to contain any null values. The dataset also contains around 24 variables, including the satisfaction variable, which will be crucial for this project's later phases.

The project deliverables will include a dashboard for understanding the data and predictive models for customer satisfaction.

**6. METHODOLOGY**

This project is built around the process paradigm known as the CRoss Industry Standard Process for Data Mining (CRISP-DM). It entails six successive steps.

Diagram

Description automatically generated

Figure 1 – The CRoss Industry Standard Process for Data Mining (CRISP-DM)

**STEP 1 – Business Understanding**

Understanding the project's goals and needs is the main emphasis of the business understanding step. As mentioned previously this project focuses on the satisfaction of customers in the airline industry and how to develop a model that can be used to predict the same. The deliverables of this project would help an airline business to make better decision based on data to further its profitability by better serving its customers.

**STEP 2 – Data Understanding**

This stage focuses on locating, obtaining, and analysing data sets that may help in accomplishing project goals in addition to improving the foundation of business understanding. As previously stated, this project will use a public dataset that consists of 103904 record. Tebleau will be used for exploring and understanding the patterns present in the dataset. This project will attempt to present different types of analysis for better data understanding.

**STEP 3 – Data Preparation**

This steps prepares the data for final model building. The data considered for this project is large in size. But fortunately, this dataset is mostly clean. This step will include actions such as outlier/skewness treatment and imputation of missing data. Variables with low worth will also be dropped during this step. The link for the publicly available datset already consists of train-test split for model building and therfore, data splitting will not incorporated for this project however, the data will be split for training and validation of the machine learning models.

**STEP 4 – Modeling**

This stage focuses on creating multiple models using a variety of modeling methodologies. This project will attempt to create a few different models that can be used for predicting customer satisfaction, namely, Decesion Tree, Regression, HP models (Tree and Regression) and different Neural Nets for comparison. These models will be created using the SAS Enterprise Miner. The models will then be assesed to choose the best two models which will be explained extensively.

**STEP 5 – Evaluation**

The assessment step examines which model best matches the business and what to do next in more detail. Questions such as – Does the model meet the business success criteria? Were all steps carried out correctly? Will be answered summarized to make any necessary corrections. This step determines whether to proceed toward deployment, iterate further, or initiate revised objectives.

**STEP 6 – Deployment**

The project objectives will be complete at this stage and the deliverables can be given to an organization in the airline industry for detrmining their marketing strategies and implementing the findings of this project in a real world scenario. This step is beyond the scope of this project.

**7.** **DATA VISUALIZATION**

Data visualization of this dataset has been coverd in the preceeding presentation. The visualizations that were done are taken from the tableau workbook and shown below.

Graphical user interface, text, application

Description automatically generated

Figure 2 – Objectives of Data Visualization

Chart, bar chart, histogram

Description automatically generated

Figure 3 – Visualization of Objective Variables (1)

Chart, bar chart

Description automatically generated

Figure 4 – Visualization of Objective Variables (2)

Chart, bar chart

Description automatically generated

Figure 5 – Visualization of Subjective Variables (1)

Chart

Description automatically generated

Figure 6 – Visualization of Subjective Variables (2)

Graphical user interface

Description automatically generated

Figure 7 – Visualization of Subjective Variables (3)

By analysing the abouve graphs and by using filters accompanying them, certain findings can be noticed and recommendations to the airline company are to be made based on them.

