Data Analysis and Statistical Modeling on Server High-Speed Interface

MASTER THESIS

Presented to

Technische Hochschule Köln, Gummersbach campus
in the degree program of Automation and IT

Written by:

Chinmay Devaru Hegde
(Matr. No.: 11127498)

First Examiner:	**Prof. Dr-Ing. Michael Bongards** TH Köln University of Applied Sciences **(Institute for Automation and IT)**
Second Examiner:	**Dr. Xiaomin Duan** Hardware Engineer IBM **(IBM)**

Gummersbach, August 2020

Abstract

Statistics is the building block of Data analysis. Statistics in combination with Data mining leads to the powerful machine learning algorithms. Through the Satistical methods, data mining methods and different machine learning models, the branch of Artificial intelligence is in the verge of achieving various applications like Autonomous driving, automated customer support etc which were improbable few decades ago. This paper presents one such application of Machine learning on the data of IBM mainframes. The IBM mainframes are the driving force of many billion transactions taking place at this point. There is plethora of data running through the system. The high speed interface like PCI express carries data running through the different components. Such data is analysed and a solution is presented in this thesis.

The motivation was to analyse and find the relationship between the variable eye width and other independent variables of PCIe data. Various statistical approaches like F-Test, P-value, ANOVA were applied. Machine learning models like OLS, Random forest and Neural Networks were tested for finding linear and non-linear relationship between variables. Unique Visualizing techniques like TSNE plot was implemented. All these methods pointed that there is no relationship that exists between Eye width and the other variables. Hence the randomness of eye width were analysed based on Central limit theorem, Probability density function and Confidence interval. Automation of data downloading from the data warehouse was achieved by implementing ISDW function. Necessary works and concepts were introduced which laid the foundation for future work.

Keywords: Random Forest, Neural Network, F-test, ANOVA, TSNE plot, P-value, OLS model, Confidence interval

Acknowledgements

I would like to take this opportunity to thank all the people who helped me and guided me to successfully complete this thesis. Foremost, I would like to thank God for being me in the entire time of my strength, peace of mind and good health throughout my Life events.

I would like to show my greatest gratitude and thank to Prof. Michael Bongards for the inputs while working towards my thesis. The support, advices, guidance and valuable suggestions that he provided helped me to successfully complete this study.

I want to thank my guide from IBM Dr. Xiaomin Duan and Dierk Kaller for providing me a continuous support, encouragement and guidance throughout my thesis. Furthermore, I also want to thank Wolfgang Bolz for providing the opportunity to work on this project in IBM.

Finally, I must express my profound gratitude to my parents and friends for supporting me in every phase of my life and for good wishes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The merging of statistics and computer science has been evident for the rising of Data Science. In this century, the data is considered as a great source for creating valuable information and values. The data obtained from any system can be used to develop a better efficient system by statistical analysis of the features. It helps to find and even predict fault in the system due to sensor failure, and also explains the behavior and relationship between features. Hence statistical analysis of data is considered to improvise through a creative and innovative process.

Big Data, Deep Learning, Data Mining are considered as the branches within the data science field. Statistics form the root of the development of data science. The area of data science makes us possible to achieve data prediction, visualization, and statistical analysis of the latest data. This analysis creates more value for the existing system with the added benefits of accelerating a more efficient system.

The modern era of technology, along with Artificial Intelligence(AI), has a significant impact on computer science with smart and intelligent algorithms. This enables the leading technology companies to invent devices that have the capability to revolutionize the technology world. IBM is one of the leading companies that continuously reinvent its mainframe server. This thesis will focus on the analysis on one of the IBM's mainframe servers with respect to high-speed interfaces.

IBM's mainframes are the data servers, which is capable of processing up to 1 trillion web transactions each day with high reliability and security. The fig 1.1 shows the recent IBM Mainframe Z15[57]. At their core, mainframes are high-performance computers with large amounts of memory and processors.

To perform highly complex tasks, it is enabled with powerful processors consisting of high-speed components. These high-speed components are brought together by the high-speed interface.

Figure 1.1: IBM Mainframe Z15 model

[Source:[57]]

The high-speed interface is the interface standard that connects high-speed components. PCIe(peripheral component interconnect express)[58] is one such high-speed interface that is located in the motherboard of mainframes, and the PCIe contains several slots that enables to plugin different components like SSD, GPU to upgrade the functionality of the servers. The attempt to analyze the PCIe data is needed to improve the performance of the upcoming mainframes of IBM. This paper serves for that purpose.

The fig 1.2 shows general PCIe slots.

