CT045-3-M-ABAV

ADVANCED BUSINESS ANALYTICS AND VISUALIZATION

INDIVIDUAL ASSIGNMENT

TECHNOLOGY PARK MALAYSIA

CSSE_CT045-3-M-ABAV_L__2022-11-04__PT

Part C: Predictive analytics on Understanding Revenue of GBI bike company by doing sales prediction on the revenue USD

HAND OUT DATE :04 NOVEMBER -2022

HAND IN DATE : 2 January -2023

Student's name : Muhammad Arif Bin Jamaluddin
Student's ID number:TP067696
Lecturer's name : Raheem Mafaas

INSTRUCTIONS TO CANDIDATES:

1 This assignment should be submitted through outline facilities madeavailable to the students.

2 Students are advised to underpin their answers with the use of references(cited using the Harvard Name System of Referencing).

3 Late submission will be awarded zero (0) unless ExtenuatingCircumstances (EC) are upheld.

4 Cases of plagiarism will be penalized.

5 You must obtain 50% overall to pass this module.

Abstract

The topic proposed for further understanding business analytics and visualization is “Predictive analytics on Understanding Revenue of GBI bike company by doing sales prediction on the revenue USD’’, whereby SAS Enterprise Miner is used to further analyse and study the datasets chosen which is GBI datasets. As the main objective is to predict or forecast the revenue of the bicycle thus in the target variable the revenue USD is selected and for the statistical output the machine learning model that has the lowest average square error is chosen. In this analysis HP decision tree has the lowest average squared error. Thus, decision tree is the best model to predict the revenue of GBI bike company for the best model that can be used for predictive modelling.

Keywords: Revenue, Bicycle, Machine learning, Prediction, Forecasting

Table of Contents

1.0 Introduction

This part is to study and understand the predictive analytics method of datasets chosen for this topic which is the GBI datasets. For this part the proposed topic is bikes sales analysis and prediction on the GBI bike datasets using machine learning method to help this company improve their business decision making. Some of the variables for this GBI datasets are country, year, quarter, month, date, price, state, quantity, profit, currency and more. As some of the variable is not significant it will be drop or deleted later on during the analysis using the SAS enterprise miner software.

Thus, with the variables from the datasets it is known that target variables that can be selected is to predict bicycle sales of this company. SAS enterprise miner is used to perform exploratory data analysis on the datasets to gain valuable insights and apply machine learning algorithms later on. As the main core of the datasets is related to bicycle sales data, thus the predicted sales can be forecasted from the bike data sales of GBI bike company from United States and Germany and from the variables given which will be explored later on. There are two sections for this assignment to understand the problem faced by GBI bike company which the first one is using the SAS enterprise miner to do predictive modelling using decision tree, linear and logistic regression, and neural network, and after that is descriptive modelling using SAS Enterprise miner as well to do the clustering analysis and market basket analysis.

Thus, in part A it discusses about the business understanding of the datasets chosen where it talks about the domain of the datasets where it is about sales domain. GBI has records of data regarding their sales, revenue, profit of their bicycle both in Euro and United States dollars and another variable such as unit cost of goods issue that is removal of goods or materials out of the warehouse, the division description whether it is selling bicycle or the accessories, the quantity of bicycles sold and more. In the part B it talks about the data visualization using Tableau software to explore the GBI datasets using bar charts, graphs, pie chart and more and part C is using the SAS enterprise Miner to do predictive modelling.

Before that, brief explanation is done to understand the history of the GBI Bike inc. Global Bike INC was founded 20 years ago where it roots come from both off-road trail-racing and long-distance racing sports as it founders want to developed bicycle that last long, durable and can withstand extreme weather and conditions. Thus, later on this company found a huge success to continuously deliver high class and quality performance for riders that demands high quality bicycle. This company was founded by John Davis and Peter Schwarz where both of them meet in 2000 and form a business partner where both of their company merged together and formed Global Bike Incorporated where both of them acts as co-founder of the company where John responsible for sales, marketing, service & support, IT, finance and human resources groups and Peter is responsible for research, design, procurement, and manufacturing groups from an organizational reporting perspective.

As the datasets is related about bicycle and cycling thus it is known that trend of predicting of bicycle demand can be done using the GBI datasets. For example, one research paper discussed regarding the folding bicycle prospective buyer prediction model where it is known that during the covid pandemic in 2020, cycling has become much popular among the public and many people shops bicycle through offline and online shops. This is due to people starts to realize to be healthy during covid and for certain individuals it is an excellent form to develop physical fitness and minimize the risk of health problems (P. J. W. Van Den Noort, 2016)

Google trends shows that the there is a trend the number of folding bikes has increased to 900% and mountain bikes saw an increase from 680% and the last rank is increasing trend to 300% for road bikes or racing bicycles (T. K. Yunianto, 2020). It is same with GBI bike company that sells various type of bicycles such as

- Professional touring bike
- Off-road bike
- Deluxe touring bike
- Other outdoor bicycle accessories

Where it has both division for both male and female and going back to the folding bicycle prospective buyer prediction model it shows that the data is mostly are triggered from the folding bike keywords and often the time people or customers who might be interested to buy the folding bikes has few criteria for example if the bicycle is small, lightweight, and foldable. It noted that, the one of the criteria is that it is easily transportable to other cars and vehicles for long trips and another advantage is that for folding bike it can go through passage or pathway that other bicycle can’t passed through. Thus, with folding bike it can be easily lifted and continue with the journey again.

GBI bike customers in general or people in general before deciding to purchase any of the bicycle, typically it will start with reading online reviews of the products on the internet but often the time this online review may cause confusion. The confusion is due to biased towards certain models or less experienced in reviewing products online. Another research paper discusses about forecasting of the bicycle sales prediction titled “Perfect casting for cycling” where it can be implemented in the GBI datasets to predict the revenue of the GBI bicycle sales.

2.0 Related works

There are several works that are related to the bicycle business domain for this proposed topic such as using machine learning techniques to predict the forecasted sales of the bicycles and also bike sharing.

One of the research papers discuss about the domain of the bicycle industry in Indonesia which can be related with the GBI as the domain is selling bicycles where it is founded that the in Indonesia the demand of bicycle has increased by 1000% where it is often indicates that bicycles are not only used as transportation but also daily transportation. It also can be seen that for the global data for bicycle industry indicates higher demand during the pandemic and thus number of sales is increasing and for example demand for bicycles increasing by 40 percent in the United States. On the other hand, in the United Kingdom the trend of bicycle demands for personal used is increasing by 33 %, bike sharing increased by 12 percent and in France there is increasing budget for the bicycle parking facilities. This indicates that due to covid 19 pandemic, bicycle has become option or alternative for other people as mode of transportation and also maintain their health by exercising (F. Pradolo, 2020).

There are several bicycle types sold for example sold such as road bikes, mountain bikes and folding bikes (A. B. Tamtomo, 2020) as shown in figure below.

Bike Type	Terrain	Speed	Mobility
Road Bike	On-road	Fast	Low
Mountain Bike	Off-road	Medium	Medium
Folding Bike	On-road	Slow	High

Figure 1

Hence, this table shows the type of the bicycle where the price varies from cheap to more premium section and also the speed, terrain, and the mobility. This also can be seen at the GBI datasets where in the material master description parts there is several types of bikes sold by the GBI company such as professional touring bikes and also GBI also sells accessories for bicycle safety gears such as off-road helmet, knee pads, water bottle, elbow pads and more. There are few problems in selecting the best bicycle for the folding bike prediction buyer model whereby the variable is not only limited to the practicality, speed, size but it is also correlate with the prospective of the buyer where the variables are

- Budget
- Gender
- Age
- Body posture

Most of the time someone who is an entry level buyer do not have the experience choosing the best or have the knowledge on selecting the appropriate bicycle and hence predictive models can be used to choose the right bicycle model. In this research paper it uses machine learning method to predict whether the buyers will buy either folding bicycle or another type of bicycle that is suitable for them.

There are several problems faced in order to select the type of bicycle whereby Zaki et al. proposed a new solution where it used binary classification method where it is implemented to classify the difference between motorized bicycle and non-motorized bicycle as one the crucial variables are speed, and it indicates the performance analysis is around 93%. (M. H. Zaki, T. Sayed, X. Wang, 2016). The next researched is done by using various machine learning algorithms such as Random Forest, Decision Tree, Support Vector Machine, KNN, Logistic Regression where Jaya Prada et al. study the classification of the bicycle buyer using these algorithms for model prediction and found out the Random Forest classifier has the best accuracy at 0.86 (S. Jaya Prada, A. Geetha Sri, B. Venkateswarlu, C. Vineesha, P. Lakshmi Teja, 2020).

2.1 GBI bicycle sales exploratory data analysis and prediction

The GBI datasets contain bicycle sales in Europe which is in Germany and in America which is in the United States and thus, the relationship between the variables can be studied and future prices can be predicted or forecasting the sales using machine learning techniques.

2.2.0 Problem statement

GBI company involves in making bicycle and thus as a company that selling products it deals with business problems such understanding the profit and revenue made from selling the bicycle and also forecasting or predicting the revenue profit of selling the bicycles to the customers. There are several challenges that faced by the GBI maybe due to business challenges such as trying to understand the revenue trend, predicting the sales forecast of the GBI bicycle products, understanding GBI customer segmentation and more. Later the datasets will be processed by performing exploratory data analysis to gain valuable insights and further apply in machine learning algorithms.

2.3.0 Aim and objectives

2.3.1 Aim

The aim of this assignment is to use the GBI datasets and to understand and explore the revenue of the GBI bike company by doing exploratory data analysis and do forecasting or sales prediction of the revenue and profit in terms of the bicycles sold in the region of both United States and Germany of the company. Thus, by using SAS Enterprise Miner this can be achieved as this software provides deeper and detail understanding of the datasets by exploring the predictive analytics using machine learning method such as decision tree, artificial neural network, and logistic regression where each model will give different results in terms of accuracy, precision, F1 score and more.

2.3.2 Objectives

1. To provide graphical and visual presentation of the GBI datasets regarding their profit, revenues, and sales of the bicycle using SAS Enterprise Miner
2. To provide forecasting or sales prediction or in laymen terms revenue of selling their bicycle
3. To do segmentation of the customers according to the type of bicycles sold based on the customers preferences

2.3.3 Scope

The purpose of sales prediction of the bicycles is important as GBI company need to have better understanding to predict their revenue and profit that they made and as GBI as a company is complex where SAP process order of this company starts with the GBI company accounting team made purchase order and after that enquiry about the quotation of products and parts from the customers confirmation. Hence after the sales has been made, it notifies the warehouse to check the stock inventory and after that logistic department starts to deliver the products after the sales office received the invoice/ bills from the customer through accounting process. This whole process is a part of ERP (Enterprise Resource Planning) where SAP and Tableau are used as tools for CRM (Customer Relationship Management) to understand customers interactions in a business organization.

The variables that can be studied or analysed are cost of goods, price, profit, revenue, quantity, and many more and the GBI datasets is created by Epistemy Press books to understand business process using SAP ERP(https://medium.com/codex/global-bike-inc-is-in-need-of-help-d1b9abf411ce).

3.0 Predictive analytics method

The GBI datasets is used to study the sales prediction for the GBI bike company for both revenue in USD using the machine learning technique method available in the SAS ENTERPRISE Miner software.

3.1 Datasets

This part discussed the brief description of the datasets of the GBI where after file import the excel file into the SAS Enterprise Miner the output of the variable's summary shows the datasets contain set of numbers contain 47992 observations and 51 variables. Figure below shows the table of datasets that contain the summary and the results of the metadata and figure 2 shows how the file for the GBI datasets in excel format is imported into the SAS Enterpriser Miner.

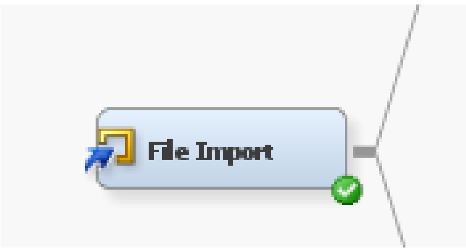


Figure 2

The figure 3 below shows the variable summary of the excel files imported after running the results of the output variables where the target variable is set for revenue USD.

```
Variable Summary

           Measurement    Frequency
Role          Level         Count

ID          INTERVAL           2
ID          NOMINAL            3
INPUT       INTERVAL           8
INPUT       NOMINAL            7
REJECTED    INTERVAL          15
REJECTED    NOMINAL           15
TARGET      INTERVAL           1
```

```
The CONTENTS Procedure

Data Set Name        EMWS5.FIMPORT_DATA                                    Observations          47992
Member Type          DATA                                                  Variables             51
Engine               V9                                                    Indexes               0
Created              14/12/2022 12:03:34                                   Observation Length    480
Last Modified        14/12/2022 12:03:34                                   Deleted Observations  0
Protection                                                                 Compressed            NO
Data Set Type                                                              Sorted                NO
Label
Data Representation  SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64
Encoding             utf-8  Unicode (UTF-8)


                              Engine/Host Dependent Information

Data Set Page Size          131072
Number of Data Set Pages    177
First Data Page             1
Max Obs per Page            272
Obs in First Data Page      254
Number of Data Set Repairs  0
Filename                    /home/u61522473/ABAV112022/Workspaces/EMWS5/fimport_data.sas7bdat
Release Created             9.0401M6
Host Created                Linux
Inode Number                5436250208
Access Permission           rw-r--r--
Owner Name                  u61522473
File Size                   22MB
File Size (bytes)           23330816
```

Figure 3

Figure 4 below shows the variables of the GBI datasets after imported into the SAS Enterprise Miner. It shows the list of variables that the software identifies from the datasets and also indicates the type of each of the variables whether it is a character or numeric.

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
25	Accounting_Document_Number	Char	10			
22	Billing_Date	Num	8			
23	Billing_Document_Number	Char	8			
7	City	Char	13	$13.	$13.	City
48	Cost_of_Goods_Sold_EUR	Num	8			
47	Cost_of_Goods_Sold_USD	Num	8			
1	Country	Char	13	$13.	$13.	Country
11	Currency	Char	3	$3.	$3.	Currency
6	Customer	Num	8	BEST.		Customer
26	Customer_Name	Char	20			
5	Day	Num	8	BEST.		Day
21	Delivery_Number	Char	8			
44	Discount_EUR	Num	8			
43	Discount_USD	Num	8			
36	Distribution_Channel	Char	2			
37	Distribution_Channel_Description	Char	9			
9	Division	Char	2	$2.	$2.	Division
35	Division_Description	Char	11			
51	Exchange_Rate_at_Quote__USD_Euro	Num	8			
12	Layer_Number	Num	8			
14	Layer_Order_Concatenated	Char	8			
15	Layer_Order_Line_Concatenated	Char	10			
29	Material_Group	Char	5			
30	Material_Group_Description	Char	14			
28	Material_Master_Description	Char	34			
27	Material_Number	Char	8			
4	Month	Num	8	BEST.		Month
13	Order_Number	Num	8			
24	Payment_Receipt_Date	Num	8			
20	Post_Goods_Issue_Date	Num	8			
40	Price_EUR	Num	8			
39	Price_USD	Num	8			
50	Profit_Margin_EUR	Num	8			
49	Profit_Margin_USD	Num	8			
10	Quantity	Num	8	BEST.		Quantity
3	Quarter	Num	8	BEST.		Quarter
16	Quote_Date	Num	8			
17	Quote_Number	Char	8			
46	Revenue_EUR	Num	8			
45	Revenue_USD	Num	8			
33	Sales_Area	Char	10			
34	Sales_Area_Description	Char	40			
18	Sales_Order_Create_Date	Num	8			
19	Sales_Order_Number	Char	5			
32	Sales_Org_Description	Char	18			
31	Sales_Organization	Char	4			
8	State	Char	2	$2.	$2.	State
42	Unit_Cost_at_Goods_Issue_EUR	Num	8			
41	Unit_Cost_at_Goods_Issue_USD	Num	8			
38	Unit_of_Measure	Char	2			
2	Year	Num	8	BEST.		Year

Figure 4

The table 1 below shows the summary of the GBI bike company datasets with the descriptions.