Graphical user interface, application, Word

Description automatically generated

Figure 8 – Findings and Recommendations through Visualizations

**8. EXPLORATORY DATA ANALYSIS (EDA) AND DATA PREPARATION**

**8.1. INTRODUCTION OF THE DATASET**

An airline customer satisfaction survey is the dataset that was selected for this investigation. There are 103904 entries in the dataset, the majority of which don't seem to have any null values. Around 24 other factors, including the satisfaction measure, are also included in the dataset. Below is a table with a summary of the metadata. **The variable name, description, data type, and example data are all listed in the table. 4 nominal, 17 ordinal, 2 interval, and 1 ratio variable are included in the dataset's total of 24 variables.**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Description | Data Type | Sample Data |
| Id | Customer ID | Nominal | 70172, 5047.... |
| Gender | Gender of the passengers (Female, Male) | Nominal | Male, Female |
| Customer Type | The customer type (Loyal customer, disloyal customer) | Ordinal | Loyal Customer, disloyal Customer |
| Age | The actual age of the passengers | Nominal | 13, 25.... |
| Type of Travel | Purpose of the flight of the passengers (Personal Travel, Business Travel) | Nominal | Personal Travel, Business travel |
| Class | Travel class in the plane of the passengers (Business, Eco, Eco Plus) | Ordinal | Eco , Eco Plus, Business |
| Flight Distance | The flight distance of this journey | Ratio | 460, 235.... |
| Inflight wifi service | Satisfaction level of the inflight wifi service (0:Not Applicable;1-5) | Ordinal | 1,2,3,4,5 |
| Departure/Arrival time convenient | Satisfaction level of Departure/Arrival time convenient | Ordinal | 1,2,3,4,5 |
| Ease of Online booking | Satisfaction level of online booking | Ordinal | 1,2,3,4,5 |
| Gate location | Satisfaction level of Gate location | Ordinal | 1,2,3,4,5 |
| Food and drink | Satisfaction level of Food and drink | Ordinal | 1,2,3,4,5 |
| Online boarding | Satisfaction level of online boarding | Ordinal | 1,2,3,4,5 |
| Seat comfort | Satisfaction level of Seat comfort | Ordinal | 1,2,3,4,5 |
| Inflight entertainment | Satisfaction level of inflight entertainment | Ordinal | 1,2,3,4,5 |
| On-board service | Satisfaction level of On-board service | Ordinal | 1,2,3,4,5 |
| Leg room service | Satisfaction level of Leg room service | Ordinal | 1,2,3,4,5 |
| Baggage handling | Satisfaction level of baggage handling | Ordinal | 1,2,3,4,5 |
| Checkin service | Satisfaction level of Check-in service | Ordinal | 1,2,3,4,5 |
| Inflight service | Satisfaction level of inflight service | Ordinal | 1,2,3,4,5 |
| Cleanliness | Satisfaction level of Cleanliness | Ordinal | 1,2,3,4,5 |
| Departure Delay in Minutes | Minutes delayed when departure | Interval | 25, 1..... |
| Arrival Delay in Minutes | Minutes delayed when Arrival | Interval | 18, 6...... |
| Satisfaction | Airline satisfaction level(Satisfaction, neutral or dissatisfaction) | Ordinal | neutral or dissatisfied, satisfied |

**8.2. PROCESS FLOW**

Figure – 9 shows the total process flow for the exploration, pre-processing, and modelling processes.

Diagram

Description automatically generated

Figure 9 – SAS Process Flow

**The highlighted areas show the importing of the dataset into SAS enterprise miner, EDA and the pre-processing of the data.**

**8.3. STATISTICAL EXPLORATION**

A dataset may be better understood through exploration, which also makes it simpler to explore and utilise the data in the future. Class variable statistics for the dataset of this study is shown in figure 10. This shows the number of levels for each variable.

A picture containing chart

Description automatically generated

Figure 10 – Class Variable Summary Statistics

**It is important to note that many variables show levels of 6 when the customer survey should have only 5 posiible levels, i.e., 1 to 5. Additionally, the output variable – satisfaction, has two levels of 56.67% and 43.34% mode percentage and as the dataset is sufficiently large, there is no need for class balancing to take place** (also shown graphically in figure 11).

Chart, bar chart

Description automatically generated

Figure 11 – Satisfaction Frequencies

**Exploring the numeric values in the dataset, it is noticed that there exists 300 missing values in Arrival\_Delay\_in\_Minutes. It is also noticed that Arrival\_Delay\_in\_Minutes and Departure\_Delay\_in\_Minutes have very high skewness.**

Calendar

Description automatically generated

Figure 12 – Interval Variable Summary Statistics

Another important concept that is shown by exploring data in SAS enterprise miner is the variable worth. Variable worth of the independent variables is shown with regards to predicting satisfaction. **Gender shows the least amount of worth and is therefore discarded in the pre-proccing step.** Variable worth histogram is shown in figure 13.

Chart, bar chart, histogram

Description automatically generated

Figure 13 – Variable Worth

**8.4. PREPARATION OF THE DATA**

The crucial stage in addressing missing critical values, consistency issues, noise, mistakes, and outliers is data preparation. These data flaws would persist in data science without data preparation, which would reduce the quality of the final model created.

Diagram

Description automatically generated

Figure 14 – Data Preparation Highlighted

From the exploration of the dataset, it is concurred that there are four main steps to be conducted for the preparation of this dataset.

* Drop the variable – Gender.
* Impute the missing values in Arrival\_Delay\_in\_Minutes.
* Replace the 0’s in the survey data to 3’s (median of 1,2,3,4 and 5) to reduce the survey data value to levels of 5. By using median, the overal data reliablity is least affected.
* Transform the variables Arrival\_Delay\_in\_Minutes and Departure\_Delay\_in\_Minutes to reduce their skewness.