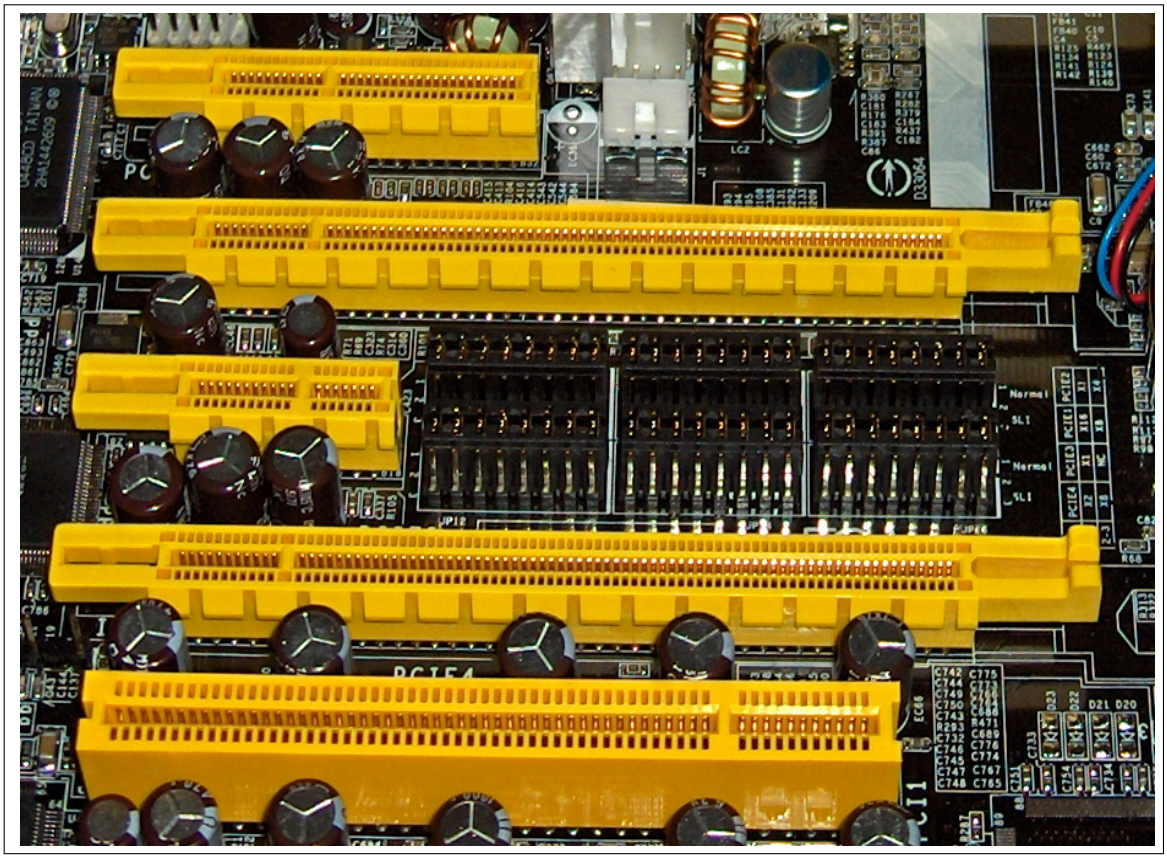

Figure 1.2: PCIe with slots

[Source:[58]]

The important facts to note here are:

1. The mainframe has stored the data relevant to performance
2. The data used in this paper are primarily PCIe metric data
3. The amount of data is large to be manually processed
4. There is a demand for automation of PCIe Data analysis.

1.1 Thesis Goal and Research questions

The fig 1.3 shows the simple server architecture, where drawers are arranged parallelly. A drawer is a container which is the room for the Central processors.

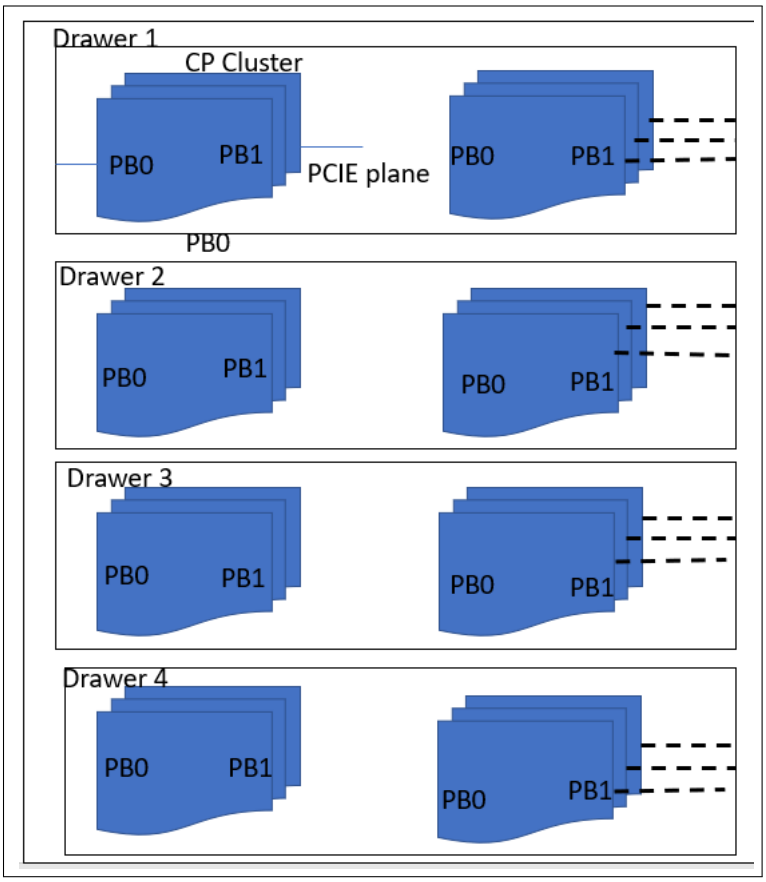

Figure 1.3: CP clusters in drawers

The main component in the drawer is **Central Processor(CP)**. Each CP is associated with an identifier called **Component ID**. Each drawer consists of two CP clusters. A single CP cluster possesses 3 CPs. Each CP has two ports called **Plug Position**. The Plug positions are named as PB0 and PB1. Each Plug position has a PCIe bus connected to it. The PCI express data runs through different paths called **Lanes**. And the PCIe bus consists of 16 lanes to carry the signals. The lanes are numbered from 0 to 15. The wiring of lanes is of different lengths, and it is measured. This parameter is called as **Etch lengths**. There are total 20 slots in the complete server machine where the CP can be fixed. This is called the **Slot ID**. And the numbering of slot id varies from N0CP0 to N7CP2, the number 0 to 7 in Slot ID is associated with which drawer it is belonging. The data we are focusing on here is the data coming into the CP rather than data going out of CP.

The fig 1.4 shows a single drawer and the 6 Cps having two Plug positions PB0 and PB1 each. PB here stands for PCIe Bridge which enables the CP to get connected to PCIe slots

through lanes, as shown by red arrow indication in the figure.