Table 1

Variables	Description
Accounting_Document_Number	Document number of key the system uses to access the accounting document of GBI company
Billing_Date	The date generation of statement transactions for previous billing cycle of the company
Billing_Document_Number	This contains billing data of the company for more than one business transactions
City	The city where the products or bicycles was sold to the customers
Cost_of_Goods_Sold_EUR	This is the direct costs of producing goods sold by GBI in euro
Cost_of_Goods_Sold_USD	This is the direct costs of producing goods sold by GBI in USD
Country	The demographic region where the bicycles were sold
Currency	The currency of the country whether it is in euro or dollars
Customer	Numeric values of the customer
Customer_Name	The customer’s name such as Peach Tree Bikes, Silicon Valley Bikes, Big apple Bikes, Northwest Bikes, Furniture City Bikes, DC Bikes, Rocky Mountain Bikes, Socal Bikes, Philly Bikes, Beantown Bikes, Windy City Bikes and more
Day	The recorded day data of the GBI company
Delivery_Number	Delivery number ID
Discount_EUR	The discount given in euro
Discount_USD	The discount given in dollars

Distribution_Channel	The products (bicycle) of the GBI get from the manufacturer to the end user/ customers
Distribution_Channel_Description	The description of the distribution channel where in this case it is a wholesale
Division	It is divided into which are ‘AS’ and ‘BI’
Division_Description	The division of the description which is Accessories and Bicycles
Exchange_Rate_at_Quote__USD_Euro	The exchange rate quote between euro and united state dollars
Layer_Number	The layer number value ID
Layer_Order_Concatenated	The layer order concatenated value
Layer_Order_Line_Concatenated	The layer order line concatenated value
Material_Group	The material group which can categorized as safety and bikes
Material_Group_Description	The material group description where it can be categorized as safety gear and finished bikes
Material_Master_Description	This includes some of the material sold by GBI such as knee pads, deluxe touring bike red, Men's Off-Road Bike, water bottle cage and more
Material_Number	This one represents the material number in form of codes for example knee pads are represented by KPAD1000
Month	The number of months
Order_Number	Represents the tracking number of products delivered by GBI
Payment_Receipt_Date	Date goods or services were received or contractually due
Post_Goods_Issue_Date	Date on which the goods must physically leave the shipping point
Price_EUR	Price of products/ bicycle sold in euro
Price_USD	Price of products/ bicycle sold in dollars

Profit_Margin_EUR	Income over revenues of the GBI company in euro
Profit_Margin_USD	Income over revenues of the GBI company in dollars
Quantity	The quantity of bicycle of GBI company
Quarter	The recorded quarter data of the GBI company
Quote_Date	The quote date ID
Quote_Number	The number allocated to each Quote as set out in the Quote.
Revenue_EUR	The revenue made in euro
Revenue_USD	The revenue made in dollars
Sales_Area	The region of the sales area
Sales_Area_Description	The description of the sales area whether it is located on the east or on the west
Sales_Order_Create_Date	The date on which the order has been created
Sales_Order_Number	An order created for selling the product to the customer
Sales_Org_Description	The region in which where the organization sells the products
Sales_Organization	The sales organization that sells the GBI bicycle products
State	The states in that particular country
Unit_Cost_at_Goods_Issue_EUR	The unit price of moving the goods out of the warehouse in euro
Unit_Cost_at_Goods_Issue_USD	The unit price of moving the goods out of the warehouse in dollars
Unit_of_Measure	Unit measure ID
Year	The year data recorded

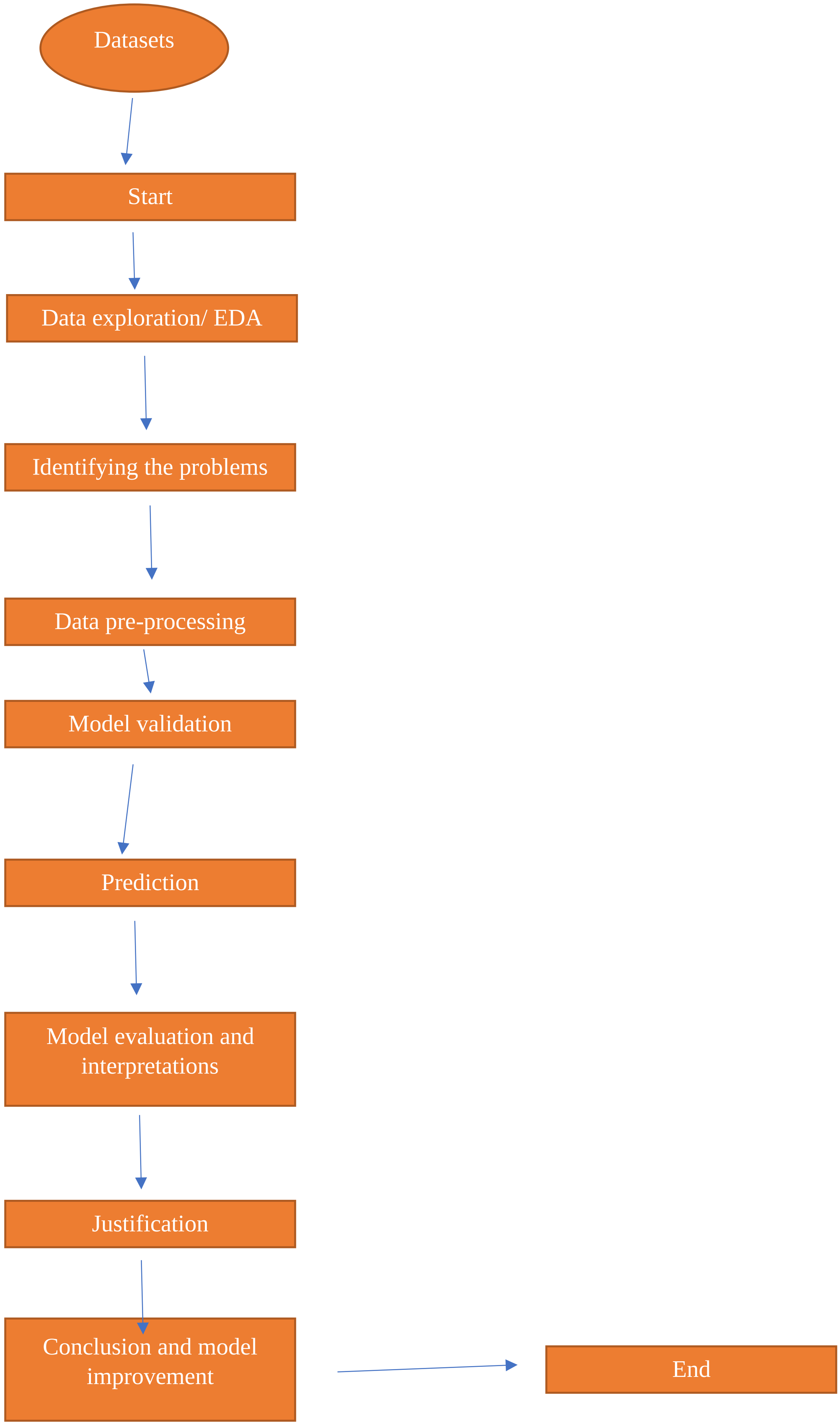
3.2 Methodology

The methodology used for this analysis is SEMMA method where in this case there is a slight change occurs on the methodology selected previously which was CRISP-DM in part A but for this one SEMMA method is selected. This technique was developed by SAS where the acronym SEMMA stands for Sample, Explore, Modify, Model, Assess which when referring to a Data Mining project. Below shows the 5 stages of SEMMA process with some brief explanation on each of the stages or steps.

1. Sample = This part involves the data sampling by getting a few portions of large datasets for example in this case the GBI datasets which contain significant information.
2. Explore = This part is the data exploration whereby anomalies or outliers in the trends of the datasets are explored to gain insights and ideas about the anomalies
3. Modify = This part undergoes data modification whereby it undergoes modify the data by creating, selecting, and doing the transformation variables so that it can focus on the model selection section.
4. Model = This part is the data modelling where SAS Enterprise Miner are allowed to automatically search and find the combination of data that are good and very much reliable to predicts the targeted or desired outcome
5. Assess = This part is the assessment stage where it assessed the data by evaluating the reliability or usefulness of the any of the findings gathered from the data mining process and do estimation on how it performs.

The SEMMA process is pretty much linked with the SAS Enterprise Miner software as although the SEMMA method is not related with the data mining tools chosen. The process flow of SEMMA method is very easy to understand as it allows organized and sufficient data mining project along with good maintenance and development. Thus, with this method it helps to solve to any business problems and find the purpose of data mining business goals (Santos, M & Azevedo, C, 2005).

Below shows how the methodology of predicting revenues or sales of GBI bicycle correlated with SEMMA method where it comprises of the flowchart and brief overview of the model building in SAS Enterprise Miner.


Datasets
Start
Data exploration/ EDA
Identifying the problems
Data pre-processing
Model validation
Prediction
Model evaluation and interpretations
Justification
Conclusion and model improvement
End

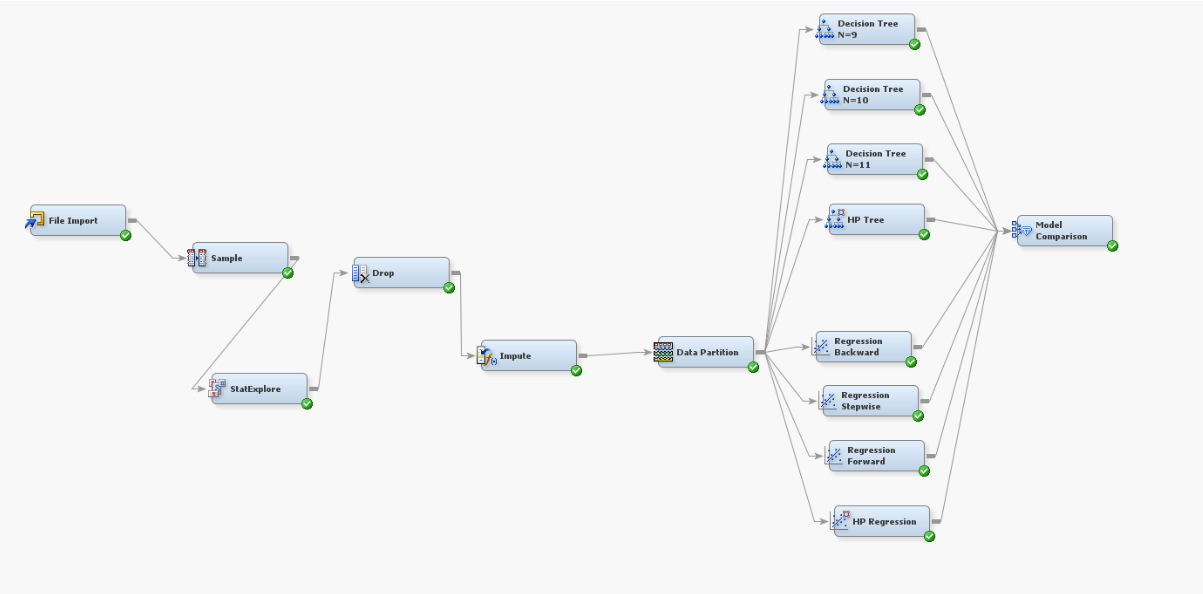


Figure 5

Figure 5 above shows the example of SAS data mining workflow where two machine learning model is used which are decision tree and linear regression.

3.3 Data pre-processing

Data pre-processing is one of the core component of data preparation where it is part of the phase that processed raw data. This also can be called as data transformation where it is a process of converting, cleansing and structuring data into a usable format where the data is analysed to support decision making process. Data preparation is very crucial to produce an accurate machine learning model whereby under the pre-processed part the normalization, noise removal and feature selection are one and to train model it is divided into three parts such as training, validation, and testing. Figure 5 and 6 shows the GBI SAS Missing values and the correlation statistics value where it should be known that the larger the datasets is it took more time to crunch the data. Typically, it not required to do pre-processing from data sampling, dropping the variables and imputation but it is shown for this assignment where proved need to be shown that the datasets are not cleaned. The process of cleaning or cleansing the raw data is done to get rid off duplicate and outliers and getting the best input variable for further analysis using machine learning algorithm

Variable Levels Summary
(maximum 500 observations printed)

Variable	Role	Frequency Count
Customer	ID	25
Material_Number	ID	19
Order_Number	ID	128
Quote_Number	ID	9906
Sales_Order_Number	ID	9818
dataobs	ID	23996

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City	INPUT	24	26	Denver	11.66	Irvine	11.13
TRAIN	Division_Description	INPUT	3	26	Bicycles	53.62	Accessories	46.27
TRAIN	Material_Group_Description	INPUT	3	26	Finished Bikes	53.62	Safety Gear	46.27
TRAIN	Material_Master_Description	INPUT	19	26	Deluxe Touring Bike (silver)	7.19	Deluxe Touring Bike (red)	7.15
TRAIN	Sales_Area_Description	INPUT	9	26	United States West-Wholesale-Bic	21.36	United States East-Wholesale-Bic	20.84
TRAIN	Sales_Org_Description	INPUT	5	26	United States West	37.94	United States East	37.67

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Cost_of_Goods_Sold_USD	INPUT	3852.574	4117.233	23970	26	0	2400	21675.5	0.723072	-0.57047
Discount_USD	INPUT	27.43238	286.9335	23970	26	0	0	4499.834	10.61177	113.4619
Month	INPUT	6.114101	3.431849	23970	26	1	6	12	0.209316	-1.20354
Price_USD	INPUT	1557.536	1455.043	23970	26	14.75	2165	4495.15	0.054782	-1.70487
Profit_Margin_USD	INPUT	4317.061	4615.49	23970	26	19.764	3030	36720	1.001062	1.079353
Quantity	INPUT	8.159783	3.936244	23970	26	1	8	21	0.278618	-0.88374
Quarter	INPUT	2.383813	1.130017	23970	26	1	2	4	0.191153	-1.35137
Unit_Cost_at_Goods_Issue_USD	INPUT	748.169	701.466	23444	552	0	1095	2132.41	0.039822	-1.7394
Revenue_USD	TARGET	8197.068	8503.316	23970	26	40.626	5890	40498.5	0.675323	-0.67415

Figure 6

Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Revenue_USD

Input	Correlation
Profit_Margin_USD	0.96807
Cost_of_Goods_Sold_USD	0.96514
Price_USD	0.86518
Unit_Cost_at_Goods_Issue_USD	0.85689
Discount_USD	0.21430
Quarter	-0.02394
Month	-0.02565
Quantity	-0.49650

Figure 7

3.4 Settings

This part discusses the settings used in the SAS Enterprise Miner.

1) Initial stages

The settings selected is using the FileImport

Property	Value
General	
Node ID	FIMPORT
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Import File	D:\ABAV\GBI_Dataset.xlsx
Maximum Rows to Import	1000000
Maximum Columns to Import	10000
Delimiter	,
Name Row	Yes
Number of Rows to Skip	0
Guessing Rows	500
File Location	Local
File Type	xlsx
Advanced Advisor	No
Rerun	No
Score	
Role	Train
Report	
Summarize	No
Status	
Create Time	12/11/22 6:31 AM
Run ID	031a9278-e617-eb4c-947c-4d682763dd96
Last Error	
Last Status	Complete
Last Run Time	12/11/22 2:31 PM
Run Duration	0 Hr. 0 Min. 56.09 Sec.
Grid Host	
User-Added Node	No

Figure 8

2) Initial stages

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Material_Number	ID	Nominal	No		No	.	.
Order_Number	ID	Interval	No		No	.	.
Customer	ID	Interval	No		No	.	.
Sales_Order_Number	ID	Nominal	No		No	.	.
Quote_Number	ID	Nominal	No		No	.	.
Material_Master_Description	Input	Nominal	No		No	.	.
Material_Group_Description	Input	Nominal	No		No	.	.
City	Input	Nominal	No		No	.	.
Division_Description	Input	Nominal	No		No	.	.
Distribution_Channel_Description	Input	Nominal	No		No	.	.
Sales_Area_Description	Input	Nominal	No		No	.	.
Profit_Margin_USD	Input	Interval	No		No	.	.
Quantity	Input	Interval	No		No	.	.
Price_USD	Input	Interval	No		No	.	.
Month	Input	Interval	No		No	.	.
Quarter	Input	Interval	No		No	.	.
Unit_Cost_at_Goods_Issue_USD	Input	Interval	No		No	.	.
Sales_Org_Description	Input	Nominal	No		No	.	.
Cost_of_Goods_Sold_USD	Input	Interval	No		No	.	.
Discount_USD	Input	Interval	No		No	.	.
Unit_of_Measure	Rejected	Nominal	No		No	.	.
Unit_Cost_at_Goods_Issue_EUR	Rejected	Interval	No		No	.	.
Accounting_Document_Number	Rejected	Nominal	No		No	.	.
Profit_Margin_EUR	Rejected	Interval	No		No	.	.
Sales_Order_Create_Date	Rejected	Interval	No		No	.	.
Year	Rejected	Interval	No		No	.	.
Revenue_EUR	Rejected	Interval	No		No	.	.
Billing_Date	Rejected	Interval	No		No	.	.
Sales_Area	Rejected	Nominal	No		No	.	.
Sales_Organization	Rejected	Nominal	No		No	.	.
State	Rejected	Nominal	No		No	.	.
Quote_Date	Rejected	Interval	No		No	.	.
Discount_EUR	Rejected	Interval	No		No	.	.
Distribution_Channel	Rejected	Nominal	No		No	.	.
Delivery_Number	Rejected	Nominal	No		No	.	.
Exchange_Rate_at_Quote__USD_Euro	Rejected	Interval	No		No	.	.
Division	Rejected	Nominal	No		No	.	.
Currency	Rejected	Nominal	No		No	.	.
Cost_of_Goods_Sold_EUR	Rejected	Interval	No		No	.	.
Country	Rejected	Nominal	No		No	.	.
Day	Rejected	Interval	No		No	.	.
Customer_Name	Rejected	Nominal	No		No	.	.
Billing_Document_Number	Rejected	Nominal	No		No	.	.
Payment_Receipt_Date	Rejected	Interval	No		No	.	.
Material_Group	Rejected	Nominal	No		No	.	.
Price_EUR	Rejected	Interval	No		No	.	.
Post_Goods_Issue_Date	Rejected	Interval	No		No	.	.
Layer_Order_Line_Concatenated	Rejected	Nominal	No		No	.	.
Layer_Number	Rejected	Interval	No		No	.	.
Layer_Order_Concatenated	Rejected	Nominal	No		No	.	.
Revenue_USD	Target	Interval	No		No	.	.

Figure 9

Based on figure 9 it shows the variables that was set as ID, Input, Rejected and Target where revenue USD is set as target variable for this analysis. There are 15 variables set as input, 5 variables that is set as ID and 30 variables that is set as rejected

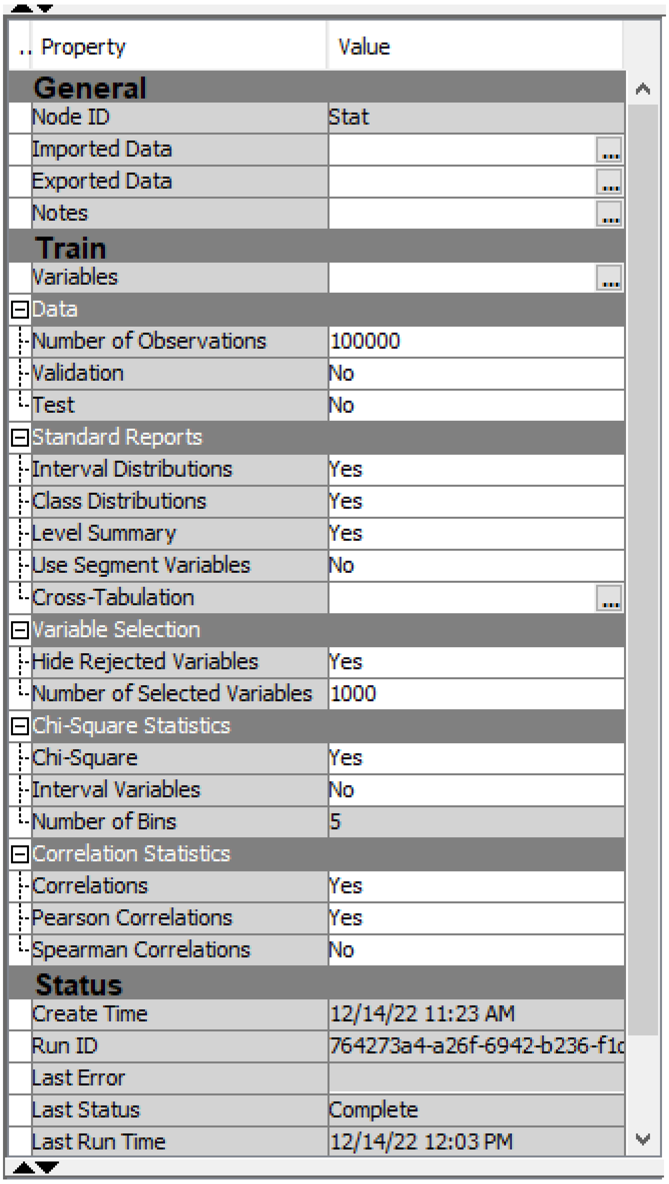
3) Pre-processing

Figure 10 below shows the data sampling method where it is done to analyze the subset of the data in order to have bigger pitcure on the whole datasets used for the machine learning analyis.