Apart from the above steps for preparation, data partition must be conducted for the validation of the models to be built.

**8.5. DATA DROP**

Graphical user interface, table

Description automatically generated

Figure 15 – Data Drop (Gender)

Figure 15 shows the use of the drop node in SAS enterprise miner to remove the variable – Gender. This is done because of its low variable worth in predicting satisfaction of customers.

**8.6. DATA IMPUTATION**

Graphical user interface, application, table, Excel

Description automatically generated

Figure 16 – Data Imputation

Figure 16 shows the use of the impute node in SAS enterprise miner to impute the variable – Arrival\_Delay\_in\_Minutes. The method chosen was median imputation to fill in the blank values. It is noticed that the data in this variable is skewed and when the data is skewed, it is optimal to use the median value for replacing the missing values.

**8.7. DATA REPLACEMENT**

Graphical user interface, table

Description automatically generated

Figure 17 – Data Replacement

Figure 17 shows the use of the replacement node in SAS enterprise miner to replace the class variables that have 0’s in the survey data when the only acceptable values are 1,2,3,4 and 5. To keep data reliablity, the value 0’s are replaced by the median – 3.

**8.8. DATA TRANSFORMATION**

Data that has been structured and validated correctly enhances data quality and shields programmes from possible pitfalls like skewness. If skewness exists, the data must be transformed. From EDA , Arrival\_Delay\_in\_Minutes and Departure\_Delay\_in\_Minutes show high skewness. Therefore, using log transformation, the skewness of the variables are reduced to an acceptable range.

Graphical user interface, table

Description automatically generated

Figure 18 – Log Transformation

The prefix IMP\_ refers to the previous imputation of the variable.

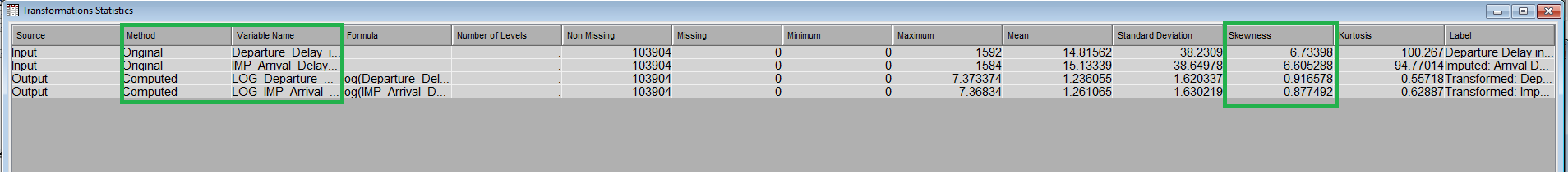


Figure 19 – Log Transformation Output

The output of log transformation shows the skewness to drop below 1, which is accepted to build a machine learning model.

**8.9. DATA PARTITION**

Data partition is the node in SAS entrprise miner that is used to separate the dataset based on user defined percentages into train, validation and test datasets. For the purpose of this study, there will be no use of test data. However, 30% of the data will be used for validation of the model that are to be built. Validation sets are chosen and the final model is tuned using them. The remaining 70% of the data will be used to train the machine learning models.

Graphical user interface, application

Description automatically generated

Figure 20 – Data Partition

**9. MODEL CONSTRUCTION AND INTERPRETATION OF MODEL RESULTS**

To build a reliable model for predicting the customer satisfaction towards the airline, several models are created and trained. These models are compared to find the best one.

The models created are namely –

1. Tree Based Models – Decision tree, HP tree and HP forest.
2. Logistic Regression Models – Forward regression, Polynomial 2nd degree regression and HP regression.
3. Neural Networks – Generalized linear model (GLM), Multilayer perceptron (MLP), HP 1 layer and HP 3 layer.

**9.1. TREE BASED MODELS**

***Decision Tree***

Figure 21 shows the tree based models highlighted in the process flow. The target variable is satisfaction.

Diagram

Description automatically generated

Figure 21 – Tree Models

**The tree is trained based on the log worth of the variables in the dataset. The greater the log worth, the more information given by the variable towards the model.**

Chart, box and whisker chart

Description automatically generated

Figure 22 – Tree Nodes

Looking at the subtree assessment plot, it is visible that at N =15, misclassification rate stabilises. **Therefore, N=15 number of tree nodes is used for model building.**

Chart, histogram

Description automatically generated

Figure 23 – Subtree Assessment Plot

***Interpretation of Outcomes***

**Event classification table is shown in the figure 24. The difference between the expected and actual numbers of successes is shown in the classification table, also known as a confusion matrix.** Similar to that, it contrasts the expected and actual numbers of failures. The table shows the true positive, false positive, true negetive and false negetive values.