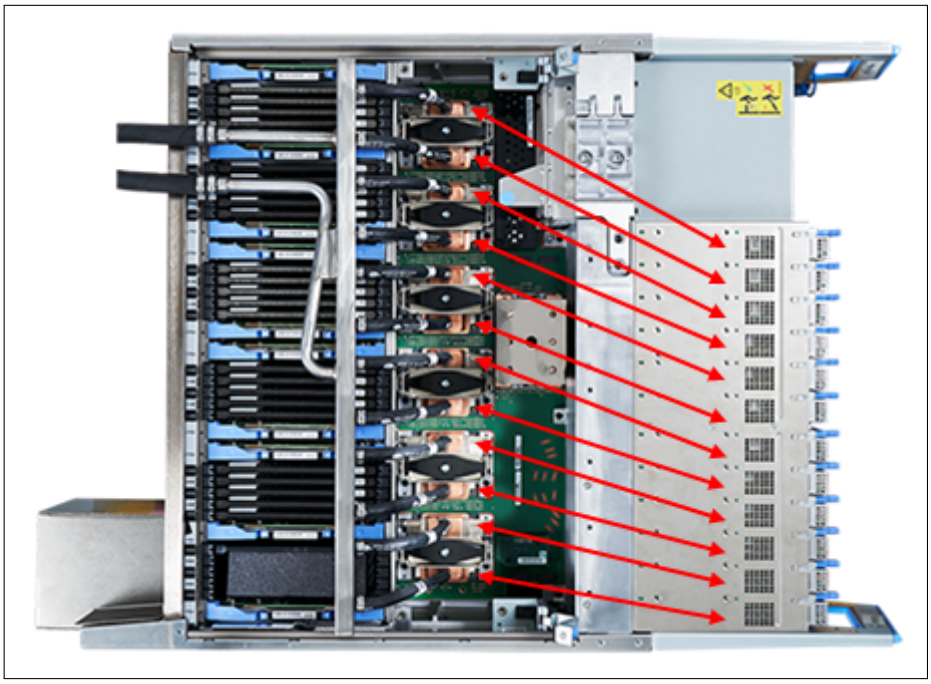

Figure 1.4: A drawer with CPs and PCIe slots

In parallel to the normal data traffic on the high-speed interface, the performance metric such as eye widths are collected and stored. Eye width is the parameter that measures the percentage of opening in the data passage. The higher the value of eye-opening, the lower is the data loss, and hence the machine performance is high. Until now, the machines which are failing to provide the desired eye-opening have been rejected and considered as invalid data input, since it is not guaranteed that the machine is in the right status as the data being collected. That collected data are not all valid since the data collection process is blind and is done without the involvement of domain knowledge. The eye openings are affected by many known and unknown factors. The design of such interfaces has been considered and the known factors are optimized, whereas there is disturbance remaining in the system that are unknown and not easy to be addressed by modeling. It is also desirable to address and quantify these unknown factors.

Manually cleaning the data set is cumbersome. It is then also desired to have an automated cleaning process without human intervention.

The data from the hardware is available to explore the relationship between the parameters. There is a lot of scope in data science to explore in this regard. Questions like, the relationship between various parameters, and the parameters affecting eye width, probability of eye width are open to explore. This work serves as first step to involve data

science in performance analysis of high-speed interfaces in IBM.

The above-mentioned terms like Lanes, Eye width, Slot ID, Component ID, Plug Position and etch lengths are some of the important features to look in the data, and the relationship of them with the eye-opening is needed to be analyzed.

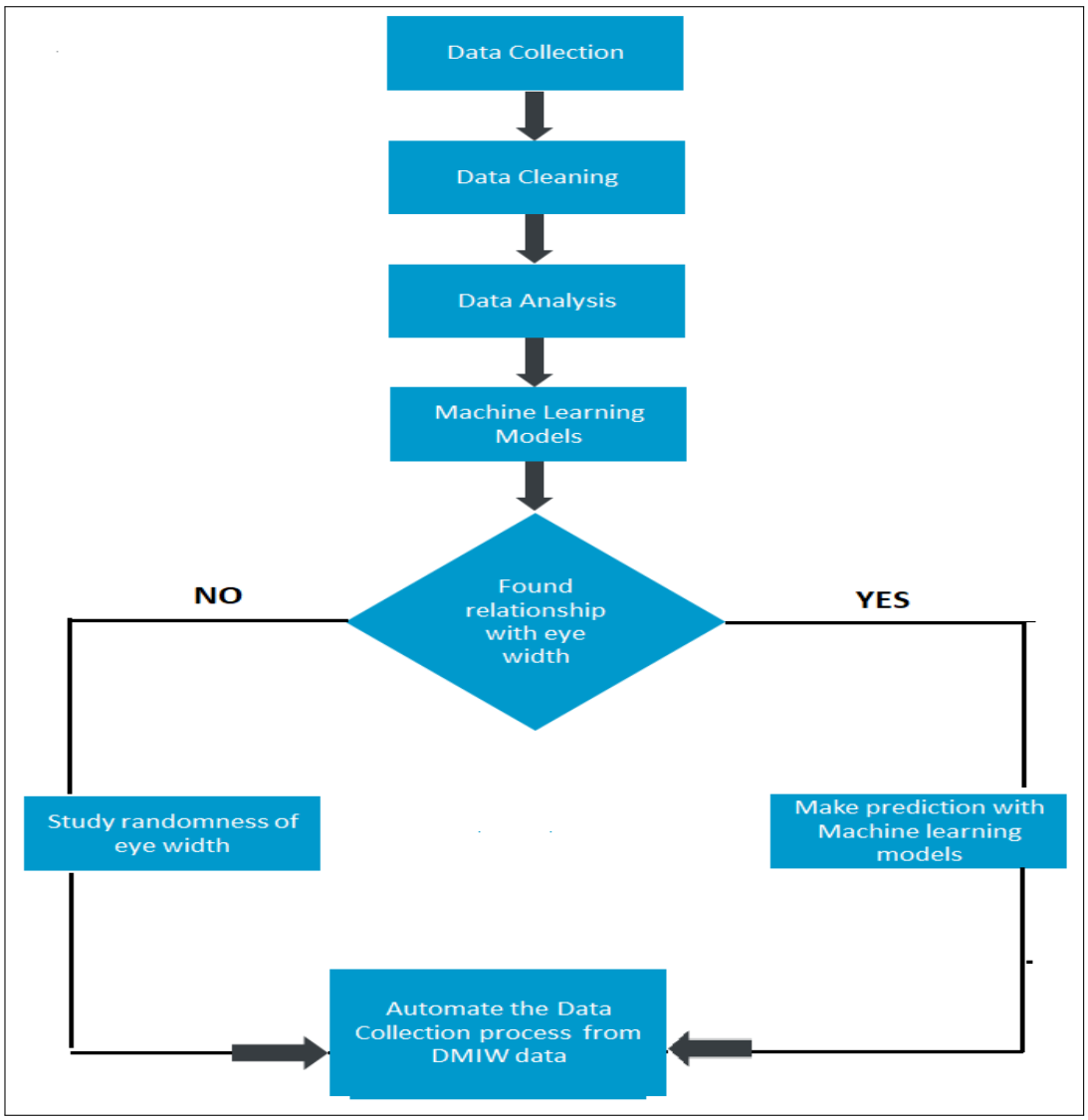

Figure 1.5: Flow chart of thesis

The fig 1.5 shows the flow chart for the solution to the thesis problem. The steps begin from data collection from the PCIe interface, which is stored in the IBM cloud. The next step is to clean the data and make ready to analyze the data. The data is then analyzed with respect to the correlations among the input and output parameters. The goal is also to find out whether a suitable machine learning model can be found. After these steps, if there is a relationship found between eye width and other parameters, then suitable machine learning models are tuned to make eye width prediction. If the eye width is found to be random, then the randomness is studied. At last, the contribution of this thesis has been

consolidated into a python script that automates the steps from data collection to final result plots.

1.2 Outline

This thesis is organized in the following chapters:

1. *Chapter1* defines the introduction, motivation, research questions, goal of the thesis, and structure of the thesis.
2. *Chapter 2* describes the theoretical background of the server high-speed interface and a literature overview of this study.
3. *Chapter 3* presents data preparation, pre-processing steps like missing value analysis, outlier detection, encoding categorical features, feature correlation, and feature importance.
4. *Chapter 4* gives the overall statistical analysis of server high-speed interface. The relationship between the variables and their correlation analysis illustrates the planning, design, and implementation of data modeling with TSNE.
5. *Chapter 5* offers an idea regarding technical details like machine learning models like Random Forest, OLS(Ordinary Least Square), and Neural Network. The model analysis has been deployed with the result of statistical study.
6. *Chapter 6* Explains the Automating of data collection from data warehouse along with normal distribution, Probability density function, and Confidence interval.
7. *Chapter 7* offers a summary of the thesis, concludes with the conclusion, and future work.