General	
Node ID	Smpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Sample Method	Random
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	50.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Proportional
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	100.0
Sample Proportion	50.0
Oversampling	
Adjust Frequency	No
Based on Count	No
Exclude Missing Levels	No
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/14/22 11:20 AM

Figure 10

Figure 11 below shows the statexplore method where this one is used to examine variable distributions and statistics in the data sets.



Property	Value
General	
Node ID	Stat
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Data	
Number of Observations	100000
Validation	No
Test	No
Standard Reports	
Interval Distributions	Yes
Class Distributions	Yes
Level Summary	Yes
Use Segment Variables	No
Cross-Tabulation	...
Variable Selection	
Hide Rejected Variables	Yes
Number of Selected Variables	1000
Chi-Square Statistics	
Chi-Square	Yes
Interval Variables	No
Number of Bins	5
Correlation Statistics	
Correlations	Yes
Pearson Correlations	Yes
Spearman Correlations	No
Status	
Create Time	12/14/22 11:23 AM
Run ID	764273a4-a26f-6942-b236-f1c
Last Error	
Last Status	Complete
Last Run Time	12/14/22 12:03 PM

Figure 11

Figure 12 show the drop method where some of the variables are dropped where the drop node is to remove variables from data sets or hide variables from the metadata. The number of variables drop is from 9 variables to 7 variables.

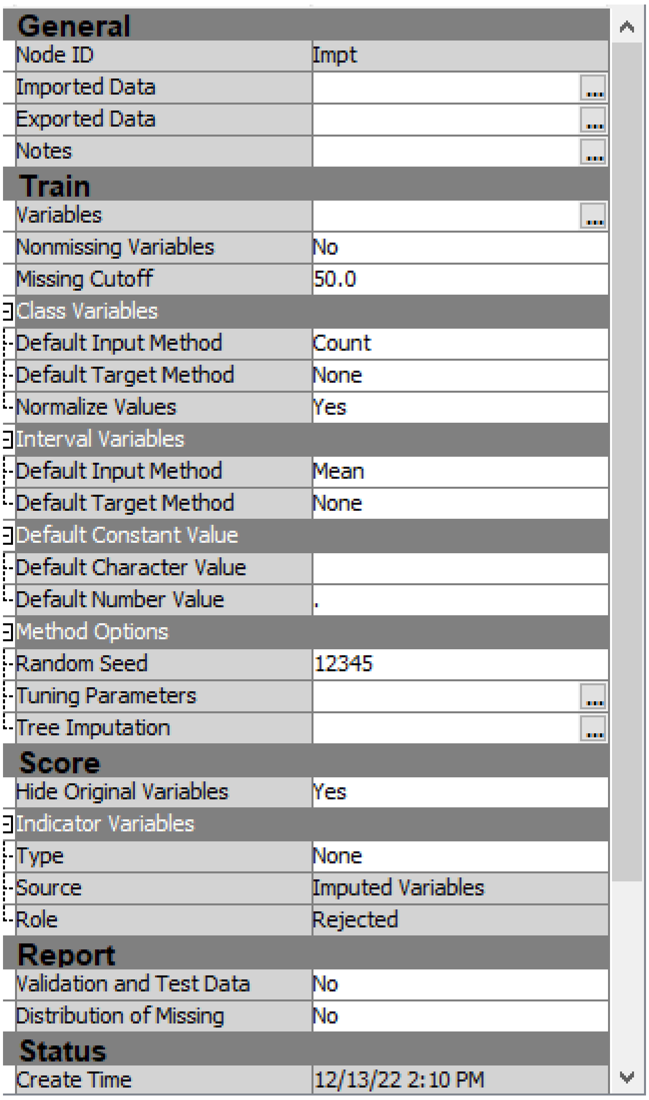
General	
Node ID	Drop
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Drop Selection Options	
Drop from Tables	No
Assess	No
Classification	No
Frequency	No
Hidden	Yes
Input	No
Predict	No
Rejected	Yes
Residual	No
Target	No
Other	No
Status	
Create Time	12/14/22 11:26 AM
Run ID	fd4ec4a5-e553-b64a-9a97-45a2
Last Error	
Last Status	Complete
Last Run Time	12/14/22 12:03 PM
Run Duration	0 Hr. 0 Min. 2.51 Sec.
Grid Host	
User-Added Node	No

Figure 12

Figure 14 shows the imputation method it is done to replace the missing data by substitute value to retain most of the information in the datasets whereby in the edit variables of the impute method, figure 13 below shows the variables that assigned with the method of either tree, count and median.

Name	Use	Method	Use Tree	Role	Level
Cost_of_Goods_Sold_USD	**Default**	**Tree**	**Default**	Input	Interval
Division_Description	**Default**	**Count**	**Default**	Input	Nominal
Material_Group_Description	**Default**	**Count**	**Default**	Input	Nominal
Material_Master_Description	**Default**	**Count**	**Default**	Input	Nominal
Price_USD	**Default**	**Tree**	**Default**	Input	Interval
Profit_Margin_USD	**Default**	**Tree**	**Default**	Input	Interval
Quantity	**Default**	**Tree**	**Default**	Input	Interval
Revenue_USD	**Default**	**Median**	**Default**	Target	Interval
Sales_Area_Description	**Default**	**Count**	**Default**	Input	Nominal
Unit_Cost_at_Goods_Issue_USD	**Default**	**Tree**	**Default**	Input	Interval

Figure 13



Property	Value
General	
Node ID	Impt
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Nonmissing Variables	No
Missing Cutoff	50.0
Class Variables	
Default Input Method	Count
Default Target Method	None
Normalize Values	Yes
Interval Variables	
Default Input Method	Mean
Default Target Method	None
Default Constant Value	
Default Character Value	
Default Number Value	.
Method Options	
Random Seed	12345
Tuning Parameters	...
Tree Imputation	...
Score	
Hide Original Variables	Yes
Indicator Variables	
Type	None
Source	Imputed Variables
Role	Rejected
Report	
Validation and Test Data	No
Distribution of Missing	No
Status	
Create Time	12/13/22 2:10 PM

Figure 14

Figure 15 shows the data partition where the data is split into training and validation set whereby it is used to observe the performance of the model on the data. Based on the figure 15, the train data is set as 70% and the validation data is set at 30% and usually the datasets is randomized beforehand to get rid of any biases. Figure 16 and 17 shows the decision tree model and 18 and 19 shows the linear regression model.

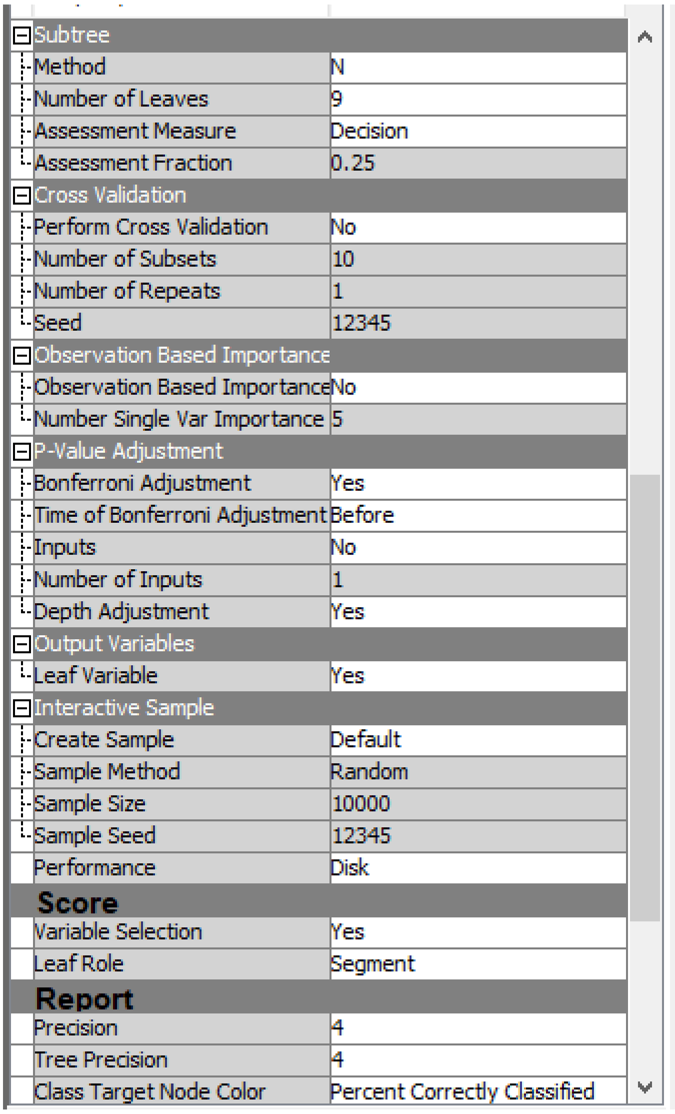
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	Yes
Class Targets	Yes
Status	
Create Time	12/13/22 1:53 PM
Run ID	2cc50d80-b178-a84a-9e9a-4f64
Last Error	
Last Status	Complete
Last Run Time	12/14/22 12:09 PM
Run Duration	0 Hr. 0 Min. 2.96 Sec.
Grid Host	
User-Added Node	No

Figure 15

4) Type of model

General	
Node ID	Tree
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	N

Figure 16 Decision tree



Subtree	
Method	N
Number of Leaves	9
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	
Leaf Variable	Yes
Interactive Sample	
Create Sample	Default
Sample Method	Random
Sample Size	10000
Sample Seed	12345
Performance	Disk
Score	
Variable Selection	Yes
Leaf Role	Segment
Report	
Precision	4
Tree Precision	4
Class Target Node Color	Percent Correctly Classified

Figure 17 Decision tree

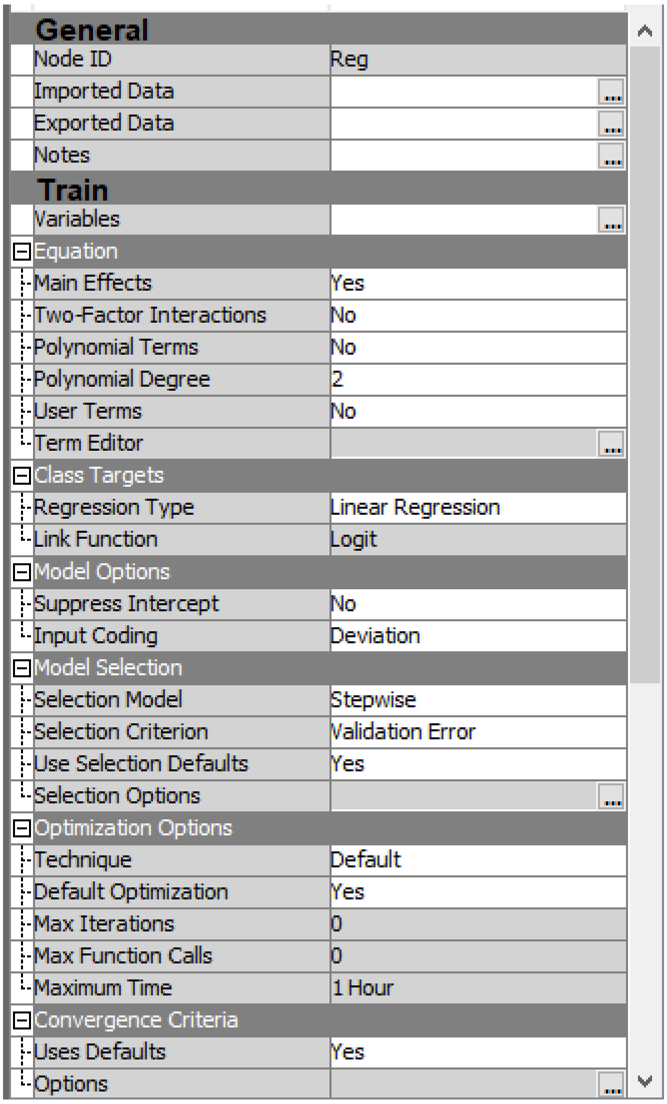


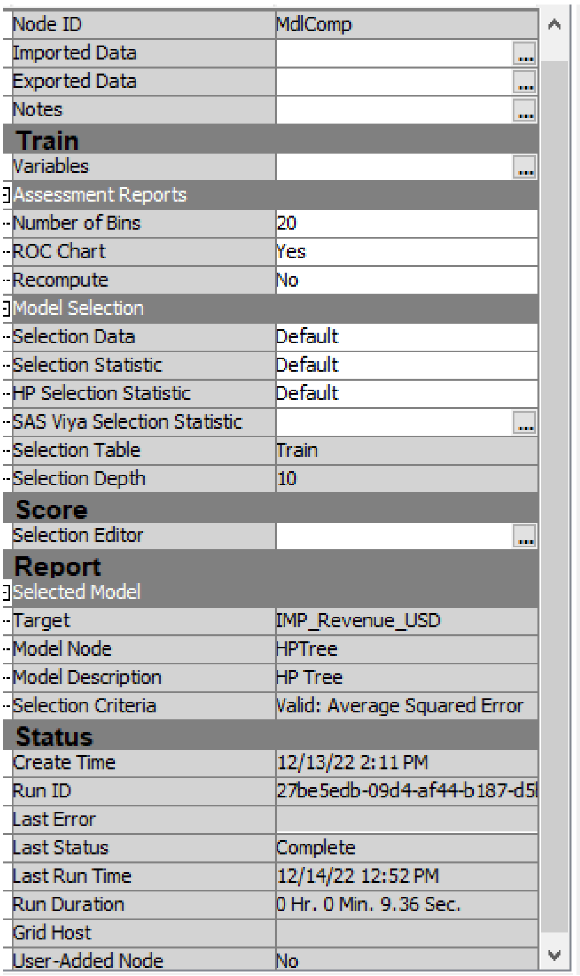
Figure 18 Linear regression

Output Options	
Confidence Limits	No
Save Covariance	No
Covariance	No
Correlation	No
Statistics	No
Suppress Output	No
Details	No
Design Matrix	No
Score	
Excluded Variables	Reject
Status	
Create Time	12/13/22 3:47 PM
Run ID	e3700832-0a36-bc44-aae8-d5
Last Error	
Last Status	Complete
Last Run Time	12/14/22 12:18 PM
Run Duration	0 Hr. 0 Min. 4.43 Sec.
Grid Host	
User-Added Node	No

Figure 19 Linear regression

Figure 20 shows the model comparison setting between the two-machine learning model proposed which is the decision tree and linear regression.

5) Model comparison



Node ID	MdlComp
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Assessment Reports	
Number of Bins	20
ROC Chart	Yes
Recompute	No
Model Selection	
Selection Data	Default
Selection Statistic	Default
HP Selection Statistic	Default
SAS Viya Selection Statistic	...
Selection Table	Train
Selection Depth	10
Score	
Selection Editor	...
Report	
Selected Model	
Target	IMP_Revenue_USD
Model Node	HPTree
Model Description	HP Tree
Selection Criteria	Valid: Average Squared Error
Status	
Create Time	12/13/22 2:11 PM
Run ID	27be5edb-09d4-af44-b187-d5l
Last Error	
Last Status	Complete
Last Run Time	12/14/22 12:52 PM
Run Duration	0 Hr. 0 Min. 9.36 Sec.
Grid Host	
User-Added Node	No

Figure 20

4.0 Experiments

4.1.0 Data visualizations

This part explained the input variables of the datasets where data visualizations is done on to study the distributions of the input variables in terms of graphs, pie charts and more. In this part only a few inputs variables are used to explain the data visualizations where Material master description, material group description, city, sales area description, profit margin USD and quantity is used to do descriptive analytics. Figure 21 shows the graph of the material master description whereby professional touring bike black has the highest frequency at 448 whereby repair kit is the lowest at 177. This also shows that professional touring bike black bring the highest revenue USD and the repair kit is the lowest for the revenue USD.

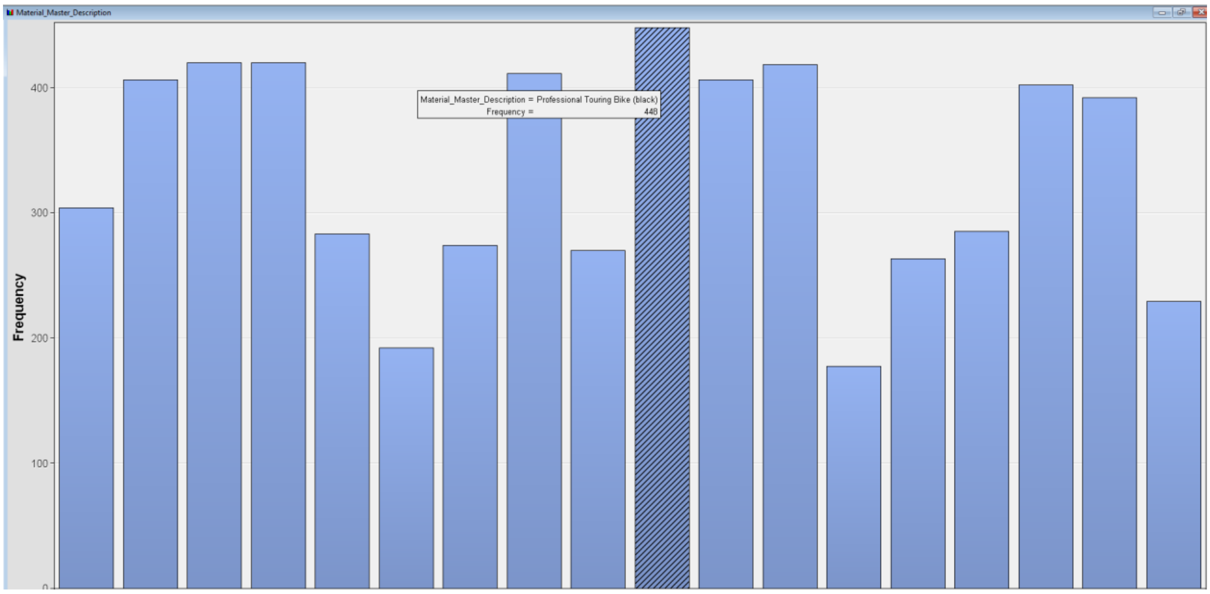


Figure 21

Figure 22 shows the pie chart for the material group description where it is divided into safety gear at 47.37% and finished bikes at 52.63% and this indicates that GBI as a company solely and mainly focus on the selling bicycles than selling other accessories such as safety gear.