Table

Description automatically generated

Figure 24 – Event Classification Table (DT)

Looking at node rules, from the top two node rules that almost predict satisfaction perfectly, it is seen that for business travel, online boarding and inflight entertainment are very importanct factors. For personal travel, online boarding and inflight wifi are the two most important factors.

Text, table

Description automatically generated

Figure 25 – Node Rules (DT)

***HP Tree***

Setting the leaf size as 5 (the smallest number of training observations that a leaf can have), the results of the model is as show.

***Interpretation of Outcomes***

The event classification table is shown in figure 26.

Table

Description automatically generated

Figure 26 – Event Classification Table (HP Tree)

The node rules however, are more complicated due to consideration of high performance. Node 55 shows that for disloyal customers, inflight wifi and other services are important toward their satisfaction.

Text

Description automatically generated

Figure 27 – Node Rules (HP Tree)

***HP Forest***

For the building of HP forest model, maximum number of trees is set as 100 and maximum depth of tree is set as 50.

***Interpretation of Outcomes***

The iteration plot of the model shows that there is no overfitting of the model.

Graphical user interface, chart, application

Description automatically generated

Figure 28 – Iteration Plot (HP Forest)

The event classification table shows really good results as there is a high value for true positive and low value for false positive.

Table

Description automatically generated

Figure 29 – Event Classification Table (HP Forest)

The model is best at number of trees equals 28 as the missclasification rate for the trained model is lowest at 0.0570.

Table

Description automatically generated with medium confidence

Figure 30 – Iteration History (HP Forest)

**9.2. REGRESSION BASED MODELS**

**Stepwise regression techniques like forward regression start with an empty model and gradually add variables.** For this specific case study, logistic regreesion is taken as this is a classification problem. The models are highlighted in the procss flow.

Diagram

Description automatically generated

Figure 31 – Regression Models

***Forward Selection***

In SAS Entermprise Miner, the first step is to set the regression type as logistic regression and link function as logit. The selection model is set as forward and selection criteria is validation misclassification (classification problem).

Table

Description automatically generated

Figure 32 – Regression Node Selection

***Interpretation of Outcomes***

Misclassification rate is the fraction of predictions that were wrong, without distinguishing the positive and negetive predictions. The misclassification rate is shown in figure 33, the summary of forward selection.

Table

Description automatically generated

Figure 33 – Summary of Forward Selection (Forward Regression)

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 20. The summary of forward selection shows the variables added during each step at model building. The final model, at step 20, has a misclassification rate of 0.0843. The analysis of maximum likelihood estimation shows assumed probability distribution.

Table

Description automatically generated

Figure 34 – Maximum Likelihood Estimation (Forward Regression)

The event classification for forward regression is shown in figure 35.

Table

Description automatically generated

Figure 35 – Event Classification Table (Forward Regression)

***Polynomial Second Degree Forward Regression***

The same steps as previous model should be repeated. However, the polynomial term is set to yes and polynomial degree is set to 2.

Table

Description automatically generated

Figure 36 – Polynomial 2nd Degree Forward Regression

***Interpretation of Outcomes***

The summary of forward selection is shown in figure 35.