Chapter 2

Literature Survey

This chapter gives information about the theoretical literature work done on various machine learning algorithms like Random Forest, OLS and TSNE Plot, Target mean encoding, and the Neural network models.

Breiman introduced the Random Forest with bagging and boosting approach in 1996. In [1], the features of Random Forest are explained with bootstrapping, OOB(out of bag), Variable importance, proximity measure, and many more. The advantage of these features in the random forest is the creation of each tree in the random forest with bootstrap mechanism of data samples which reduces overfitting of the model and creates fewer variance trees.

In [2], Random Forest(RF) is used for the classification of remote sensing. In this work, a comparison of RF with Support Vector Machine (SVM) is made and compared in various circumstances like optimization time, accuracy rate, learning criteria. They found the results were same but the parameters required to tune RF are much easier and lesser than SVM.

Variable importance is one of the vital features in RF. In[3], the variable importance has been utilized in bias samples of bioinformatics datasets. Here vias variable selection have been implemented to build the RF model and the bootstrap mechanism is used to reduce the tradeoff between accuracy and bias of dataset samples.

In [4], the targets to detect the cycle slip in satellite system using Bayesian theory and polynomial regression. Here the discontinuity in the polynomial regression functions is determined using the Bayesian distribution approach.

In [5], a unique regression technique called Evolutionary Polynomial Regression has been introduced. This works with the combination of best features of conventional and genetic programming regression strategies They interpreted the model could handle the data interpolation and could identify noise levels.

Sparse Polynomial Regression(SPORE) methodology has been implemented to predict the computational speed of the computer[6]. Here the detailed prediction has been made between response computational speed with the input features and relatively the error is less than 7% with other machine learning models and the prediction accuracy and

interpretability has been drastically improved with this methodology. Polynomial regression implemented to make data analysis of arbitrary product distributions in [7].

In[8], streamflow forecasting has been implemented with different modeling techniques like Evolutionary polynomial regression(EPR), Multi-layer perceptron neural network (MLPNN), and optimally pruned extreme learning machine(OP-ELM). In the model performance, the peak lows were predicted well with EPR and it outperformed other extreme learning models.

In[9], the Bayesian encoding has been implemented to predict the uncertainty in the brain perceptrons. The computations from the Bayesian optimal, the probability distribution study have been deployed, tested and the results from the sensory uncertainty have been determined.

Neural network estimates the future variables with the memory-based cells using continuous feed-forward loop to fetch information from the previous layers of data. In [10], the study of the general regression neural network(GRNN) has been made where the network predicts with a fast learning rate without iterative process and can be used for interpolation of datasets. GRNN is similar to a probabilistic neural network (PNN) with modification in decision boundaries. GRNN predicts the continuous value-dependent features in the model.

In [15], the neural network has been implemented for the application in face detection. The paper describes with two stages with neural network based filter and merge, overlapping detections in the second stage. This gives the basic implementation knowledge of Neural networks, which could be helpful in future work.

In [11], t-SNE have been implemented along with Tree-Based Algorithms. In the paper, t-SNE used for the visualization and interpretation of multi-dimensional data and used the Barnes-Hut algorithm with a tree-based algorithm to accelerate the t-SNE working. In the result, the conclusion involves that t-SNE had performed better than the dual tree algorithm.

In[12], t-SNE is compared with Sammon mapping, Isomap, and Locally Linear Embedding, and the output from t-SNE is considerably better than other techniques.

In[13] OLS model is used in the Probing interactions using the computational procedures

In[14] The way to detect single influential points has been demonstrated using the OLS model

In[33] The target mean encoding scheme is demonstrated to treat high-cardinality categorical features in classification and prediction problems.

Chapter 3

Data Preparation

This chapter elucidates the data preparation steps involved in the high-speed interface datasets. It also includes the functional aspects of data collection, data distribution, and unit measurements. This further describes the possible data pre-processing steps like data cleaning, feature correlation which explains the relationship between features, missing data treatment, outlier detection check, encoding categorical variables to ease the data visualization process[49] and feature importance which determines the most and least important features in the high-speed bus interface dataset. After the necessary literature survey has been done the first and foremost step in any analysis is the Data collection process.

3.1 Data Collection

Data collection is the first and major step of any data science project. Four sets of PCIe data were collected from the database.

1. The first set of data were manually downloaded from the data warehouse. The first set is comprised of all primary variables which are mentioned in table 3.1.

2. The second set of variables had values of Etch lengths and pin lengths. These values were transferred to the first set of variables based on matching of Slotid, Plugposition, and Lane values from two datasets.

3. The third set of variables contained Psro values and Crpo_Vddn values. Each Component_id has a performance value PSRO and the voltage associated(Crpo_Vddn). So by matching the Component_id values from both the datasets these variables were added to the main dataset.

4. The fourth set of variables were Delay and loss which is obtained based on the length of wires connected (Etch_length6). These variables were added to the main dataset.

3.2 Measurement Units

Measurement Units defines the magnitude of the quantity. The important information on Measurement units are listed below:

1. Each variable is measured with different units.
2. Knowledge of units is an added benefit because there could be a known existing relationship between different units[16].
3. This known relationship could then lead to derive different variables from the existing variables. The part of the original variable might be correlating with the target variable, and this part could be the derived variable from the original variable using the measuring units.

The concepts of Standardization and normalization could be necessary to make the units uniform to a machine learning algorithm[17]. The important facts to note here are:

1. *The standardization process becomes redundant for the tree-based machine learning models due to their architecture*[46].
2. *The models like neural network require standardization*[46].

The table 3.1 shows the measured units of variables used and during which phase the variable was extracted.