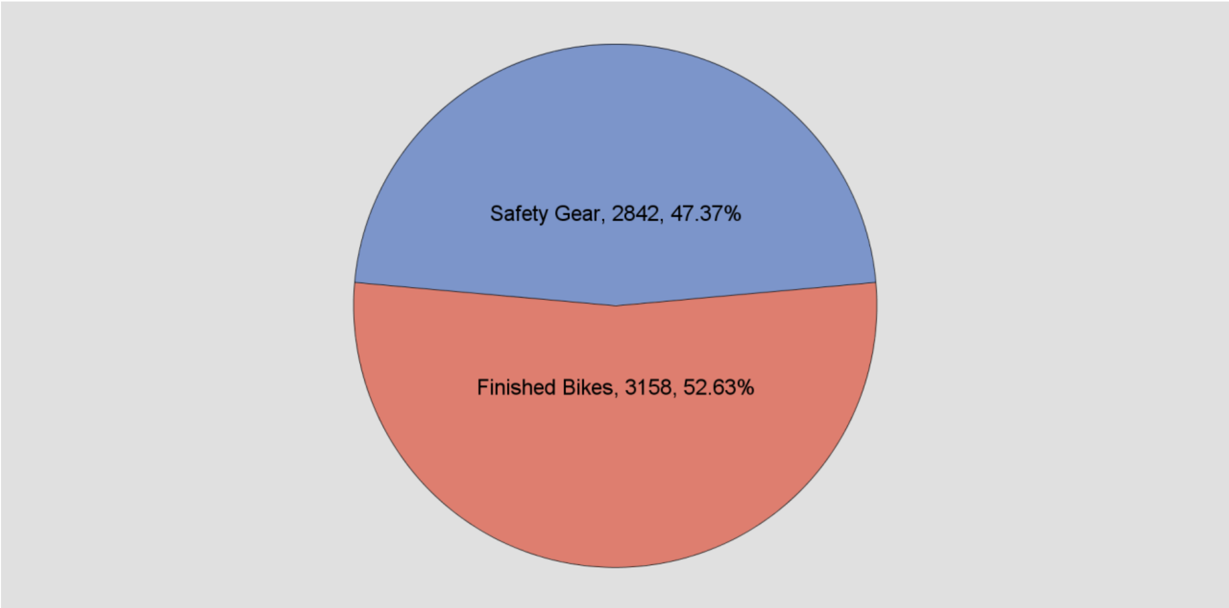


Figure 22

Based on figure 23 it shows the city description on which the most profitable or bring the most revenue USD to the company and it shows the Palo Alto bring the most revenue. It follows by the city of Denver, Irvine, Seattle and so on. The least city that made less revenue for GBI is Washington DC.

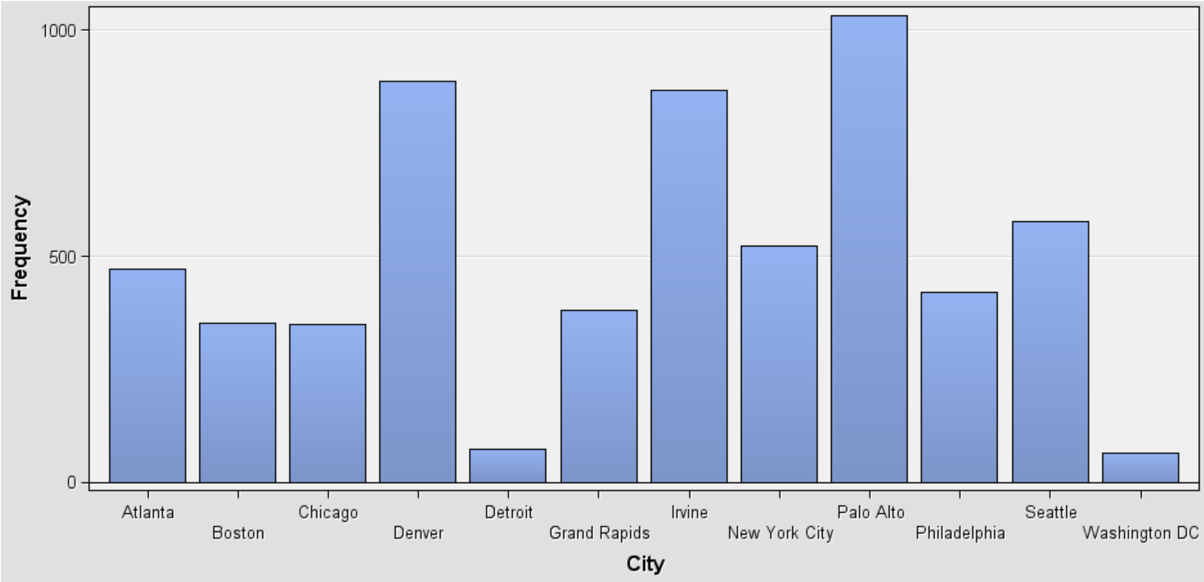


Figure 23

Figure 24 shows the pie chart of the sales area description whereby the pie chart is almost equally distributed in terms of the percentage values and the highest one is United States West Wholesale bicycles at 29.62% and the lowest is United States East Wholesale accessories at 20.87%

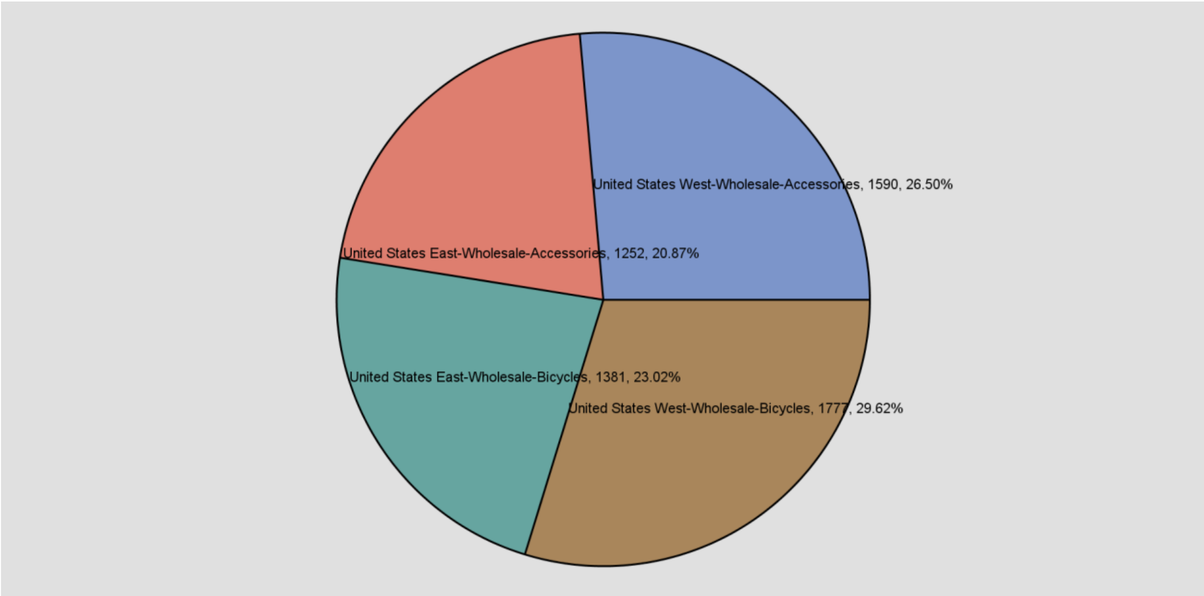


Figure 24

Figure 25 shows the graph of the profit margin USD between 32 and 2908 USD is the most frequent one for the GBI where the value of profit margin and figure 26 shows the graph distribution of the quantity where it shows that quantity 5 and 6 has the highest frequency. This also indicates that quantity 5 and 6

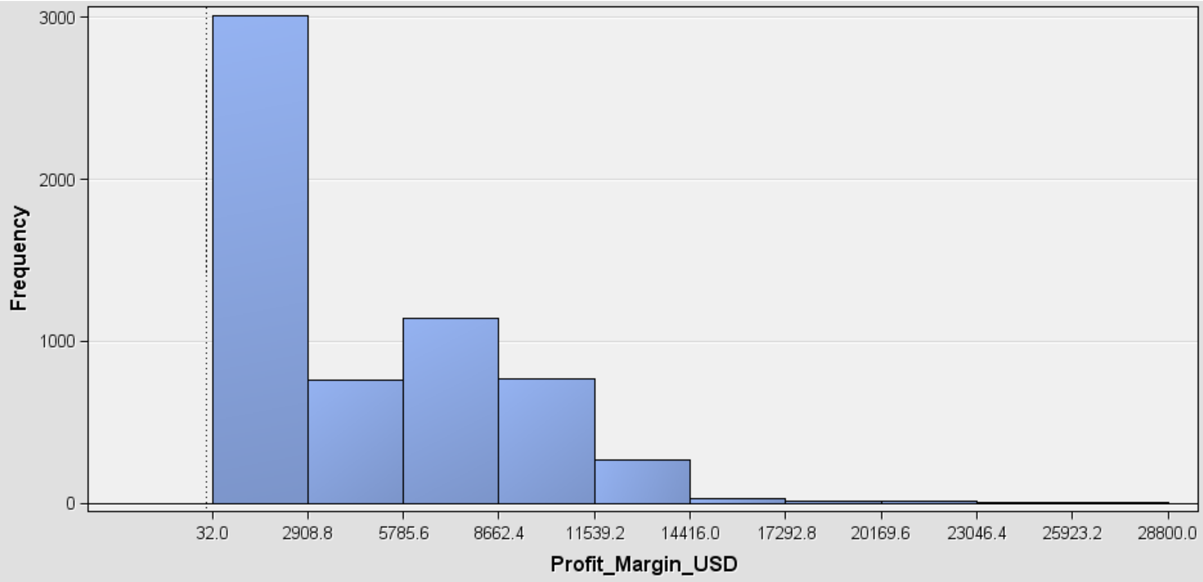


Figure 25

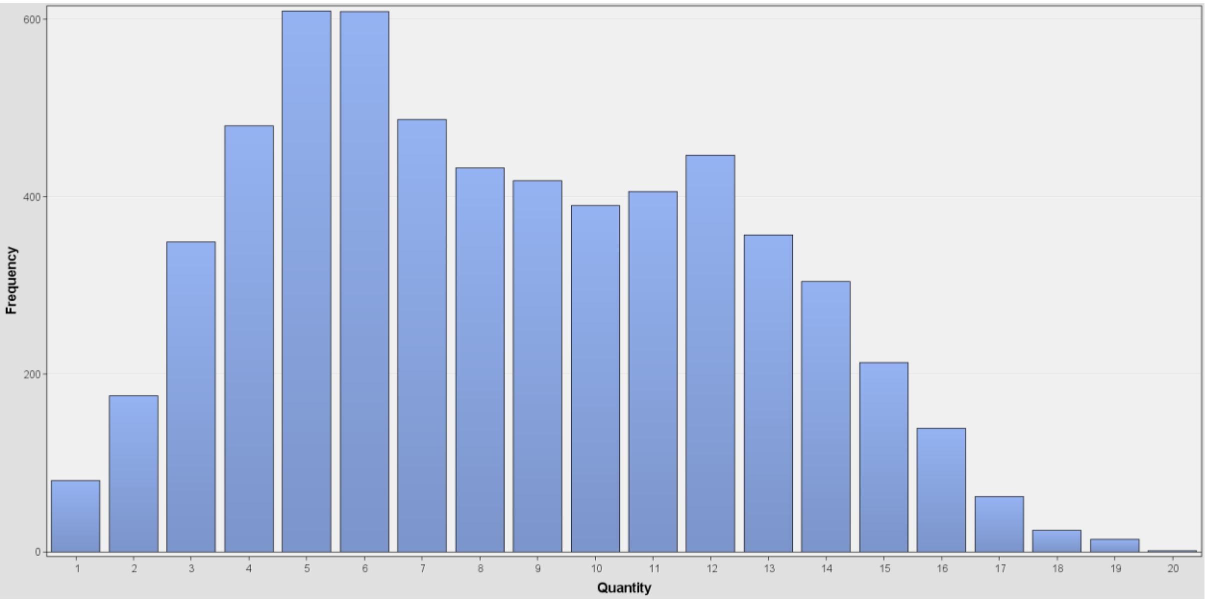


Figure 26

Figure 27 on the other hand shows the histogram graph distribution of frequency against the month where between month of 2.1 and 3.2 the frequency is the highest and it shows that in this month GBI bike make the most profit selling the bicycles to the customers. The lowest frequency of month is between 6.5 and 7.6

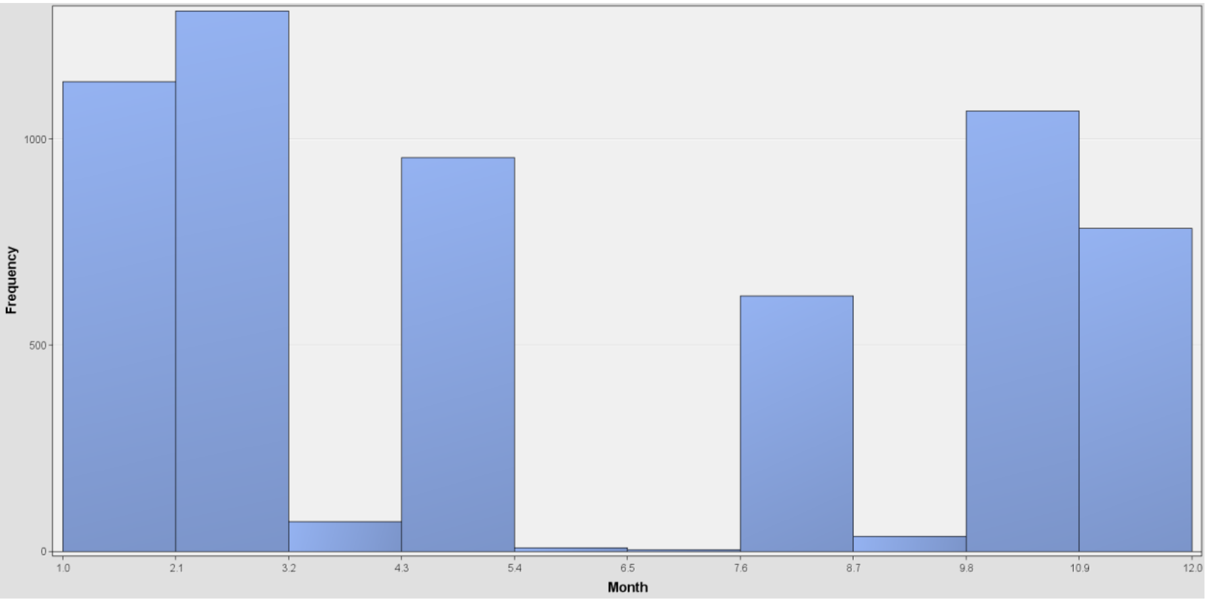


Figure 27

Figure 28 shows the boxplot of cost of goods sold in USD where the minimum whisker value for this input variable is 0, first quartile at 180, median at 2140, third quartile at 6870, maximum whisker at 16800 and mean at 3684.59

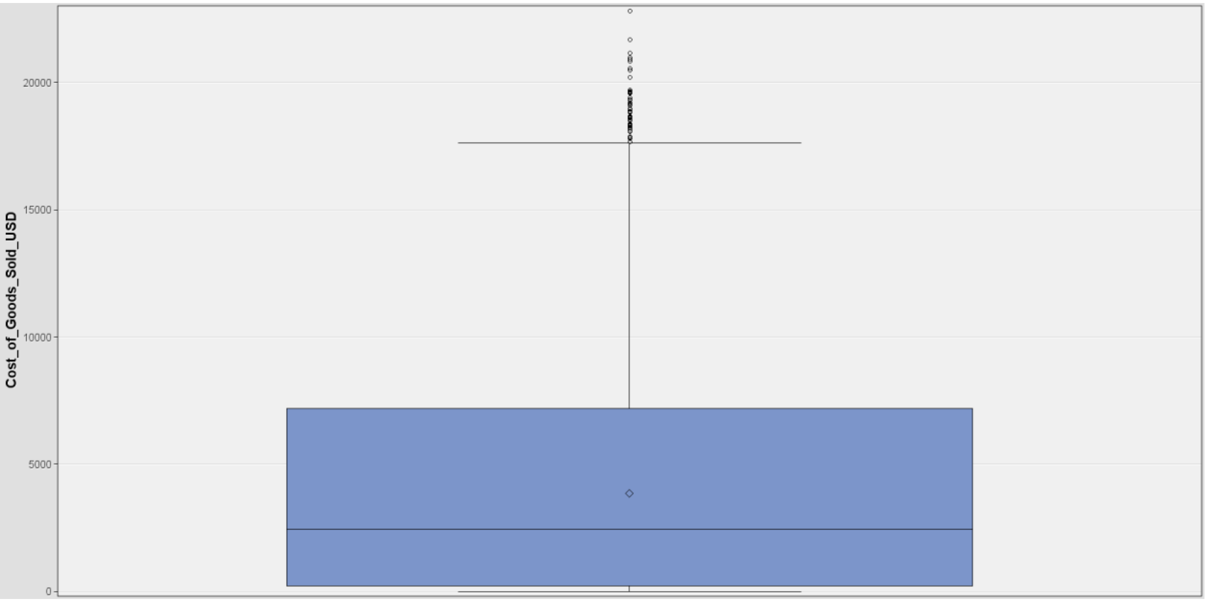


Figure 28

Figure 29 shows the graph distribution for the target variable the revenue USD whereby the highest revenue occurs between the value of 66 and 3418.2

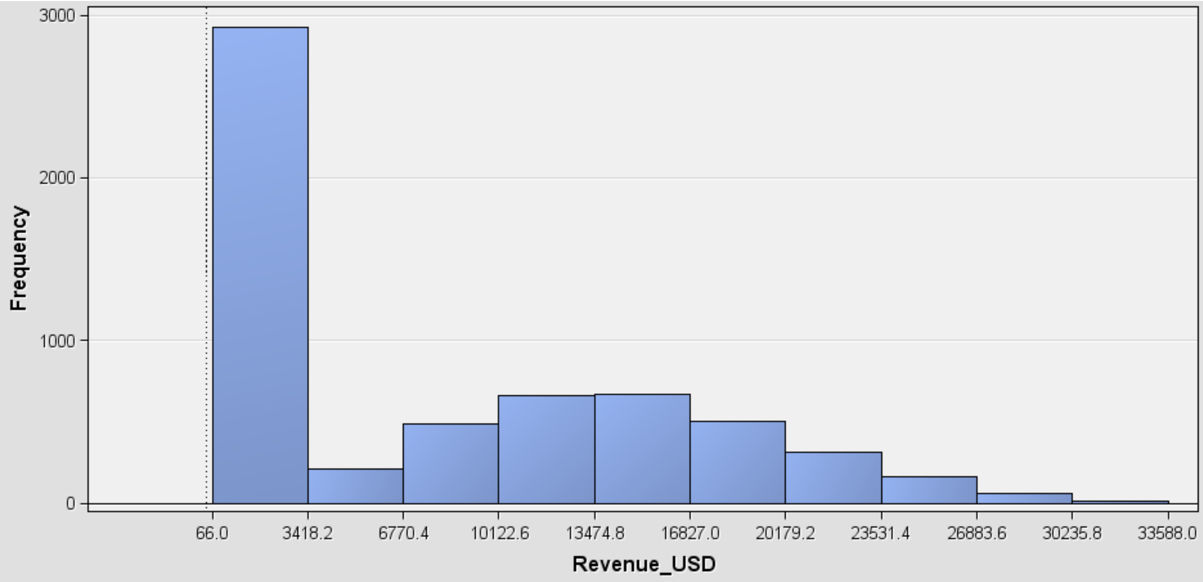


Figure 29

4.2.0 Model evaluation, validation, and optimization

There are two models being used to study the impact of the other variables on the dependent variables which in this case the revenue USD where the role is set as rejected and for the machine learning model used is decision tree and linear regression. It is known that the target variable of the revenue USD is numerical data and thus the suitable machine learning to be deployed is decision tree and regression. In this case the suitable regression is linear regression whereby the target variable for linear regression is numerical value.