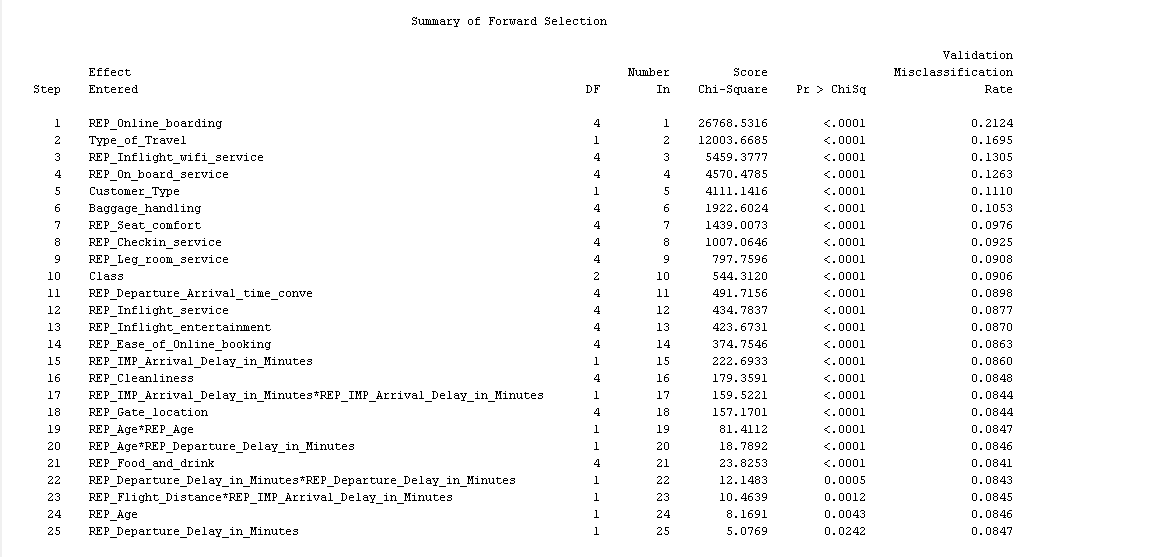


Figure 37 – Summary of Forward Selection (Forward 2nd Regression)

The selected model, based on the misclassification rate for the validation data, is the model trained in Step 21. The misclassification rate is 0.0847.

The event classification table is shown below.

Table

Description automatically generated

Figure 38 – Event Classification Table (Forward 2nd Regression)

***HP Regression***

Set the SAS node to logistic regression as regression type and logit as the link function. Selection method as forward.

***Interpretation of Outcomes***

Table

Description automatically generated

Figure 39 – Summary of Selection (HP Regression)

The model at step 20 is selected and event classification table is shown in figure 40.

Table

Description automatically generated

Figure 40 – Event Classification Table (HP Regression)

**9.3. NEURAL NETWORK MODELS**

**A neural network is a collection of algorithms that aims to identify underlying links in a piece of data using a method that imitates how the human brain functions.** In this context, neural networks are systems of neurons that can be either organic or synthetic in origin.

In SAS Enterprise Miner, a variety of neural net models can be built. The study builds four models shown in figure 41.

Diagram

Description automatically generated

Figure 41 – Neural Net Models

Graphical user interface, text, application

Description automatically generated

Figure 42 – Neural Net Node Settings

**The node architecture has to be selected by the user, the two architectures used in this study are generalized linear model and multilayer perceptron.** The generalized linear model (GLM), which generalizes linear regression, allows the linear model to be connected to the response variable via a link function and allows the variance of each measurement to be a function of the projected value. A form of feedforward artificial neural network that is fully linked is called a multilayer perceptron.

Neural nets in SAS can have multiple hidden layers. Neural networks with one to two hidden layers might function if the data is less complicated and has fewer dimensions or characteristics. Three to five hidden layers can be employed to provide the best result when the data in question has big dimensions or characteristics.

**Model selection criteria for these neural nets is misclassification.**

After the building of these models, a model comparison is done to choose the best neural net among them.

***Neural Net Model Comparison***

Table

Description automatically generated

Figure 43 – Comparison of Neural Nets

Figure 43 shows the comparison of different models built. High Performance 1 Hidden Layer Neural Network shows the least misclassification rate of 0.057325. As a note, the high-performance neural networks perform better than the MLP and GLM models.

The even classification table of the comparisons of the different models is shown.

Table

Description automatically generated

Figure 44 – Event Classification Table (Neural Nets)

**Therefore, out of the 4 neural nets that were built, High Performance1 Hidden Layer Neural Network (HP 1L NN) is the best and is chosen.**

**10. MODEL COMPARISON**

Using HP 1 layer as the best neural network, it is then compared to the tree-based and regression-based models. The best two models out of these will be studied in detail to result in the best model for the prediction of customer satisfaction.

SAS has a model comparison node which shows the different misclassification rates for each model.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 45 – Fit Statistics of Model Comparison

**HP 1-layer neural net and HP forest show the best results out of all the models that were built showing misclassification rates of 0.057325 and 0.058223 respectively.**

Table

Description automatically generated

Figure 46 – Event Classification Table (All Models)

**HP 1L neural nets and HP forest show the highest true postive values and lowest false positive values and are therefore the two best models built for this study.**

**11. CRITICAL INTERPRETATIONS OF RESULTS**

This section of the report details the outputs of the HP 1-Layer Neural Net and HP Forest and compares them to determine the best model. The business interpretations of the model is also discussed.

**11.1. HP FOREST**

A forest of decision trees is created via the high-performance HP forest node in SAS process, which also builds predictive models. An association between input variables and a target variable is defined by a predictive model.

In SAS, user provides the maximum number of trees and the maximum depth of each tree created in the forest. For the purpose of this study, the values given were 100 and 50 respectively. The variable importance method is chosen as loss method because for loss reduction, all variables that have an importance greater than 0 will be set to input. All other variables will be rejected.