Variables	Measurement Units	Phase which the variables are extracted
CORNER	String	First set
COMPONENT_ID	String	First set
DIRECTION	String	First set
COMPONENT_TYPE	String	First set
LANE	Integer	First set
COMPONENT_TYPE	String	First set
PLUG_POSITION	String	First set
TEST_DATE	DATE	First set
EYE_HEIGHT	Pico seconds	First set
EYE_WIDTH	Pico seconds	First set
EYE_FOM	Pico seconds	First set
EYE_HEIGHT_AVG	Pico seconds	First set
EYE_WIDTH_AVG	Pico seconds	First set
EYE_FOM_AVG	Pico seconds	First set
EYE_HEIGHT_MAX	Pico seconds	First set
EYE_WIDTH_MAX	Pico seconds	First set
EYE_FOM_MAX	Pico seconds	First set
EYE_HEIGHT_MIN	Pico seconds	First set
EYE_WIDTH_MIN	Pico seconds	First set
EYE_FOM_MIN	Pico seconds	First set
EYE_WIDTH_SDEV	Pico seconds	First set
EYE_FOM_SDEV	Pico seconds	First set
DATAERROR	String	First set
SLOTID	String	First set
SLOTID_DEC	String	First set
START_TIMESTAMP	DATE	First set
FINISH_TIMESTAMP	DATE	First set
PNP_FILE	String	First set
HOME_DIRECTORY	Pico seconds	First set
Etch_length_X	Millimeter	Second set
Pin_length	Millimeter	Second set
PSRO	milliwatt	Third set
CRPO_VDDN	millivolt	Third set
Delay	millisecond	Fourth set
Loss	db	Fourth set

Table 3.1: Table of variables, measurement units and the phase which the variables belong

3.3 Data description

1. CORNER : This is a variable that defines the used environmentals.
2. COMPONENT_ID : This is the unique identifier for each Central Processor. There are total of 386 in this data set.
3. DIRECTION : This is a variable that indicates whether the data is collected in the CP end or not.
4. COMPONENT_TYPE: This is a variable that indicates the component on which the data was acquired.
5. MACHINE_TYPE: This is a variable that indicates whether the machine belongs to high-end or low-end or Mid-range. Value of '3906' indiactes high-end machine. And '3907' indicates low-end.
6. MACHINE_SERIAL: This is a variable that shows the serial number of the machine.
7. LANE: This is the variable that shows which lane is active in the data transmission. The values range from 0 to 15.
8. PLUGPOSITION: This is the variable that indicates the active port in CP. There are two ports, hence PB0 and PB1 are the two possible values.
9. TEST_DATE: This is the date where the given Component Ids were tested. The date here is the whole month of January 2019.
10. EYE_HEIGHT: Eye height is a performance metric similar to eye width. Eye height along with eye width gives the total eye opening.
11. EYE_WIDTH: Eye width is the reliable performance metric in the given data set. It shows the percentage of eye-opening of a CP. The step size is 3.9 and the range in the given data set is 35.1 to 78.1.
12. EYE_FOM: First order of the Eye width.
13. EYE_HEIGHT_AVG: This is the variable that gives the average of the Eye height for the given CP on the given test date.

14. EYE_WIDTH_AVG: This is the variable that gives the average of the Eye width for the given CP on the given test date.

15. EYE_FOM_AVG: This is the variable that gives the average of the Eye FOM for the given CP on the given test date.

16. EYE_HEIGHT_MAX: This is the variable that gives the maximum value of the Eye height for the given CP on the given test date.

17. EYE_WIDTH_MAX: This is the variable that gives the maximum value of the Eye width for the given CP on the given test date.

18. EYE_FOM_MAX: This is the variable that gives the maximum value of the Eye FOM for the given CP on the given test date.

19. EYE_HEIGHT_MIN: This is the variable that gives the minimum value of the Eye height for the given CP on the given test date.

20. EYE_WIDTH_MIN: This is the variable that gives the minimum value of the Eye width for the given CP on the given test date.

21. EYE_FOM_MIN: This is the variable that gives the minimum value of the eye FOM for the given CP on the given test date.

22. EYE_HEIGHT_SDEV: This is the variable that gives the standard deviation for the eye height for the given CP on the given test date.

23. EYE_FOM_SDEV: This is the variable that gives the standard deviation for the eye FOM for the given CP on the given test date.

24. EYE_WIDTH_SDEV: This is the variable that gives the standard deviation for the eye width for the given CP on the given test date.

25. DATAERROR: This is the variable that indicates whether there is an error in the given data.

26. SLOTID: There are 20 slot ids and this variable shows on which slot the CP has been fixed in the machine.

27. SLOTID_DEC: This is the variable duplicated to SLOTID

28. START_TIMESTAMP: This is the starting date and time where the given Component Ids were tested.

29. FINISH_TIMESTAMP: This is the finishing date and time where the given Component Ids were tested.

30. PNP_FILE: This is the variable that stores the results of the tests for a given test date. It also stores the data on which machine is being tested.

31. HOME_DIRECTORY: This is the variable which gives the location of PNP_File

32. Etch_length1: Etch length of the first component.

33. Etch_length2: Etch length of the second component.

34. Etch_length3: Etch length of the third component.

35. Etch_length4: Etch length of the fourth component.

36. Etch_length5: Etch length of the fifth component.

37. Etch_length6: This is the sum of all previous etch lengths.

38. pin_length: This is the variable which gives the value of the length of pins to which components are fixed.

39. PSRO: This is the variable that indicates the speed of a CP. The lower the PSRO value, the faster the CP. So this value is unique to a given CP.

40. CRPO_VDDN: This is the variable that gives the operating voltage value for a given CP. So this value is unique to a given CP.

41. Delay: This is the variable that gives the value of delay in transmission of data in picoseconds.

42. Loss: This is the variable that gives the value of the amount of data loss in percentage during transmission.

3.4 Data Pre-processing

Data Pre-Processing involves steps of converting raw data into data that is suitable to learn by machine learning models[50]. The current data need to be treated with various techniques and should be evaluated before passing data to the machine learning model. The below methods were used:

1. Data Cleaning
2. Feature Correlation
3. Missing values
4. Outliers
5. Encoding categorical variables
6. Feature Importance

3.4.1 Data Cleaning

The dataset contained about 50% duplicates. In other words, there were 50% data rows identical to each other. This can lead to problems like high bias, memory consumption, redundancy, etc. Hence the duplicates were removed by drop_duplicates command of Pandas library[18].

- 'ROW NO' was deleted since it only gave the information about the row number which does not have anything to do with EYE_WIDTH
- 'CARD' was deleted because throughout the data it had only on value.
- 'DIRECTION' had no values. It was only NaN. So it was removed.
- 'COMPONENT_TYPE' had only a single value
- 'MACHINE_SERIAL' had only NaN values.
- 'DATAERROR' had only NaN.
- 'EYE_FOM','EYE_FOM_MIN','EYE_FOM_AVG','EYE_FOM_MAX','EYE_FOM_SDEV' had only value '0'

3.4.2 Feature correlation

Feature correlation can also be one of the decisive factors in the quality outcome of the model[19][20]. For example highly correlated variables with the target variable may be helpful in prediction. And also when the independent variables are highly correlated within themselves. This can lead to a condition called 'Multicollinearity'.This can lead to skewed or misleading results. Models like Randomforest do not get affected by multi-collinearity but the linear regression models are affected.