4.2.1 Decision tree

Decision tree interpretation is that for an instance with this algorithm the trees learn from the data to approximate the sine curve with set of if then-else decision rules. Thus, as the tree becomes much deeper, the decision rules become much more complex, and the model become much fitter. A decision tree is a supervised machine learning algorithm where it is used for both classification and regression problem whereby decision tree is a hierarchical of tree structure that consist of root node, branches, internal nodes, and leaf nodes.

In this machine learning model, there is three decision tree leaves model proposed whereby each of it has different number of leaves such as 9, 10 and 11. Figure 30 below shows the decision tree with different number of trees that will give different results for the output and after the model comparison the best decision tree will be chosen to build the model.

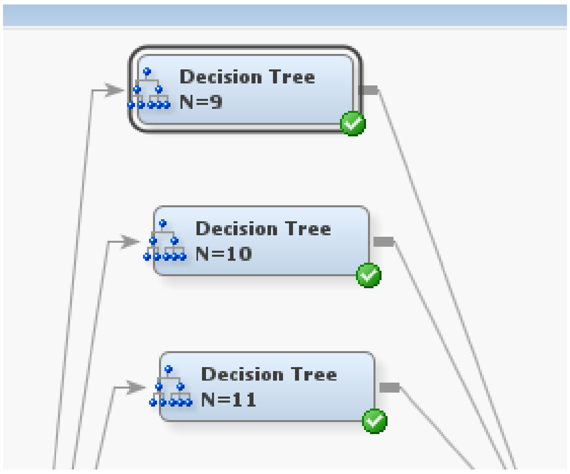


Figure 30

1) Leaf statistics

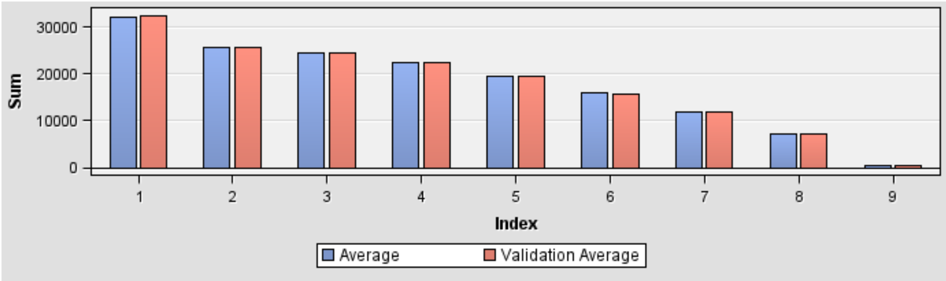


Figure 31 of N=9

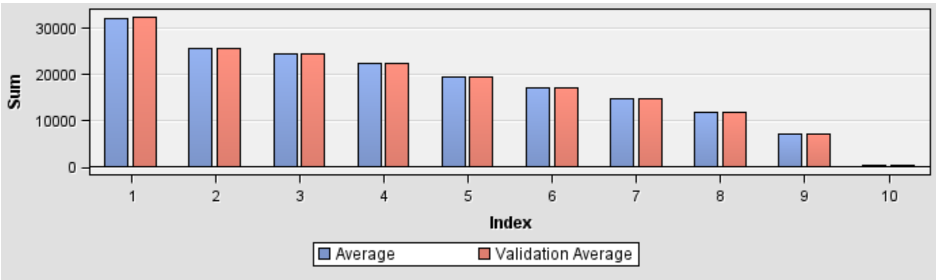


Figure 32 of N=10

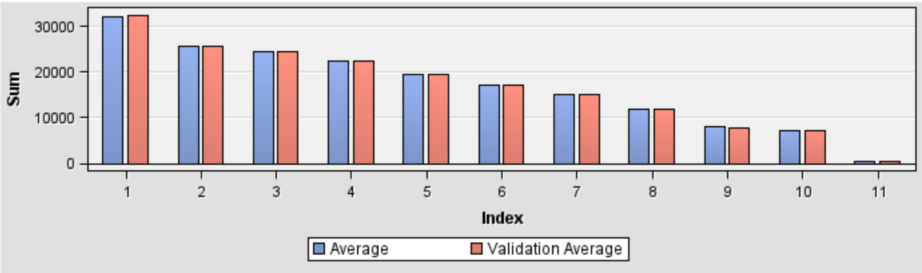


Figure 33 of N=11

Figure 31 until 33 shows the leaf statistics of the decision tree models for the results output for number of threes from 9 until 11 and the leaf statistics is to shows the data distribution by including the extremes values, outliers, median and trends. It is known that if the target variable is numeric the assessment method selected is average square error for the decision tree. On the other hand, the decision tree model with the least value of average square error is N=11 and thus this one is selected for model comparison with the linear regression model as shown below whereby the larger the number the larger the error occurs. In comparison with the higher values of ASE for N=9 and N=10 as shown in figure 35 and 36. The leaf statistics shows the bar graph of validation value and validation average.

90	_ASE_	Average Squared Error	1086212.85	1094796.71

Figure 34 ASE of N=11

87	ASE	Average Squared Error	1331105.31	1483213.49

Figure 35 ASE of N=9

88	_ASE_	Average Squared Error	1184940.37	1296124.33

Figure 36 ASE of N=10

Regarding the decision tree, it can be used for both classification and regression problem and task whereby it is mainly to solve classification problem. Decision tree can be shown as in figure 37 whereby the internal nodes represent the features of the datasets, branches represent the decision rules and each of the leaf nodes represent the outcome.

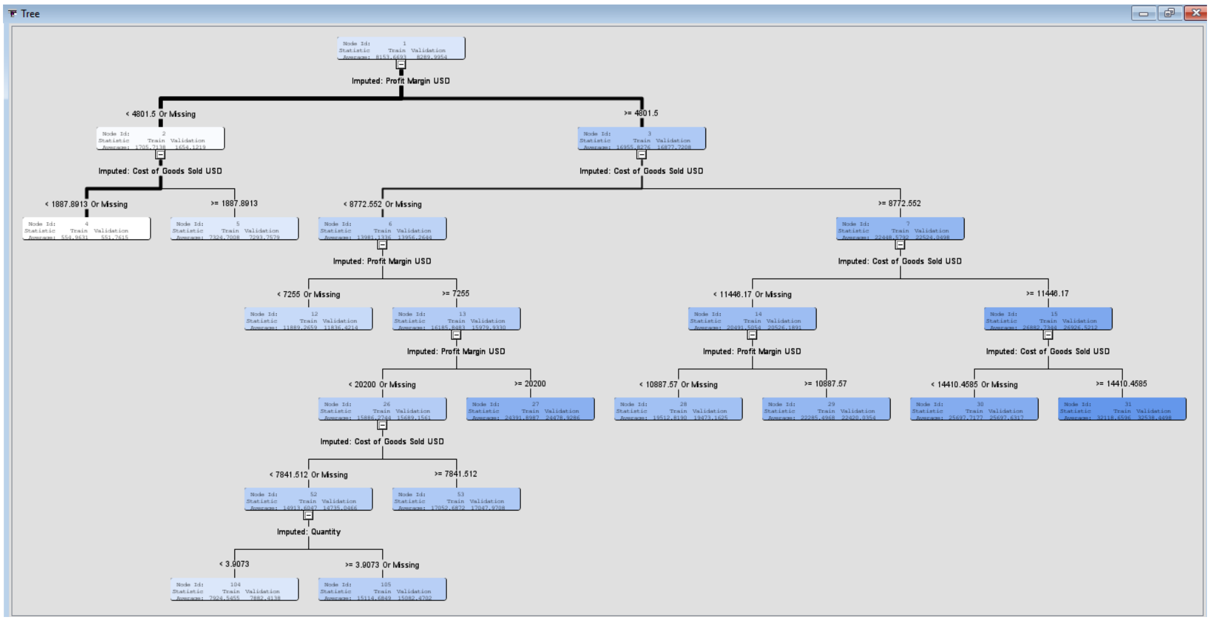


Figure 37 Decision tree of N=11

Decision tree has two nodes which are the decision node and leaf node where decision nodes are used to make any decision and contain many branches, whereby leaf nodes are the output of the decisions and does not contain further branches. It can be said that it is a graphical visualization and representation of providing all the possible solutions to a problem based on certain conditions. The main purpose of decision tree is that it can mimic the ability of human like thinking when making decision and logic behind decision tree is that it is like a tree structure. Below is the decision tree terminology whereby decision tree works by predict the class of the input datasets where the algorithm starts from the root node of the tree. Attribute selection measure on other hand is the best attribute for root node and sub nodes whereby the most common one are information gain and Gini index.

Table 2

Terminology	Explanation
Root node	Where the decision trees started
Leaf node	Final output of the node
Splitting	Split the root node into sub nodes
Branch/Sub tree	Tree formed after split the tree
Pruning	Remove unwanted branches from tree
Parent/Child node	Root node is parent node and other nodes are child nodes.

Variable Importance   Validation

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_Unit_Cost_at_Goods_Issue_USD	Imputed: Unit C...	0	0.0000	0.0000	.
IMP_Sales_Area_Description	Imputed: Sales ...	0	0.0000	0.0000	.
IMP_Quantity	Imputed: Quantity	1	0.0411	0.0585	1.4248
IMP_Profit_Margin_USD	Imputed: Profit M...	4	1.0000	1.0000	1.0000
IMP_Price_USD	Imputed: Price U...	0	0.0000	0.0000	.
IMP_Material_Master_Description	Imputed: Materia...	0	0.0000	0.0000	.
IMP_Material_Group_Description	Imputed: Materia...	0	0.0000	0.0000	.
IMP_Division_Description	Imputed: Divisio...	0	0.0000	0.0000	.
IMP_Cost_of_Goods_Sold_USD	Imputed: Cost of...	5	0.4595	0.4613	1.0040

Figure 38 of the Variable importance

Figure 38 shows the variable importance of the N=11 whereby the variable importance indicates the amount of information from certain variable used by the model which in this case the decision tree. The variable become higher importance if the model relies more on the variables and Gini index, or the other name is mean reduction in impurity mechanism as this one is by default to find the significance of the variable. One of the important measures is that when splitting the variables, the increase of the split criterion for each tree split will add up the whole forest for each tree split.

The significance of a variable is measured by how much information from that variable is "used" by a model. The more a model's reliance on a variable, the higher the importance of that variable. It's useful for a wide variety of models with varying metrics. The mean reduction in impurity mechanism (also known as gini important) is used by default to determine the significance of a given variable. Based on figure 38 the variables importance is given to the quantity, profit margin usd and cost of goods sold usd.

Observations are ranked based on their posterior probabilities or predicted target values. For an interval target, it is the average predicted target value of the top n% observations. The Average Square Error method selects the tree that has the smallest average square error whereby the target variable is numerical and thus average square error method is selected.

Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
IMP_Revenue_USD	Imputed: Revenue USD	_NOBS_	Sum of Frequencies	16797	7199	.
IMP_Revenue_USD	Imputed: Revenue USD	_MAX_	Maximum Absolute Error	12328.1	7676.123	.
IMP_Revenue_USD	Imputed: Revenue USD	_SSE_	Sum of Squared Errors	1.825E10	7.8814E9	.
IMP_Revenue_USD	Imputed: Revenue USD	_ASE_	Average Squared Error	1086213	1094797	.
IMP_Revenue_USD	Imputed: Revenue USD	_RASE_	Root Average Squared Error	1042.215	1046.325	.
IMP_Revenue_USD	Imputed: Revenue USD	_DIV_	Divisor for ASE	16797	7199	.
IMP_Revenue_USD	Imputed: Revenue USD	_DFT_	Total Degrees of Freedom	16797	.	.

Figure 39 of the Fit statistics of N=11

Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_Profit_Margin_USD	Imputed: Profit Margin USD	4	1.0000	1.0000	1.0000
IMP_Cost_of_Goods_Sold_USD	Imputed: Cost of Goods Sold USD	5	0.4595	0.4613	1.0040
IMP_Quantity	Imputed: Quantity	1	0.0411	0.0585	1.4248

Figure 40 of variable importance of N=11

Figure 40 shows the variable importance where again it indicates the number of splitting rules, importance, validation importance and ratio of validation to training importance.

Assessment Score Rankings

Data Role=TRAIN Target Variable=IMP_Revenue_USD Target Label=Imputed: Revenue USD

Depth	Number of Observations	Mean Target	Mean Predicted
5	843	26649.31	26649.31
10	1731	20491.51	20491.51
20	984	17052.69	17052.69
25	1147	15114.68	15114.68
30	2364	11889.27	11889.27
45	1681	7336.48	7336.48
55	8047	554.96	554.96

Figure 41

Figure 41 shows the assessment score rankings whereby for decision trees, one of the prediction types is ranking which the predictive model uses input measurements to optimally rank each case (order).

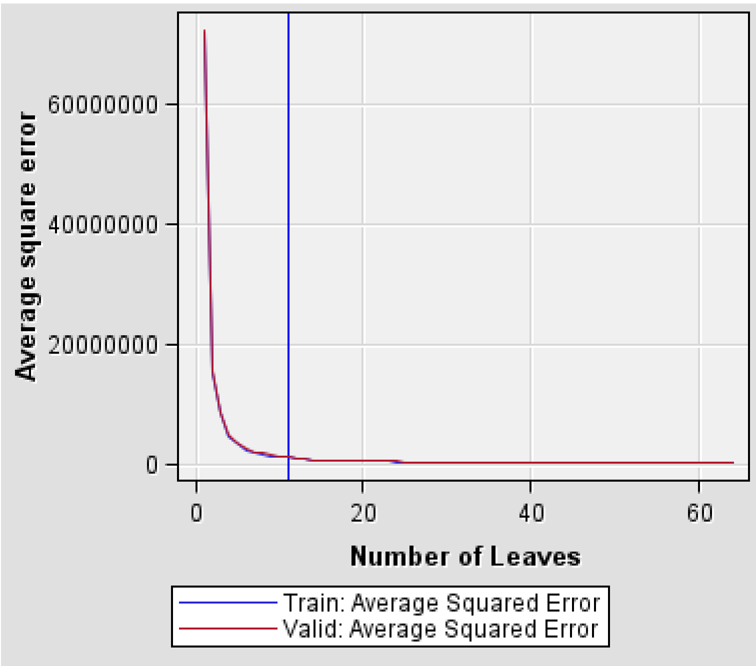


Figure 42

Figure 42 shows the subtree assessment plot for the average square error where it shows the line graph of train average square error and valid average square error.

4.2.2 Linear regression

The second machine learning model selected is linear regression and in regression model it is categorizes into logistic and linear regression. For logistic regression the target variable is categorical binary meanwhile for this analysis linear regression is used because the target variable is numerical. Regression is a different approach than decision trees whereby regressions are parametric models that assume specific association structure between the inputs and target.

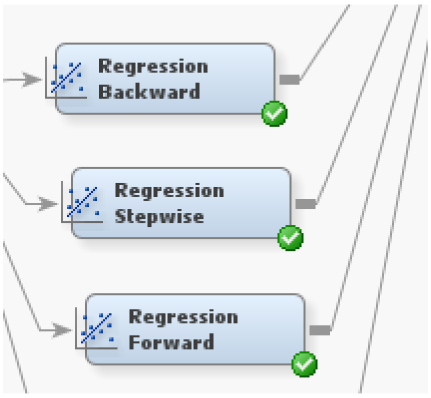


Figure 43

Figure 43 shows the linear regression method where the sequential selection can be categorized into forward, backward, and stepwise selection where the forward method creates the sequence of models of increasing complexity. The backward selection on the other hand, creates sequence of models of decreasing complexity. Stepwise on the other hand, combined the forward and backward method. For linear regression the class target set as linear regression where the link function is logit.