Upon running this node, the iteration plot shows that the model built is not overfit.

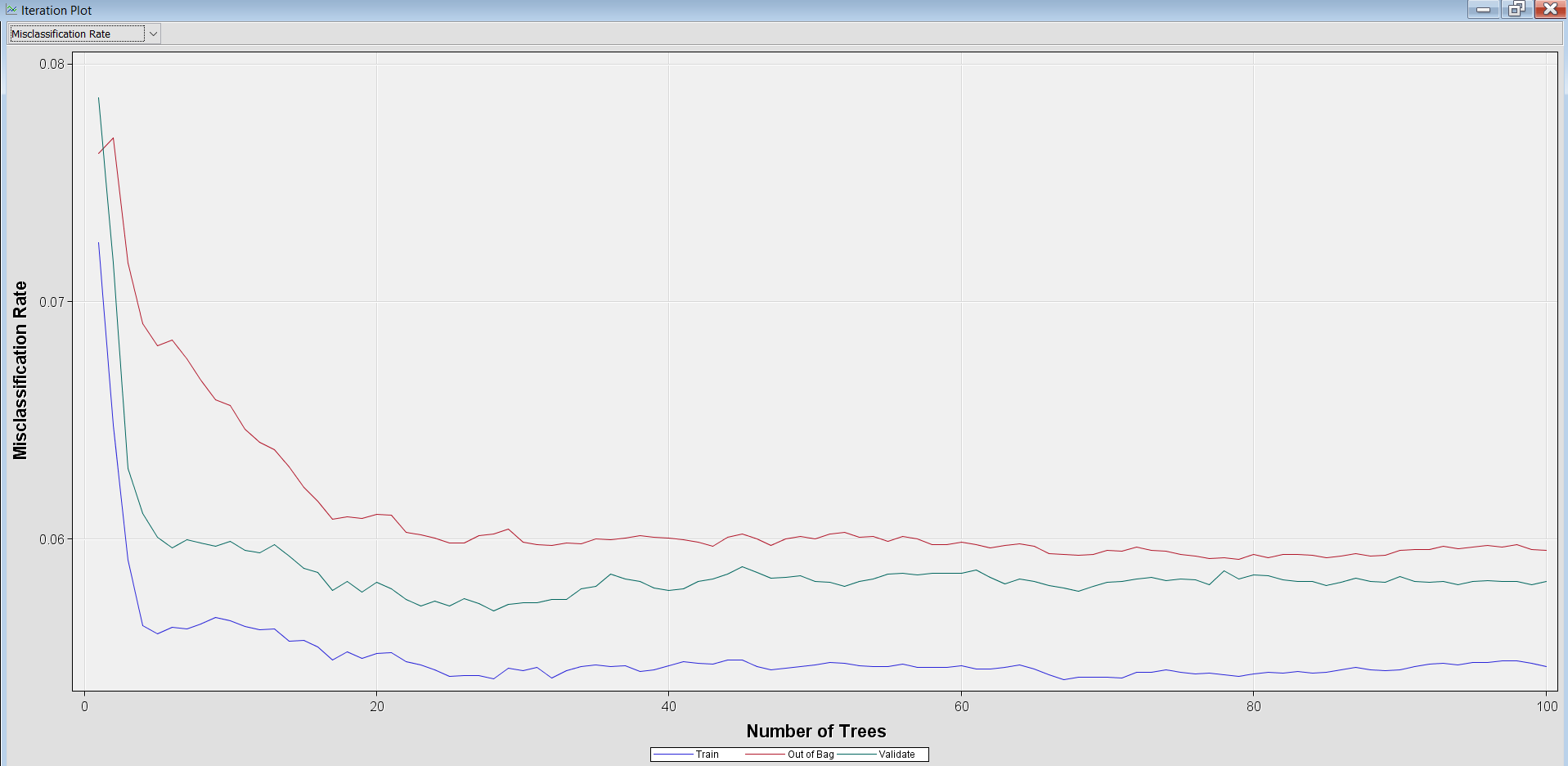


Figure 28 (repeat) – Iteration Plot (HP Forest)

The fit statistics of the model show the misclassification rate.