Pearson's correlation: Pearson's Correlation[21][22] is an efficient method to know at what extent there is relationship between two variables. The value ranges from -1 to +1. Positive sign denotes the the variables are directly proportional to each other while the negative sign indicates that the varables are inversely proportional. Fig 3.1 shows the visualization of correlation between the variables through a heat map. The numercial representation of degree of correlation is shown in the table 3.2. Other than Eye_Height variables there is no much notable correlation between Eye_Width and other varables.

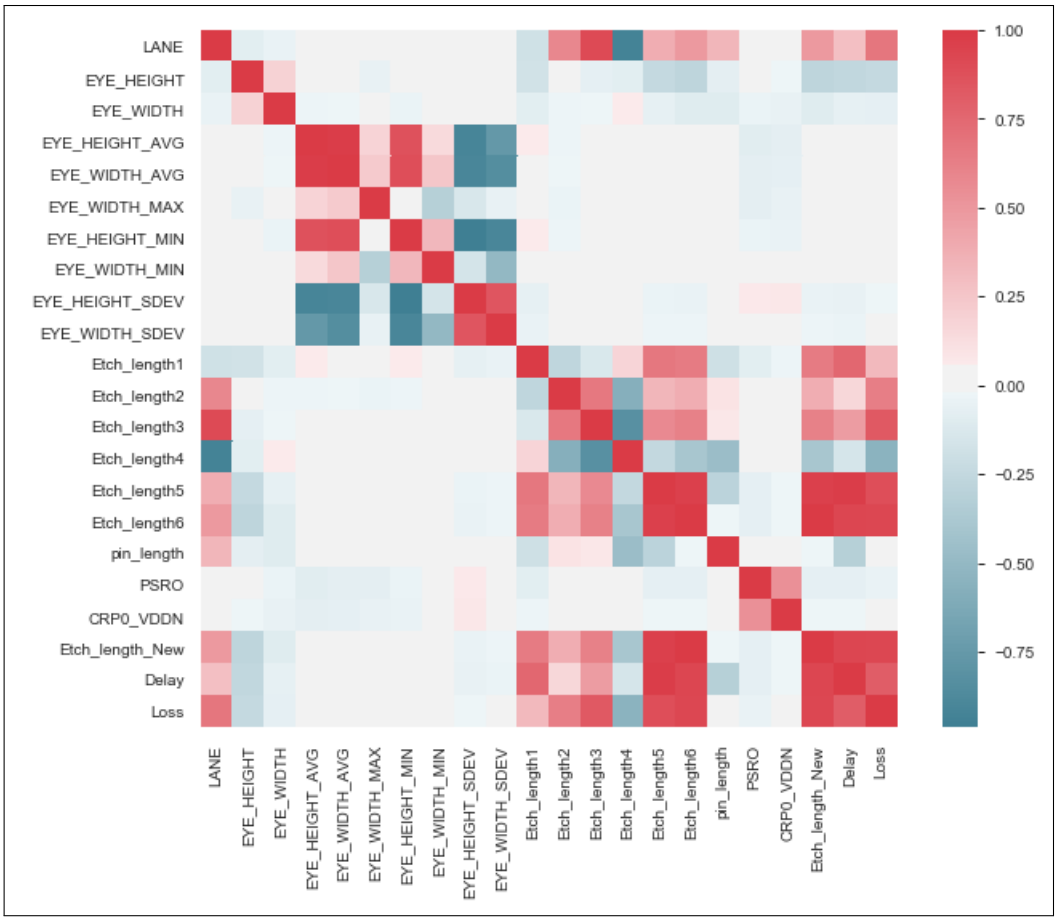

Figure 3.1: Pearson's Correlation of Variables

Variables	Correlation
LANE	-0.01765052618893341
TEST_DATE	-0.17088942608819394
EYE_HEIGHT	0.68465
EYE_HEIGHT_AVG	0.45875508088883005
EYE_WIDTH_AVG	0.4664623808600898
EYE_HEIGHT_MAX	0.4664623808600898
EYE_WIDTH_MAX	0.16174032962267476
EYE_HEIGHT_MIN	0.4109378416859971
EYE_WIDTH_MIN	0.08412028699928331
EYE_HEIGHT_SDEV	-0.4212016066582919
EYE_WIDTH_SDEV	-0.38266414897965595
START_TIMESTAMP	-0.17088942608819394
FINISH_TIMESTAMP	-0.12950163238143786
Etch_length_1	0.025819629213432328
Etch_length_2	-0.013074930128626788
Etch_length_3	-0.015028272950615774
Etch_length_4	0.01993777231046075
Etch_length_5	0.01383346521246604
Etch_length_6	0.0061315675971195865
Pin_length	-0.03387487772550808
PSRO	-0.04
CRPO_VDDN	-0.05
Delay	-0.06
Loss	-0.07

Table 3.2: Numerical interpretation of Pearson's correlation of variables with respect to EYE_WIDTH

Multicollinearity:

Multicollinearity is a term which indicates that there is high correlation between the independent variables[23]. This is not a desirbale scenario because the independent variables should be independent rather than dependent on other variables. A regression coefficient is the amount of change that occurs in the target variable when an independent variable is varied by one unit where as all other independent variables are held uninterrupted. Hence when multicollinearity occurs, determining the coefficient values becomes highly difficult because changing 1 unit in the independent variable changes the value of other correlated variables too. Multicollinearity effects the regression coefficients and p-values[24][25]. Therefore, Linear regression models are effected by this condition. For finding the degree of multicollinearity a technique called Variance Inflation Factor(V.I.F) is used.

Variance Inflation Factor : It is a software calculated value which ranges from 1 to infinity. This value is assigned to each independent varaible of dataset. A value of 1 indicates that there is no direct relationship between that variable and others. A value between 1 to 5 indicates moderate correlation. Values greater than 5 is an evidence that there is high multicolinearity and hence there could be wrong estimation of p value and coefficient value for that independent variable[26][27]. The table 3.3 shows VIF values for the features in the data. The table concludes that the data is severely affected by multicollinearity since the V.I.F values for all variables are above 5. Hence models like Random forest and Neural networks which do not get affected by multicollinearity is more suitable for this condition[60].