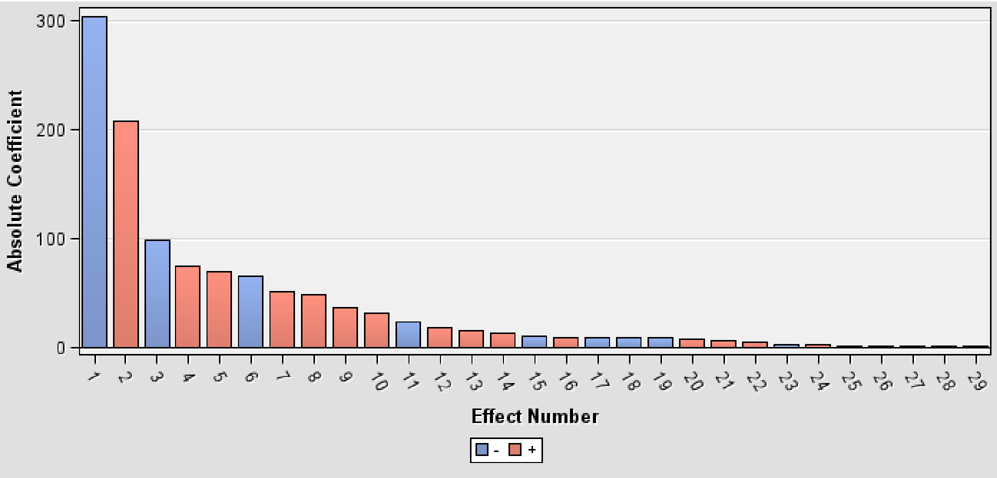


Figure 44

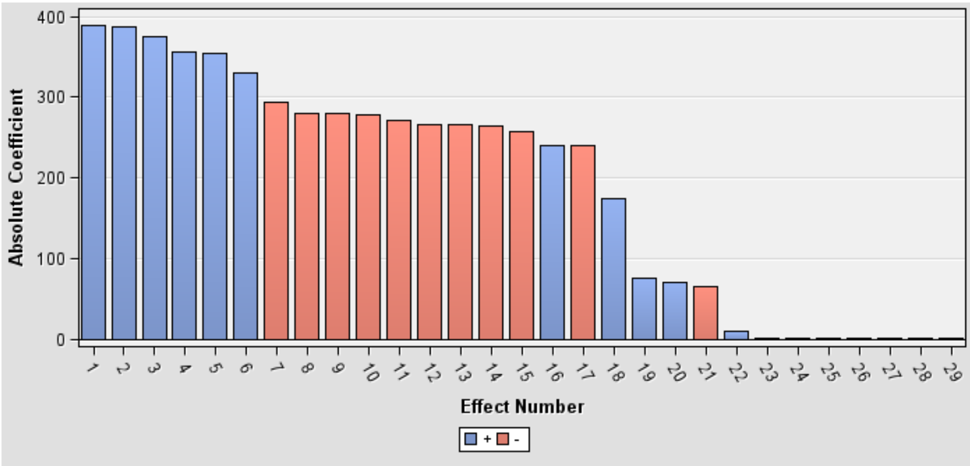


Figure 45

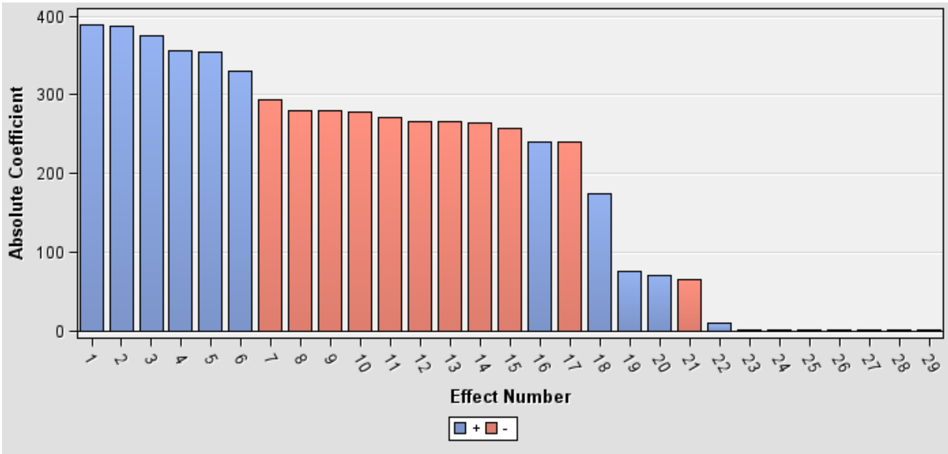


Figure 46

Figure 44,45 and 46 shows the effect plot of regression backward, regression stepwise and regression forward whereby in linear regression the selection criterion used is validation error whereas in logistic regression that has categorical binary target variable the selection criterion used is validation misclassification. Regarding the regression model, the assessment of regression model or the regression complexity can be optimized by choosing the optimal model in the sequential selection sequence. Regarding the regression it can manage the missing values, handle extreme outliers, and use nonnumeric inputs, and managing the missing values can be done using synthetic distribution methods and estimation methods.

The best method to handle extreme outliers or unusual values is to transform or regularize the offending inputs in order to eliminate extreme values where after that a standard regression model can be accurately fit using the transformed input by replaced the original input.

The next part is to show the results output of each of the backward, stepwise and forward regression method where it shows the statistical output of each of the linear regression model.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	28	1.2096977E12	43203490672	794037	<.0001
Error	16768	912345825	54410		
Corrected Total	16796	1.2106101E12			

Model Fit Statistics

R-Square	0.9992	Adj R-Sq	0.9992
AIC	183188.5370	BIC	183190.6373
SBC	183412.6767	C(p)	29.0000

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
IMP_Cost_of_Goods_Sold_USD	1	1.50058E10	275791	<.0001
IMP_Division_Description	1	11611665.9	213.41	<.0001
IMP_Material_Group_Description	0	.	.	.
IMP_Material_Master_Description	16	10662180.0	12.25	<.0001
IMP_Price_USD	1	62042647.3	1140.28	<.0001
IMP_Profit_Margin_USD	1	9848524395	181006	<.0001
IMP_Quantity	1	3505412.76	64.43	<.0001
IMP_Sales_Area_Description	6	8388683.94	25.70	<.0001
IMP_Unit_Cost_at_Goods_Issue_USD	1	240366887	4417.70	<.0001

Figure 47 of backward regression model

Figure 47 and 48 shows output analysis for the backward model where it shows the analysis of variance, model fit statistics, type 3 analysis of effects and analysis of maximum likelihood statistics. There is a slightly different in terms of the statistical output for backward and stepwise whereby for backward it did not show the number of steps as compared to stepwise. The way to analyse the output of the linear regression is based on the assumption below.

$$H_o = The\ model\ is\ not\ significant$$

$$H_1 = The\ model\ is\ significant$$

Thus, if the p-value of the model is (<0.0001) less than 0.05 the model is significant.

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept		1	207.7	20.6529	10.05	<.0001
IMP_Cost_of_Goods_Sold_USD		1	1.1770	0.00224	525.16	<.0001
IMP_Division_Description	Accessories	1	-304.5	20.8436	-14.61	<.0001
IMP_Material_Group_Description	Finished Bikes	0	0	.	.	.
IMP_Material_Master_Description	Air Pump	1	-23.1660	11.7455	-1.97	0.0486
IMP_Material_Master_Description	Deluxe Touring Bike (black)	1	18.0074	11.7764	1.53	0.1263
IMP_Material_Master_Description	Deluxe Touring Bike (red)	1	15.1310	11.8479	1.28	0.2016
IMP_Material_Master_Description	Deluxe Touring Bike (silver)	1	-8.4063	11.7958	-0.71	0.4761
IMP_Material_Master_Description	Elbow Pads	1	-8.0098	11.7135	-0.68	0.4941
IMP_Material_Master_Description	First Aid Kit	1	6.9797	12.7082	0.55	0.5829
IMP_Material_Master_Description	Knee Pads	1	31.3685	11.6067	2.70	0.0069
IMP_Material_Master_Description	Men's Off Road Bike	1	-98.6356	13.4195	-7.35	<.0001
IMP_Material_Master_Description	Off Road Helmet	1	4.4599	11.7238	0.38	0.7036
IMP_Material_Master_Description	Professional Touring Bike (black	1	48.3170	12.0718	4.00	<.0001
IMP_Material_Master_Description	Professional Touring Bike (red)	1	50.8449	12.1907	4.17	<.0001
IMP_Material_Master_Description	Professional Touring Bike (silve	1	36.7618	12.1140	3.03	0.0024
IMP_Material_Master_Description	Repair Kit	1	13.1176	13.3808	0.98	0.3269
IMP_Material_Master_Description	Road Helmet	1	5.3833	11.8153	0.46	0.6487
IMP_Material_Master_Description	T-shirt	1	-9.5292	11.4591	-0.83	0.4057
IMP_Material_Master_Description	Water Bottle	1	-8.3902	11.6881	-0.72	0.4729
IMP_Material_Master_Description	Water Bottle Cage	0	0	.	.	.
IMP_Price_USD		1	0.5437	0.0161	33.77	<.0001
IMP_Profit_Margin_USD		1	0.8623	0.00203	425.45	<.0001
IMP_Quantity		1	8.8875	1.1073	8.03	<.0001
IMP_Sales_Area_Description	Germany North-Wholesale-Accessor	1	1.5023	8.1602	0.18	0.8539
IMP_Sales_Area_Description	Germany North-Wholesale-Bicycles	1	69.7523	9.0061	7.75	<.0001
IMP_Sales_Area_Description	Germany South-Wholesale-Accessor	1	0.5274	8.5171	0.06	0.9506
IMP_Sales_Area_Description	Germany South-Wholesale-Bicycles	1	74.7706	9.0855	8.23	<.0001
IMP_Sales_Area_Description	United States East-Wholesale-Acc	1	-1.0854	6.2348	-0.17	0.8618
IMP_Sales_Area_Description	United States East-Wholesale-Bic	1	-65.7100	8.0007	-8.21	<.0001
IMP_Sales_Area_Description	United States West-Wholesale-Acc	0	0	.	.	.
IMP_Unit_Cost_at_Goods_Issue_USD		1	-1.6257	0.0245	-66.47	<.0001

Figure 48 of backward regression model

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	0	0	.	.	.
Error	16796	1.2106101E12	72077285		
Corrected Total	16796	1.2106101E12			

Model Fit Statistics

R-Square	0.0000	Adj R-Sq	0.0000
AIC	303913.3120	BIC	303911.3165
SBC	303921.0409	C(p)	22233002.556

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	8153.7	65.5063	124.47	<.0001

Figure 49 of stepwise regression model for step 0

Based on figure 49 it shows the step 0 for stepwise regression model and the step carries from step 0 until step 7 as shown below for figure 50.

Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	F Value	Pr > F	Validation Error Rate
1	IMP_Profit_Margin_USD	1	1	250987	<.0001	3.524E10
2	IMP_Cost_of_Goods_Sold_USD	1	2	1033597	<.0001	6.5973E8
3	IMP_Unit_Cost_at_Goods_Issue_USD	1	3	1380.37	<.0001	5.9969E8
4	IMP_Price_USD	1	4	3431.71	<.0001	4.9528E8
5	IMP_Material_Master_Description	17	5	7.62	<.0001	4.922E8
6	IMP_Sales_Area_Description	6	6	25.70	<.0001	4.8742E8
7	IMP_Quantity	1	7	64.43	<.0001	4.8572E8

The selected model, based on the error rate for the validation data, is the model trained in Step 7. It consists of the following effects:

Intercept IMP_Cost_of_Goods_Sold_USD IMP_Material_Master_Description IMP_Price_USD IMP_Profit_Margin_USD IMP_Quantity IMP_Sales_Area_Description IMP_Unit_Cost_at_Goods_Issue_USD

Figure 50

On the other hand, the forward has the same output as stepwise and for forward method it creates a sequence of models of increasing complexity. But typically for the best model selection stepwise linear regression is chosen because it is the combination between forward and backward model.

4.2.4 Model comparison

This part here is to the model comparison between the two-machine learning model proposed which are the decision tree and the linear regression method. Figure 51 below shows the model comparison fit statistics of the machine learning model whereby the main criteria to look for the best machine learning model is based on the average squared error value. Thus, if the model has lesser average square error it is chosen as the best model to forecast the revenue in USD for GBI company as the

Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Average Squared Error	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function
Y	HPTree	HPTree	HP Tree	IMP_Reven...	Imputed: R...	39929.04	.	25702.08	.
	HPReg	HPReg	HP Regres...	IMP_Reven...	Imputed: R...	67470.56	.	54316	.
	Reg	Reg	Regression...	IMP_Reven...	Imputed: R...	67470.56	183188.5	54316	54316
	Reg3	Reg3	Regression...	IMP_Reven...	Imputed: R...	67470.56	183188.5	54316	54316
	Reg2	Reg2	Regression...	IMP_Reven...	Imputed: R...	67470.56	183188.5	54316	54316
	Tree3	Tree3	Decision Tr...	IMP_Reven...	Imputed: R...	1094797	.	1086213	.
	Tree2	Tree2	Decision Tr...	IMP_Reven...	Imputed: R...	1296124	.	1184940	.
	Tree	Tree	Decision Tr...	IMP_Reven...	Imputed: R...	1483213	.	1331105	.

Figure 51 : Fit statistics

In the fit statistics table, there are two high performance model introduced which are HP decision tree and HP regression whereby it shows that when introducing the high-performance model, it outperforms the decision tree with N=11 and stepwise regression. This can be seen in figure 51 above whereby HP decision tree has the lowest average squared error and thus the lower the average squared it has better forecast or prediction on the target variable which is the Revenue USD. Thus, HP decision tree is selected as the best model

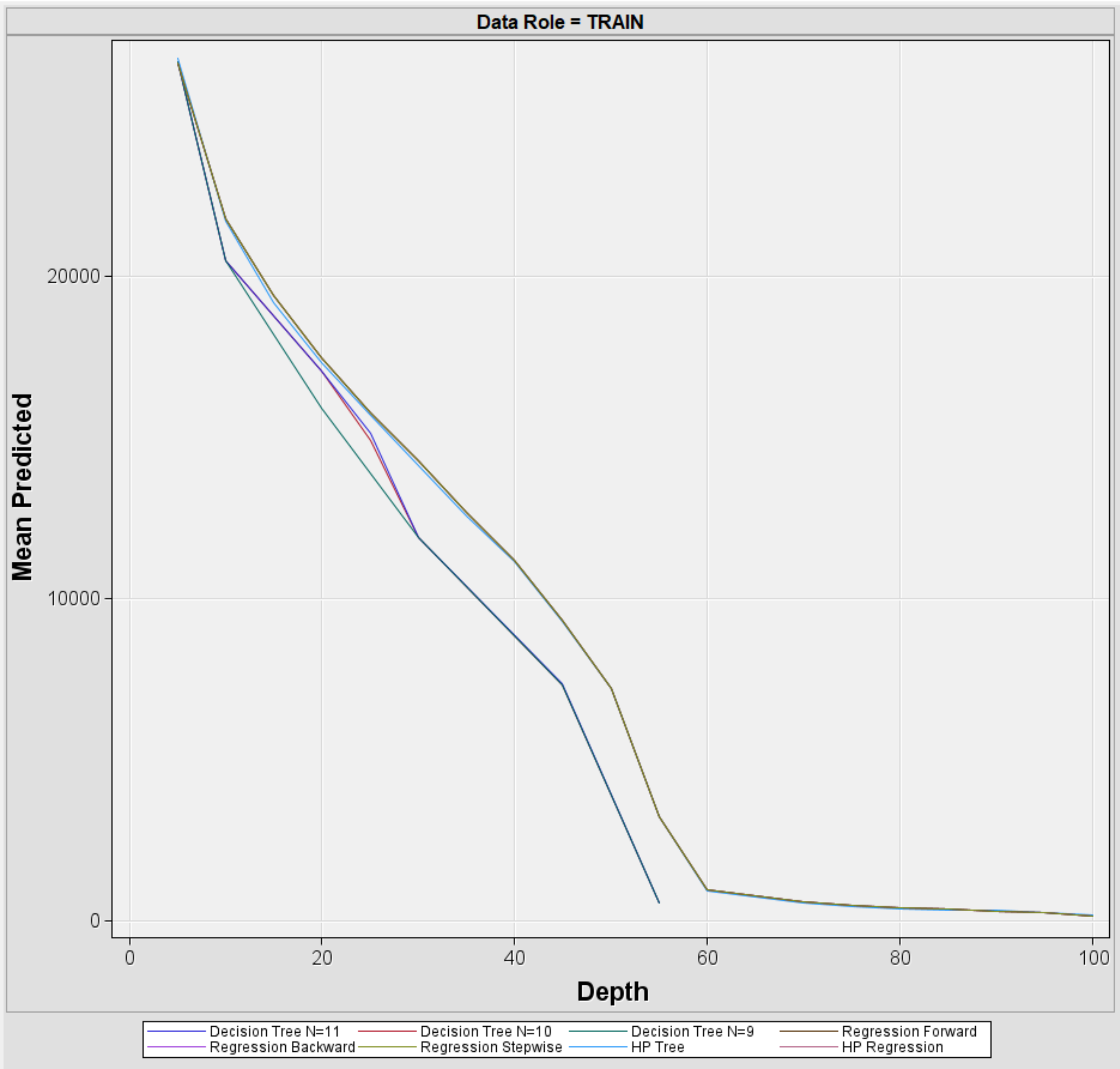


Figure 52 : scores ranking overlay Train

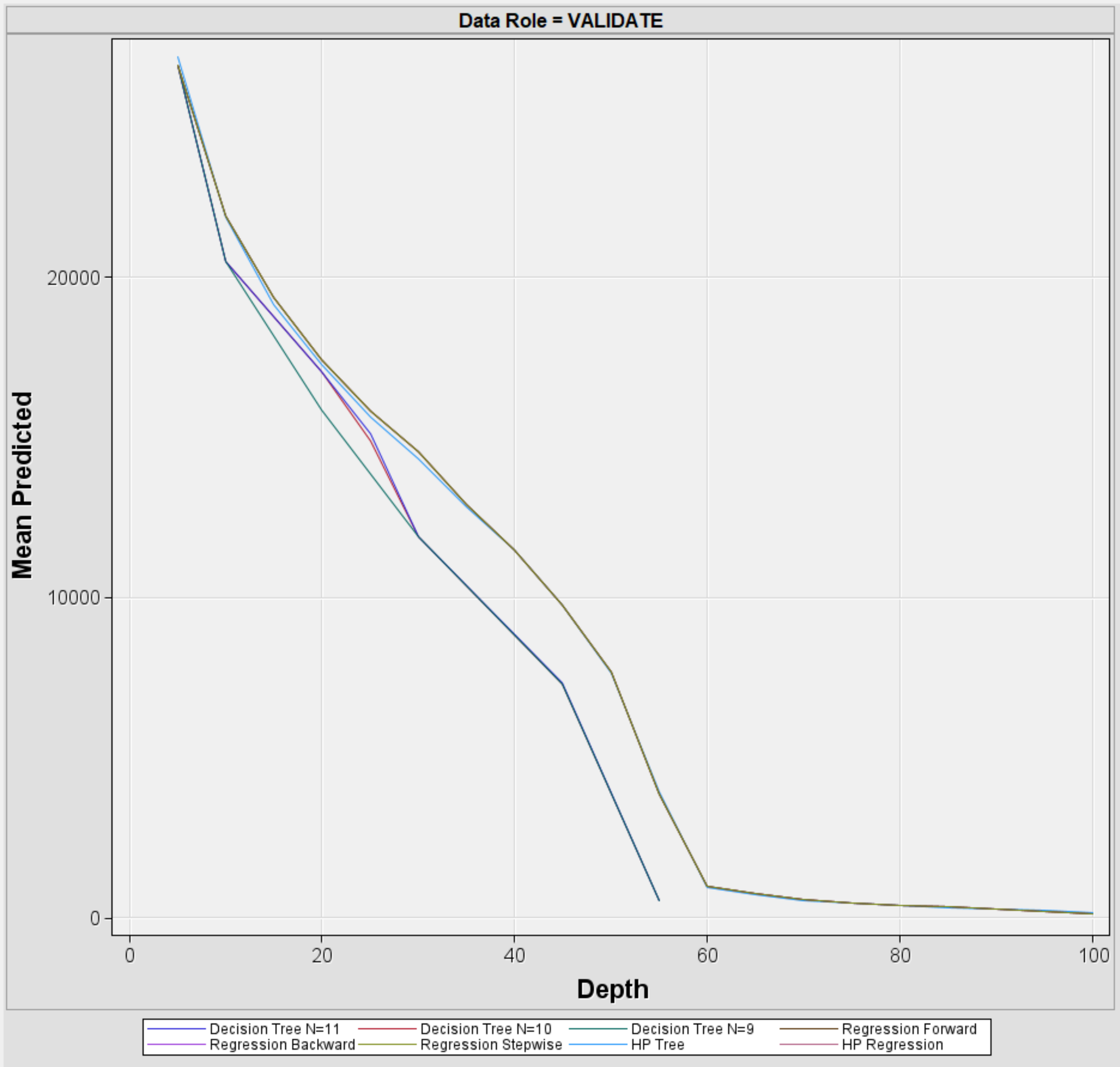


Figure 53 Validate

Figure 52 and 53 shows score rankings overlay for the revenue USD for mean predicted for both train and validate and on the other hand figure 54 shows the score distribution for both high HP regression and decision tree. Figure 55 below shows the score rankings matrix for the revenue USD whereby it again shows the HP regression and decision tree.