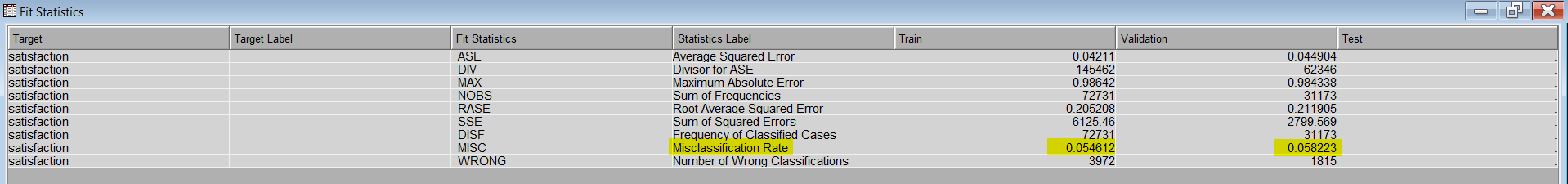


Figure 47 – Fit Statistics of HP Forest

**The validation misclassification is given as 0.058223.** The model information is given in figure 48.

A screenshot of a computer

Description automatically generated with low confidence

Figure 48 – Model Information of HP Forest

**The event classification table shows really good results as there is a high value for true positive and low value for false positive. The correctly predicted values for the validation set is 29,358 out of 31,173 which shows that the model shows good reliablity.**

Table

Description automatically generated

Figure 29 (repeat) – Event Classification Table (HP Forest)

**11.2. HP 1-LAYER NEURAL NETWORK**

In order to map an input to a certain category or anticipated value, the HP Neural node builds multilayer neural networks that transfer information from one layer to the next. You can create neural networks on enormous data sets quickly thanks to the HP Neural node, which enables this mapping to take place in a distributed computing environment.

For the HP neural node in SAS., the architecture has to be set (1 hidden layer) and the number of hidden neurons in that layer (3 neurons).

Since the neural net is a black box model, the only results it shows are as follows,