Variables	VIF
LANE	8.317600e+03
TEST_DATE	1.541854e+05
EYE_HEIGHT	2.315000e+02
EYE_HEIGHT_AVG	2.153490e+04
EYE_WIDTH_AVG	2.411227e+05
EYE_WIDTH_MAX	1.495740e+04
EYE_HEIGHT_MIN	1.149500e+03
EYE_WIDTH_MIN	4.400000e+00
EYE_HEIGHT_SDEV	2.155900e+03
EYE_WIDTH_SDEV	7.518000e+02
START_TIMESTAMP	9.007199e+15
FINISH_TIMESTAMP	9.007199e+15
Etch_length_1	inf
Etch_length_2	1.801440e+15
Etch_length_3	inf
Etch_length_4	inf
Etch_length_5	9.007199e+15
Etch_length_6	9.007199e+15
Pin_length	4.503600e+15
PSRO	2.088100e+03
CRPO_VDDN	8.922000e+02
Delay	4.895848e+09
Loss	5.690926e+10

Table 3.3: Numerical interpretation of Multicollinearity

3.4.3 Missing data

When the data was tested for missing values, it was found that there was no mising data available because the variables which had NaN values were removed in the data cleaning step[28][29].

Many machine learning algorithms do not operate if there are missing values. And also missing values cause wrong interpretation during analysis. Some of the methods to treat missing values are:

1. Impute missing value with mean or median value.
2. Delete the rows containg missing values.
3. Use any algorithm to predict the missing values.
4. Remove the variable if there are very high number of missing values.

Since when checked for the presence of missing values in the data set, it was found that three variables were only filled by missing values. It is more suitable to remove that variable itself instead of treating them with any other methods, since the other methods of imputation not possible on empty set. The variables which were cleaned are:

- DIRECTION
- MACHINE_SERIAL
- DATAERROR

3.4.4 Outliers

Outliers are the data sets which are outside the certain margin. These observations indicates the quality of the data, the range of the data and the problems in certain data. From the research studies, it is noted that outlier detection and certain measures taken with regards are always better datasets for analysis and predictions[30][31]. The figure 3.2 shows the box plot example for the month of January 2019 on the eye width value. The major outlier is the point indicating the Eye width value equal to 0. This is an improbable value. Hence it is important to remove this value from dataset.

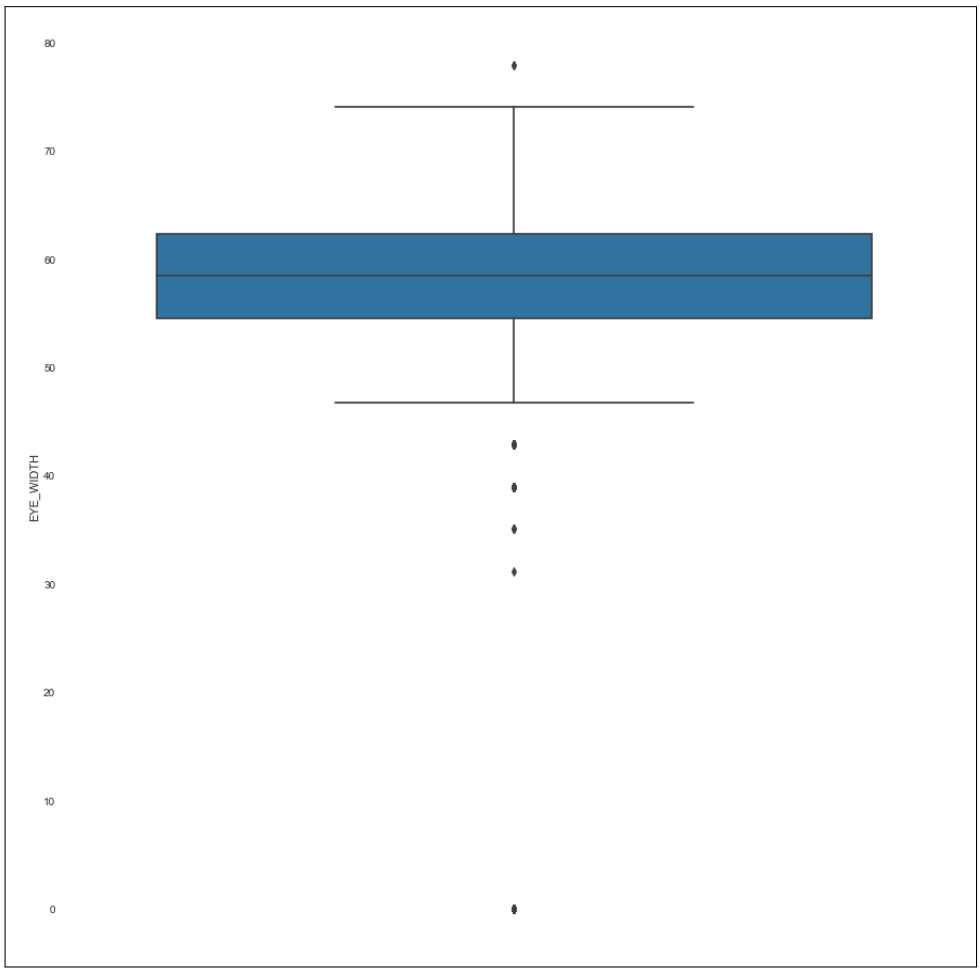

Figure 3.2: Boxplot of Eye width data from January month

3.4.5 Encoding categorical variables

Target Mean Encoding:
This technique is used to convert categorical features into numerical feature[33]. In regression models, the categorical features and numerical cannot be combined to provide input to regression or tree based model. Hence target mean encoding is used to convert category feature into numerical ones.

The benefit of mean encoding over other encoding like label encoding are these features are encoded based on the target values, that leads to fetch more information from datasets. Due to this behavior, encoding may leads to the overfitting problem, hence regularization are used to reduce the overfitting of the datasets[32]. This technique acts on the target feature and the respective category features. The table 3.4 shows an example of working of this technique:

Variable1	Target feature
a	3
a	6
a	12
b	3
b	5

Table 3.4: Example Table before applying target mean encoding

The table 3.4 consists of a categorical variable. There are two values 'a' and 'b'. The target feature is numerical. Now when we observe, the mean value of the target feature associated with category 'a' would be 7. And similarly '4' for 'b'. This values are then substituted to the new table by replacing 'a' by 7 and 'b' by '4' as shown in table 3.5

Variable1	Target feature
7	3
7	6
7	12
4	3
4	5

Table 3.5: Example Table after applying target mean encoding

3.4.6 Feature Importance

Feature importance is the process of extracting the set of features which are important in predicting the dependent varaible. However the variables in our dataset is a mixture of categorical and numerical features. This can be treated separately with respect to Eye_Width. One method for numercial feature selection is F statistic and P value. And ANOVA(Analysis of Variance) can be used for categorical features.

F-Test :
A F-Test is a statistical technique which incorporates the concepts of Null hypothesis and Alternate hypothesis. This helps in finding relevant variables for the model[34][35]. Some of the terms:

Null Hypothesis :
It is a condition which assumes the regression coefficients of the variables of the model is zero[36].