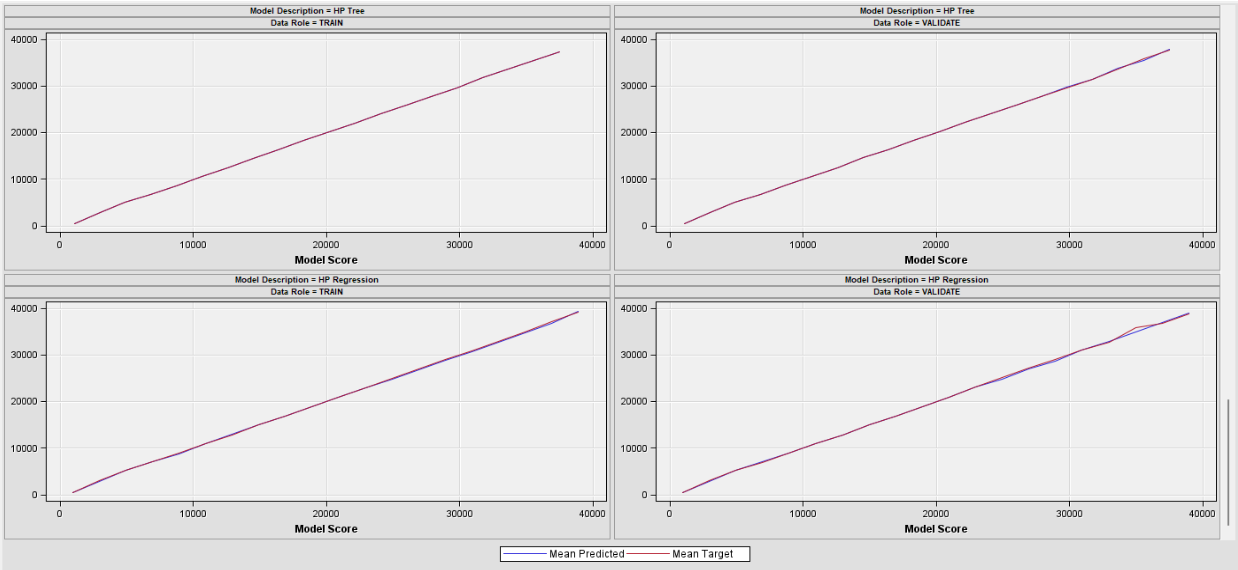


Figure 54 score distribution

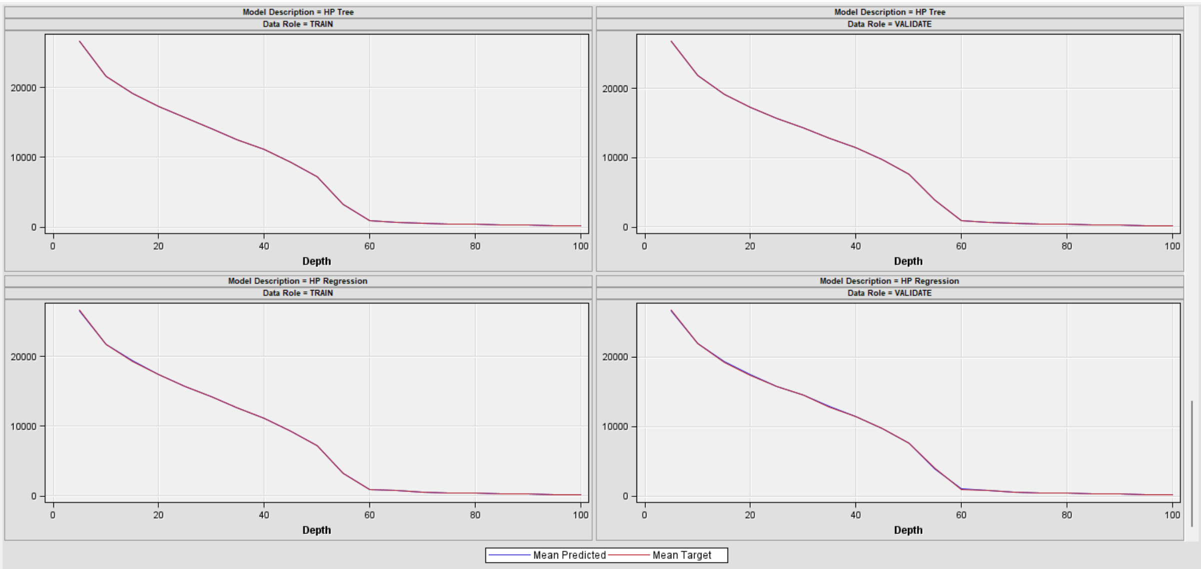


Figure 55 score rankings matrix

Statistics Comparison

Data Role	Target Variable	Target Label	Fit Statistics	Statistics Label	HPTree	HPReg	Reg	Reg3	Reg2	Tree3	Tree2	Tree
Train	IMP_Revenue_U...	Imputed: Revenue USD	_AIC_	Train: Akaike's Information Criterion	.	.	183188.5	183188.5	183188.5	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_ASE_	Train: Average Squared Error	25702.08	54316	54316	54316	54316	1086213	1184940	1331105
Train	IMP_Revenue_U...	Imputed: Revenue USD	_AVERR_	Train: Average Error Function	.	.	54316	54316	54316	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_CRITERION_	Selection Criterion: Valid: Average Squared Error	39929.04	67470.56	67470.56	67470.56	67470.56	1094797	1296124	1483213
Train	IMP_Revenue_U...	Imputed: Revenue USD	_DFE_	Train: Degrees of Freedom for Error	.	.	16768	16768	16768	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_DFM_	Train: Model Degrees of Freedom	.	.	29	29	29	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_DFT_	Train: Total Degrees of Freedom	.	.	16797	16797	16797	16797	16797	16797
Train	IMP_Revenue_U...	Imputed: Revenue USD	_DIV_	Train: Divisor for ASE	16797	16797	16797	16797	16797	16797	16797	16797
Train	IMP_Revenue_U...	Imputed: Revenue USD	_ERR_	Train: Error Function	.	.	9.1235E8	9.1235E8	9.1235E8	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_FPE_	Train: Final Prediction Error	.	.	54503.88	54503.88	54503.88	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_MAX_	Train: Maximum Absolute Error	3060	3161.152	3161.152	3161.152	3161.152	12328.1	12328.1	12328.1
Train	IMP_Revenue_U...	Imputed: Revenue USD	_MSE_	Train: Mean Square Error	.	.	54409.94	54409.94	54409.94	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_NOBS_	Train: Sum of Frequencies	16797	16797	16797	16797	16797	16797	16797	16797
Train	IMP_Revenue_U...	Imputed: Revenue USD	_NW_	Train: Number of Estimate Weights	.	.	29	29	29	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_RASE_	Train: Root Average Sum of Squares	160.3187	233.0579	233.0579	233.0579	233.0579	1042.215	1088.55	1153.735
Train	IMP_Revenue_U...	Imputed: Revenue USD	_RFPE_	Train: Root Final Prediction Error	.	.	233.4607	233.4607	233.4607	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_RMSE_	Train: Root Mean Squared Error	.	.	233.2594	233.2594	233.2594	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_SBC_	Train: Schwarz's Bayesian Criterion	.	.	183412.7	183412.7	183412.7	.	.	.
Train	IMP_Revenue_U...	Imputed: Revenue USD	_SSE_	Train: Sum of Squared Errors	4.3172E8	9.1235E8	9.1235E8	9.1235E8	9.1235E8	1.825E10	1.99E10	2.236E10
Train	IMP_Revenue_U...	Imputed: Revenue USD	_SUMW_	Train: Sum of Case Weights Times Freq	.	.	16797	16797	16797	.	.	.
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VASE_	Valid: Average Squared Error	39929.04	67470.56	67470.56	67470.56	67470.56	1094797	1296124	1483213
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VAVERR_	Valid: Average Error Function	.	.	67470.56	67470.56	67470.56	.	.	.
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VDIV_	Valid: Divisor for VASE	7199	7199	7199	7199	7199	7199	7199	7199
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VERR_	Valid: Error Function	.	.	4.8572E8	4.8572E8	4.8572E8	.	.	.
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VMAX_	Valid: Maximum Absolute Error	4533.09	3124.101	3124.101	3124.101	3124.101	7676.123	9023.605	9996.274
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VMSE_	Valid: Mean Square Error	.	.	67470.56	67470.56	67470.56	.	.	.
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VNOBS_	Valid: Sum of Frequencies	7199	7199	7199	7199	7199	7199	7199	7199
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VRASE_	Valid: Root Average Squared Error	199.8225	259.7509	259.7509	259.7509	259.7509	1046.325	1138.475	1217.873
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VRMSE_	Valid: Root Mean Square Error	.	.	259.7509	259.7509	259.7509	.	.	.
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VSSE_	Valid: Sum of Square Errors	2.8745E8	4.8572E8	4.8572E8	4.8572E8	4.8572E8	7.8814E9	9.3308E9	1.068E10
Valid	IMP_Revenue_U...	Imputed: Revenue USD	_VSUMW_	Valid: Sum of Case Weights Times Freq	.	.	7199	7199	7199	.	.	.

Figure 56

Figure 56 shows the statistics comparison of this model comparison whereby in the statistics label it shows the value train and validation whereby in train it shows the sum of squared errors, average square error, divisor for ASE, maximum absolute error, root average sum of squares, Akaike information criterion, average error function, degrees of freedom for error, model degrees of freedom, total degrees of freedom, error function, final prediction error, mean square error, number of estimate weights, root final prediction error, root mean squared error, schwarz Bayesian criterion and sum of case weights times freq. On the other hand, for valid it shows the sum of squared errors, average squared error, divisor for VASE, sum of frequencies, maximum absolute error, root average squared error, average error function, error function, mean error square, root mean square error and sum of case weights times freq.

4.3.0 Critical interpretations of the results

This part is to discuss the results of the decision tree machine learning modelling, linear regression, and the model comparison.

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	HPTree	HP Tree	39929.04	25702.08
	HPReg	HP Regression	67470.56	54316.00
	Reg	Regression Stepwise	67470.56	54316.00
	Reg3	Regression Forward	67470.56	54316.00
	Reg2	Regression Backward	67470.56	54316.00
	Tree3	Decision Tree N=11	1094796.71	1086212.85
	Tree2	Decision Tree N=10	1296124.33	1184940.37
	Tree	Decision Tree N=9	1483213.49	1331105.31

Figure 57 fit statistics

Figure 57 shows the fit statistics of the different machine learning model proposed for the HP decision tree, HP linear regression, regression forward, stepwise, backward and decision tree for number of leaves of 9,10 and 11. Several observations can be made for example for the regression model for all the three-regression type, stepwise, forward and backward has the same value for valid and train average square error which is 67470.56 and 54316. Among the decision tree, the number of leaves 11 has the lowest average square error at 1094796.71 for valid and 1086212.85 for train.

Figure 58 shows the sample statistics of the variable whereby it shows variables name of each observation, the label, the type, percentage missing, minimum, maximum, mean, number of levels, mode percentage and mode. Based on figure 58, sales order number is the only variables after exploring the variables that has missing values of 5.747% whereas the rest has no percentage missing and typically the variables that has percentage missing more than 30 % is drop from further analysis.

Sample Statistics

Obs #	Variable Name	Label	Type	Percent Missing ▼	Minimum	Maximum	Mean	Number o...	Mode Per...	Mode
8	Sales_Order_Number		CLASS	5.747126	.	.	.	128+	2.298851	4729
2	IMP_Division_Description	Imputed: Division Description	CLASS	0	.	.	.	2	53.75	BICYCLES
3	IMP_Material_Group_Description	Imputed: Material Group Description	CLASS	0	.	.	.	2	53.75	FINISHED BIKES
4	IMP_Material_Master_Description	Imputed: Material Master Description	CLASS	0	.	.	.	18	8.25	PROFESSIONAL TOURING BIKE (BLACK
5	IMP_Sales_Area_Description	Imputed: Sales Area Description	CLASS	0	.	.	.	4	30.55	UNITED STATES WEST-WHOLESALE-...
6	Material_Number		CLASS	0	.	.	.	18	8.25	PRTR1000
7	Quote_Number		CLASS	0	.	.	.	128+	2.484472	20004729
9	Customer	Customer	VAR	0	1000	12000	6479.5	.	.	
10	IMP_Cost_of_Goods_Sold_USD	Imputed: Cost of Goods Sold USD	VAR	0	0	17810	3781.394	.	.	
11	IMP_Price_USD	Imputed: Price USD	VAR	0	14.75	3200	1463.407	.	.	
12	IMP_Profit_Margin_USD	Imputed: Profit Margin USD	VAR	0	32	27000	4201.616	.	.	
13	IMP_Quantity	Imputed: Quantity	VAR	0	1	19	8.417	.	.	
14	IMP_Revenue_USD	Imputed: Revenue USD	VAR	0	66	33403.5	8027.693	.	.	
15	IMP_Unit_Cost_at_Goods_Issue_USD	Imputed: Unit Cost at Goods Issue USD	VAR	0	0	1500	689.4055	.	.	
16	Order_Number		VAR	0	1	7	3.676	.	.	
17	_dataobs_	Observation Number	VAR	0	1	3928	1987.442	.	.	
1	_WARN_	Warnings		.	.	.	.	.	.	

Figure 58 of sample statistics

This part here is to discuss the detail explanation of the graph for both of HP model for decision tree and linear regression from figure 59 until 62.