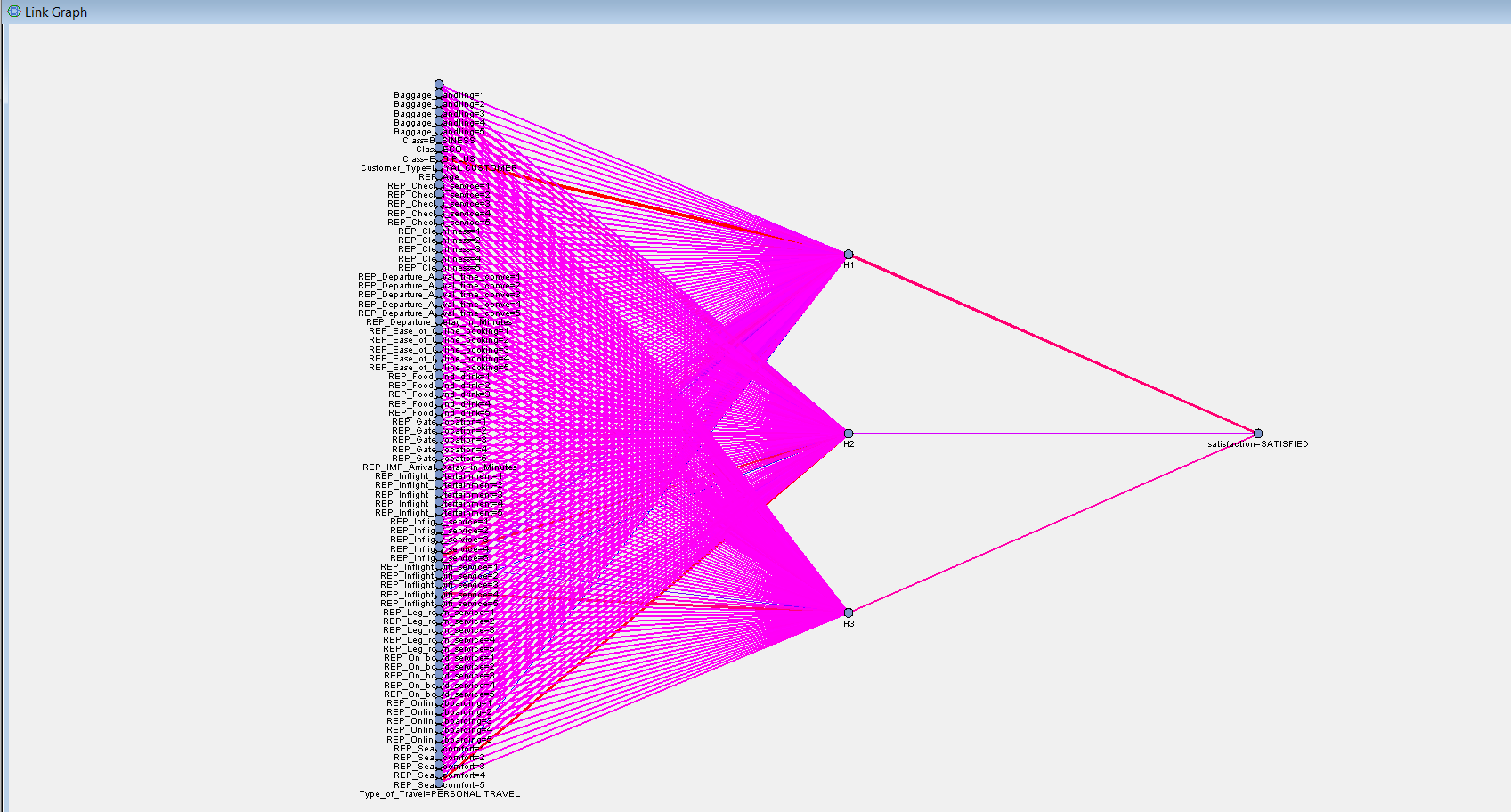


Figure 49 – Link Graph

A picture containing graphical user interface

Description automatically generated

Figure 50 – Fit Statistics (HP 1 Layer Neural Net)

**The validation misclassification rate is 0.057325.**

**The event classification table shows that out of 31,173 validation observations, 29,386 observations were correctly predicted. This shows that the model is highly reliable. However, the working of the model is still a mystery. To understand the neural network, a surrogate model in the form of a decision tree was used.**

Graphical user interface, application

Description automatically generated

Figure 51 – Surrogate Model (HP 1 Layer Neural Net)

Using the node rules of the surrogate model, business interpretations and recommendations can be made.

Text

Description automatically generated

Figure 52 – Surrogate Model Node Rule (1)

Text

Description automatically generated

Figure 53 – Surrogate Model Node Rule (2)

**12. DISCUSSIONS**

**12.1. BUSINESS ANALYSIS/RECOMMENDATIONS (HP FOREST)**

The variable importance of the HP forest model is,

Table

Description automatically generated

Figure 54 – Variable Importance of HP Forest

The model suggests that the airline company should focus mainly on the provision of inflight Wi-Fi services to significantly boost the satisfaction of their customers. Similarly, their online boarding services provide tremendous convenience to the flight passengers, and this satisfies them. Considerable efforts can also be put toward their baggage handling and inflight services.

Food and drink services don’t mainly affect the satisfaction of the customers and passengers also tend to be understanding toward delays in the part of the airline company.

**12.2. BUSINESS ANALYSIS/RECOMMENDATIONS (HP 1 LAYER NEURAL NETWORK)**

The nodes 9 and 10 shown in figure 53 emphasize that the higher rating of online boarding and inflight Wi-Fi services lead to the prediction of a customer to be satisfied. Conversely, looking at nodes 13 and 15, the lower these variables are rated, the higher the possibility of a passenger being dissatisfied overall.

**Therefore, these are the most important factors the company must focus on.**

Figure below shows the node 25 rule. This tells us that for passengers travelling for personal reasons online boarding is of highest importance as they are likely to be dissatisfied with over all service even if the Wi-Fi service is of the highest quality.

Text

Description automatically generated

Figure 55 – Surrogate Model Node Rule (3)

Other node rules also state the importance of inflight entertainment and other services.

**However, similar to the understanding obtained from the HP forest model, customers' pleasure isn't primarily impacted by food and beverage offerings, and passengers also tend to be understanding of delays on the side of the airline operator.**

**Since the HP 1 Layer Neural Net has a lower validation misclassification rate, ultimately, this model is suggested over the HP forest model. Neural networks are highly efficient and continuously learning and thus the base model by itself, shows an advantage over HP forest model.**

**13. CONCLUSIONS/PERSONAL REFLECTION**

The study conducted in this report details the building of a machine learning model based on airline passenger survey conducted to predict the overall satisfaction of the customer. The dataset was sufficiently large and challenging due to the high number of independent variables present and observations. The work conducted in this study includes understanding toward the domain of the airline sector and why it is important to understand the need for predicting customer satisfaction.

Data visualization was also conducted using Tableau software which provided the experience in using the software and understanding the need for better visualizations and the benefits it provides.

The dataset used was challenging in many ways. The output variable – satisfaction, has 2 levels, namely, satisfied, and neutral or dissatisfied. The challenge posed is the vague understanding of whether the passenger is neutral and/or dissatisfied. Other challenges include the overall size of the dataset (>100,000 observations) and the problems of data imputation and preparation. In the process of building and comparing the models presented in this report, my understanding of machine learning and its applications has greatly improved. The use of SAS Enterprise Miner for the purpose of this study created new avenues in my mind towards machine learning that were previously absent.