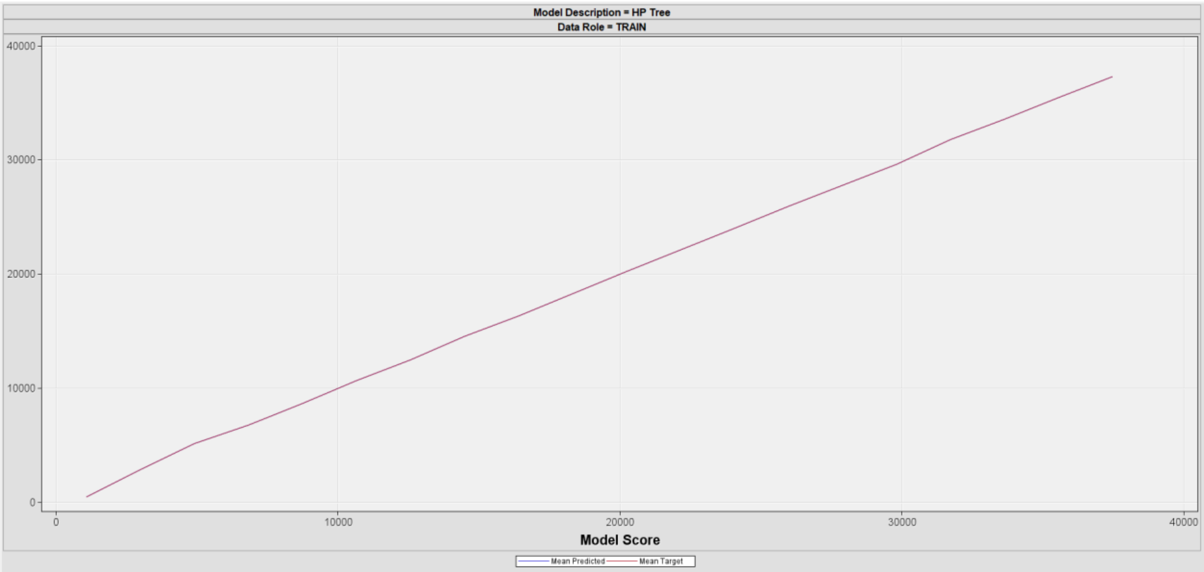


Figure 59

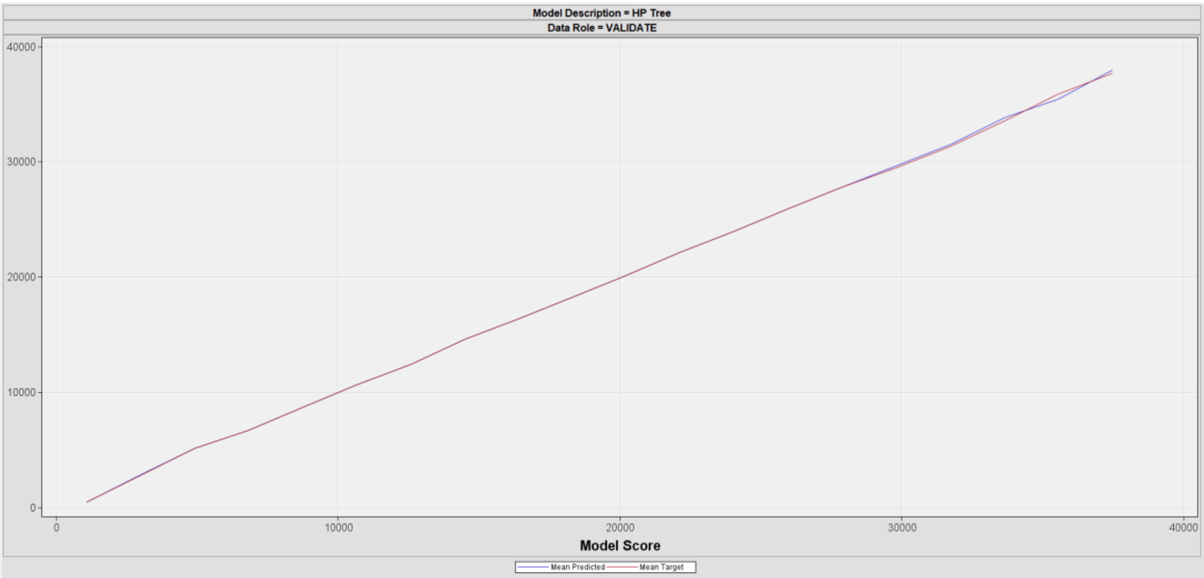


Figure 60

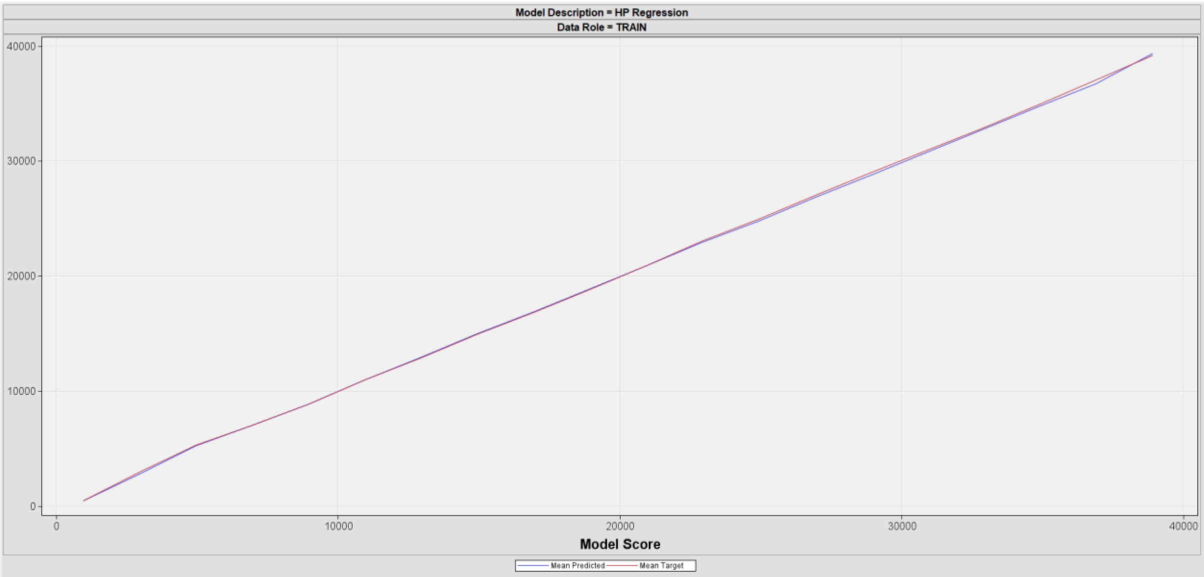


Figure 61

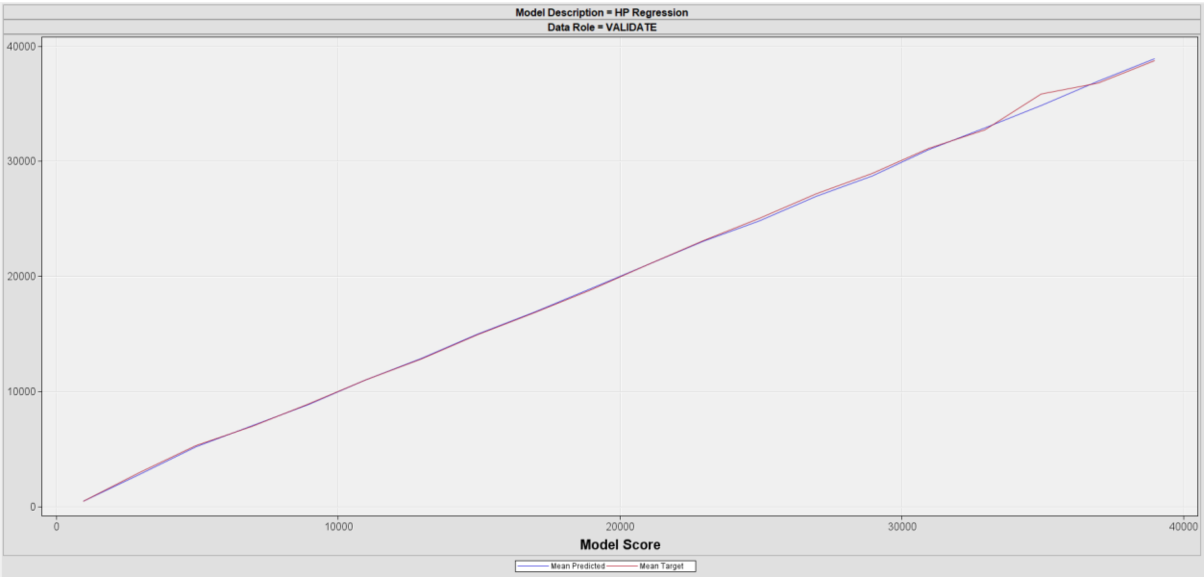


Figure 62

By looking at the graph it can be deduced that machine learning algorithm proposed have the same linear relationship for the graph of the model score except for figure 62 there is some deviation for the validate HP regression on the mean target line graph. Figure 63 until 65 shows the output of the StatExplore whereby it shows the graph of class variation, variable worth and correlation plot (pearson). Based on figure 63, division description and material group description have the same value of percent availability and material master description has the lowest.

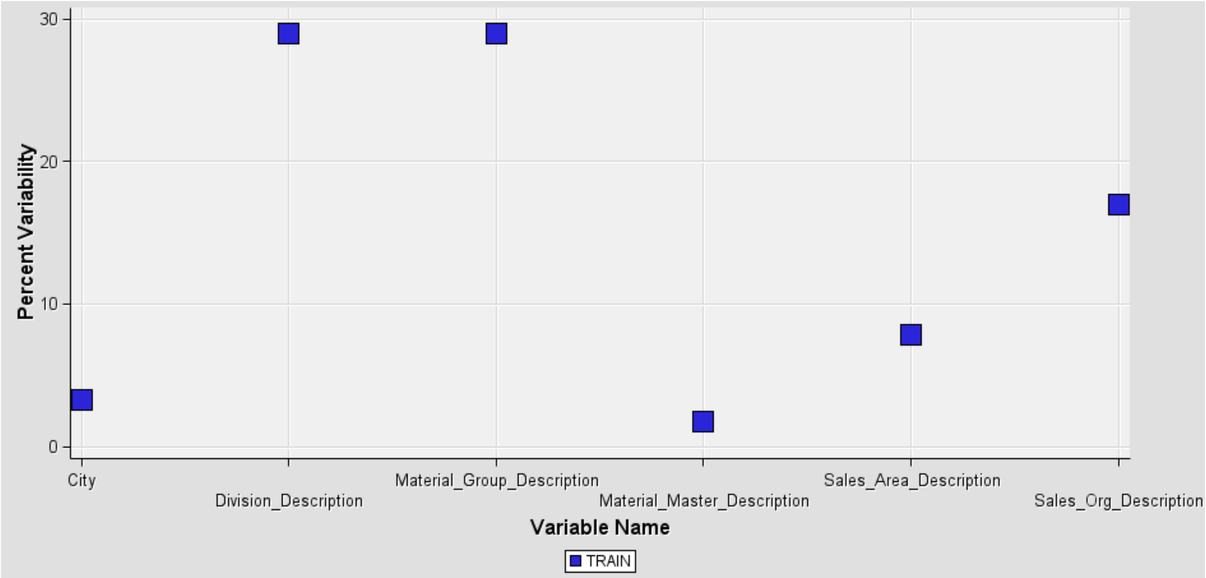


Figure 63 of class variation

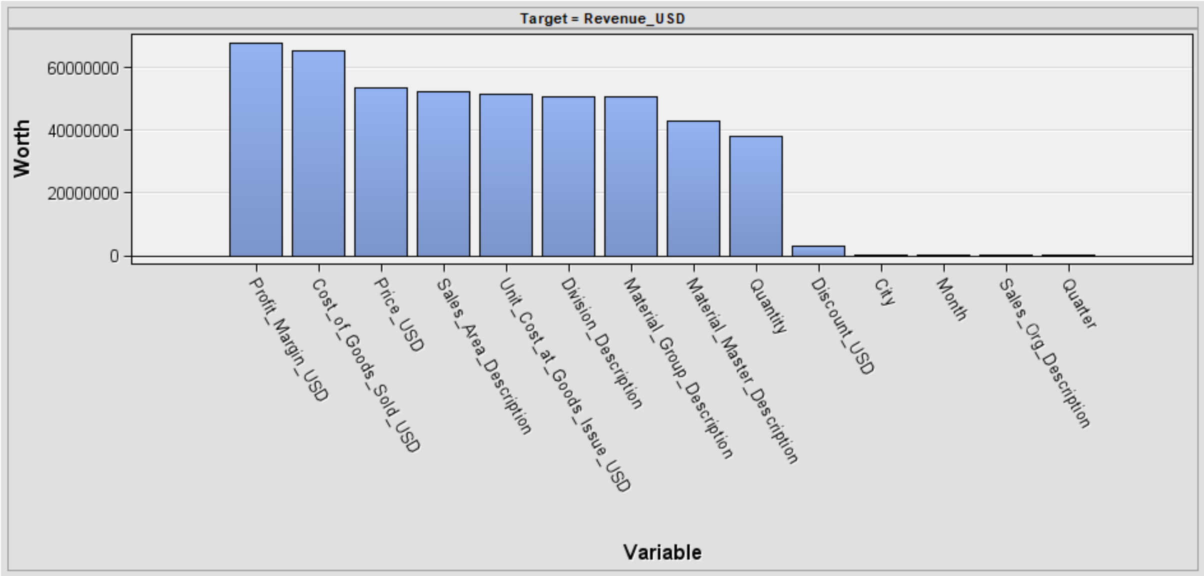


Figure 64 of variable worth

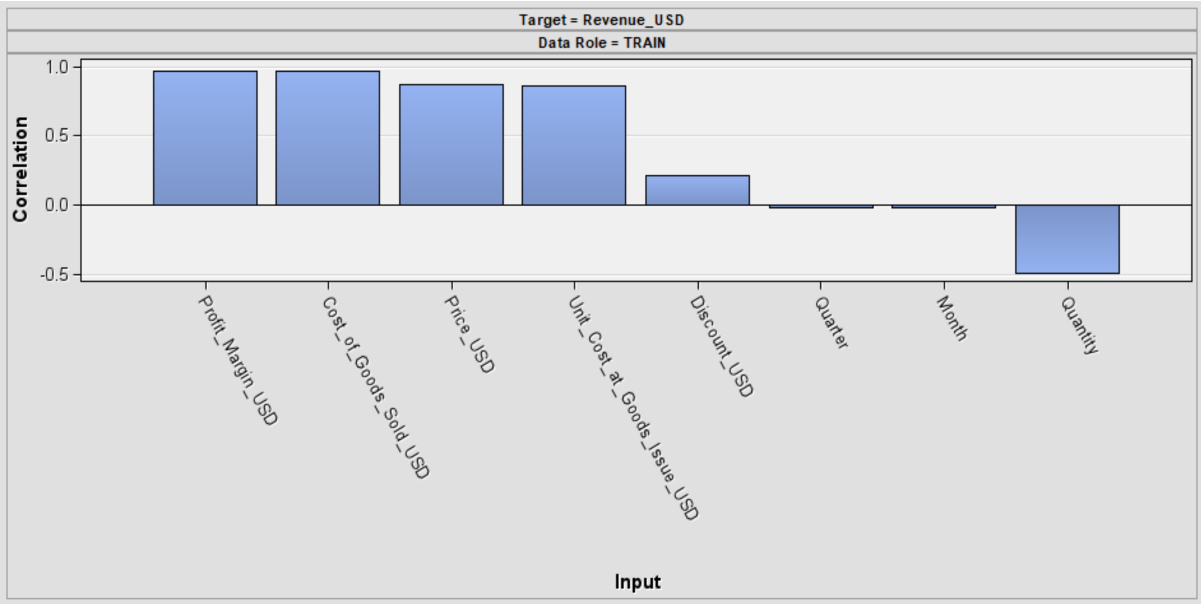


Figure 65 of correlation plot (Pearson)

Class Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City	INPUT	24	26	Denver	11.66	Irvine	11.13
TRAIN	Division_Description	INPUT	3	26	Bicycles	53.62	Accessories	46.27
TRAIN	Material_Group_Description	INPUT	3	26	Finished Bikes	53.62	Safety Gear	46.27
TRAIN	Material_Master_Description	INPUT	19	26	Deluxe Touring Bike (silver)	7.19	Deluxe Touring Bike (red)	7.15
TRAIN	Sales_Area_Description	INPUT	9	26	United States West-Wholesale-Bic	21.36	United States East-Wholesale-Bic	20.84
TRAIN	Sales_Org_Description	INPUT	5	26	United States West	37.94	United States East	37.67

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Cost_of_Goods_Sold_USD	INPUT	3852.574	4117.233	23970	26	0	2400	21675.5	0.723072	-0.57047
Discount_USD	INPUT	27.43238	286.9335	23970	26	0	0	4499.834	10.61177	113.4619
Month	INPUT	6.114101	3.431849	23970	26	1	6	12	0.209316	-1.20354
Price_USD	INPUT	1557.536	1455.043	23970	26	14.75	2165	4495.15	0.054782	-1.70487
Profit_Margin_USD	INPUT	4317.061	4615.49	23970	26	19.764	3030	36720	1.001062	1.079353
Quantity	INPUT	8.159783	3.936244	23970	26	1	8	21	0.278618	-0.88374
Quarter	INPUT	2.383813	1.130017	23970	26	1	2	4	0.191153	-1.35137
Unit_Cost_at_Goods_Issue_USD	INPUT	748.169	701.466	23444	552	0	1095	2132.41	0.039822	-1.7394
Revenue_USD	TARGET	8197.068	8503.316	23970	26	40.626	5890	40498.5	0.675323	-0.67415

Correlation Statistics
(maximum 500 observations printed)

Data Role=TRAIN Type=PEARSON Target=Revenue_USD

Input	Correlation
Profit_Margin_USD	0.96807
Cost_of_Goods_Sold_USD	0.96514
Price_USD	0.86518
Unit_Cost_at_Goods_Issue_USD	0.85689
Discount_USD	0.21430
Quarter	-0.02394
Month	-0.02565
Quantity	-0.49650

Figure 66

Figure 66 shows the summary statistics whereby it shows that the data is not clean as there are some missing values of 26 and 552 and thus imputation or data transformation need to be done to remove the missing values. It also shows the correlation statistics whereby cost of good sold usd has the highest correlation and the lowest correlation is month.

4.4.0 Discussion and conclusion

As the main goal of this analysis is to select the optimum machine learning algorithm that has the best statistical output or the least average square error. As the main goal is to have better accuracy and predicting or forecasting the revenue USD and thus HP decision tree is used as the best model for validation and train. The model assessment starts from the file being imported in the GBI assignment whereby in the file import the GBI datasets is inserted and after data sampling is done, stats explore is used to see the class variation, correlation plot pearson, variable worth and statistical output. Next is the drop variable whereby few variables are dropped from the analysis such as month, sales organization, city, discount usd and quarter and hence after that imputation and data partition. The main goal is to choose the best machine learning model which is in this case HP decision tree and thus this model is the best choice to forecast the revenue USD which is the target variable.

5.0 References

P. J. W. Van Den Noort, "Promoting Cycling for Public Health," L. Use Transp. Res., pp. 105–132, 2016.

T. K. Yunianto, "Tren Gowes Kerek Penjualan Sepeda Hingga 30% Selama Pandemi," Katadata, 2020. https://katadata.co.id/ekarina/berita/5f157dbd397ca/tren-gowes-kerek-penjualan-sepeda-hingga-30-selama-pandemi (accessed May 20, 2021).

F. Pradolo, "Selama Pandemi Covid-19, Permintaan Sepeda Meroket 1.000 Persen," Liputan 6, 2020. https://www.liputan6.com/bisnis/read/4384886/selama-pandemi-covid-19-permintaan-sepeda-meroket-1000-persen

A. B. Tamtomo, "INFOGRAFIK: Jenis-jenis Sepeda dan Tips Membeli Sepeda," kompas.com, 2020. https://www.kompas.com/tren/read/2020/06/20/161659465/infografik-jenis-jenis-sepeda-dan-tips-membeli-sepeda (accessed May 20, 2021).

M. H. Zaki, T. Sayed, and X. Wang, "Computer vision approach for the classification of bike type," J. Adv. Transp., vol. 50, no. 3, pp. 348–362, 2016, doi: http://dx.doi.org/10.1002/atr.1327.

S. Jaya Prada, A. Geetha Sri, B. Venkateswarlu, C. Vineesha, and P. Lakshmi Teja, “Bike Buyer Prediction,” Int. J. Comput. Eng. Technol., vol. 11, no. 3, pp. 45–51, 2020.

Santos, M & Azevedo, C (2005). Data Mining – Descoberta de Conhecimento em Bases de Dados. FCA Publisher.