Alternate Hypothesis :
It is a condition which assumes that null hypothesis is wrong and that the model performance is improved by including independent variables[37].

F Statistic :
A value which helps in determining whether null hypothesis is true or Alternate hypothesis[38].

F Value :
It is a value associated with each variable and it helps in identifying its importance in determining model performance[41].

P value :
P value is a probability of the null hypothesis being true. Usually the threshold value is 0.05. If a variable is having less than P value of 0.05, then it means there is less than 5% probability of the variable being not important[39][40].
The F statistic is calculated from OLS model of Statsmodel.api module in Python. The F value and P value is calculated from f_regression from sklearn.feature_selection module in Python. When the F value is more than F statistic and P value is lesser than the threshold of 0.05 for a variable, then it is evident that the variable cannot be discarded from the model. The F statistic was found to be having value of 1504. And the ta-

ble 3.6 shows the values of F value and P value for all numerical variables of dataset. And the table concludes that the variables Eye_height, Eye_height_avg, Eye_height_sdev, Eye_height_min, Eye_width_avg, Eye_width_min, Eye_width_max, Eye_width_sdev were found to be important because the F-value for them were more than 1504 and P value less than 0.05. However these are the variables which co-exist with the Eye_width variable.

Variables	**F value**	**P value**
LANE	6.04024598e+00	1.39925157e-002
TEST_DATE	4.04754465e+02	4.63559011e-089
EYE_HEIGHT	1.59930303e+04	0.00000000e+000
EYE_HEIGHT_AVG	4.82280613e+03	0.00000000e+000
EYE_WIDTH_AVG	5.02889159e+03	0.00000000e+000
EYE_WIDTH_MAX	4.89550352e+02	4.47514734e-107
EYE_HEIGHT_MIN	3.67309900e+03	0.00000000e+000
EYE_WIDTH_MIN	1.29856798e+02	5.55549022e-030
EYE_HEIGHT_SDEV	3.89849666e+03	0.00000000e+000
EYE_WIDTH_SDEV	3.09448175e+03	0.00000000e+000
START_TIMESTAMP	4.04754465e+02	4.63559011e-089
FINISH_TIMESTAMP	2.58251485e+02	1.02253581e-057
Etch_length_1	3.82064150e+00	5.06400170e-002
Etch_length_2	3.54772081e-01	5.51431469e-001
Etch_length_3	4.90940490e+00	2.67230691e-002
Etch_length_4	7.01645543e+00	8.08331080e-003
Etch_length_5	4.21622446e-01	5.16136711e-001
Etch_length_6	3.08526918e-01	5.78592309e-001
Pin_length	2.13281002e+01	3.89643432e-006
PSRO	2.13281002e+01	3.89643432e-006
CRPO_VDDN	2.13281002e+01	3.89643432e-006
Delay	2.13281002e+01	3.89643432e-006
Loss	2.13281002e+01	3.89643432e-006

Table 3.6: Numerical interpretation of F value and P value

Analysis Of Variance(ANOVA) :

One way Analysis of variance[41] is a statistical method used to evaluate the relationship between categorical features and the continuous target variable. It checks whether different groups of categorical feature have equal variance with respect to target variable[42].

If there is equal variance for all groups, then simply it means the feature does not have any impact on target feature. The Table 3.7 shows the Variance value and P value associated with the variable. More the Anova value, more the variance explained by that variable with respect to Eye_Width. Hence that variable can be more important.

Variables	**ANOVA value**	**P value**
CORNER	50.916294667768135	9.189334111614058e-33
COMPONENT_ID	74.24970651387062	0.0
PLUGPOSITION	7.60794288233897	0.00581692702237843
SLOTID	57.341103528921025	7.421620986642383e-202 2.153490e+04
SLOTID_DEC	57.341103528921025	7.421620986642383e-202
PNP_FILE	196.25063804712056	0.0
HOME_DIRECTORY	194.0161425592365	0.0

Table 3.7: Numerical interpretation of ANOVA value and P value

3.5 Summary and Conclusion for the chapter

The following points were concluded from this chapter:

1. Introduction to dependent and independent variables were given along with measurement units and data collection process.

2. Data pre-processing steps like data cleaning, finding missing values and outlier detection were done. About 8 variables were cleaned due to empty values. 3 variables were removed because of only missing values. The value equal to '0' was decided as outlier and was removed from Eye width variable.

3. Pearson's correlation was applied on variables with respect to eye width. Eye height was found to be highly correlated. However since Eye width and Eye height exist together, this finding is not very important.

4. Multicollinearity between variables were found by Variance Inflation factor. The data is severely affected by multicollinearity since the V.I.F values for all variables are above 5

5. Target mean encoding was applied to convert categorical features to numerical.

6. Statistical tests like F-Test, P value and ANOVA were conducted. By ANOVA values it was found that different categorical variables explained different amount of variance.

Chapter 4

Feature Engineering and Data exploration

In *Chapter 2* Pre-processing steps like finding the degree of correlation, cleaning data, finding and removing extreme outliers, finding relevant features by tests like one way ANOVA, F test was carried. This chapter introduces data exploring techniques like checking the data distribution over a certain range of conditions, creation of new features with available data features to create new data, relationship analysis between feature groups, and t-sne plot analysis where high dimensional state space are converted to 2D or low dimensional space data.

The process in which a new variable is extracted from the existing variable by applying the knowledge of domain area is known as Feature Engineering[51]. The extracted features are then used to increase the accuracy obtained by the machine learning models.

Data exploration is the process of understanding the data through visualization and by this the distributions and characteristics of data is analysed[52]. The steps followed here to achieve the task is :

- Finding distributions of necessary variables
- Extracting useful information from exsiting feature to form a new variable
- Analysing the relationship of variables in various combination

Now lets go through each step and perform few experiments:

4.1 Finding distributions of necessary variables

The purpose of finding the distribution of variables is that, if multiple variables are normally distributed individually, then the combination of the variables will also be normaally distributed. Normal distribution of variables allow us to forecast the variable and also to find the probability of a value in the variable using probability density function. The distribution is found for some important variables here and the resullts are shown below:

In Fig 4.1 the distribution of Eye width is drawn along the smooth gaussian curve. This data includes outliers. And outlier is the Eye width equal to 0 as seen in figure. With outlier included the mean is 54.25 and the standard deviation for the data is 18.78.

In Fig 4.2 the distribution of Eye width is drawn along the smooth gaussian curve. This data is with removal of outlier. The distribution approximately fits the normal distribution. Hence the mean is shifted to 59.56 with a standard deviation of 7.53.

The shifts in the mean and standard deviation is one example on the impact of outlier on a data set and distribution.

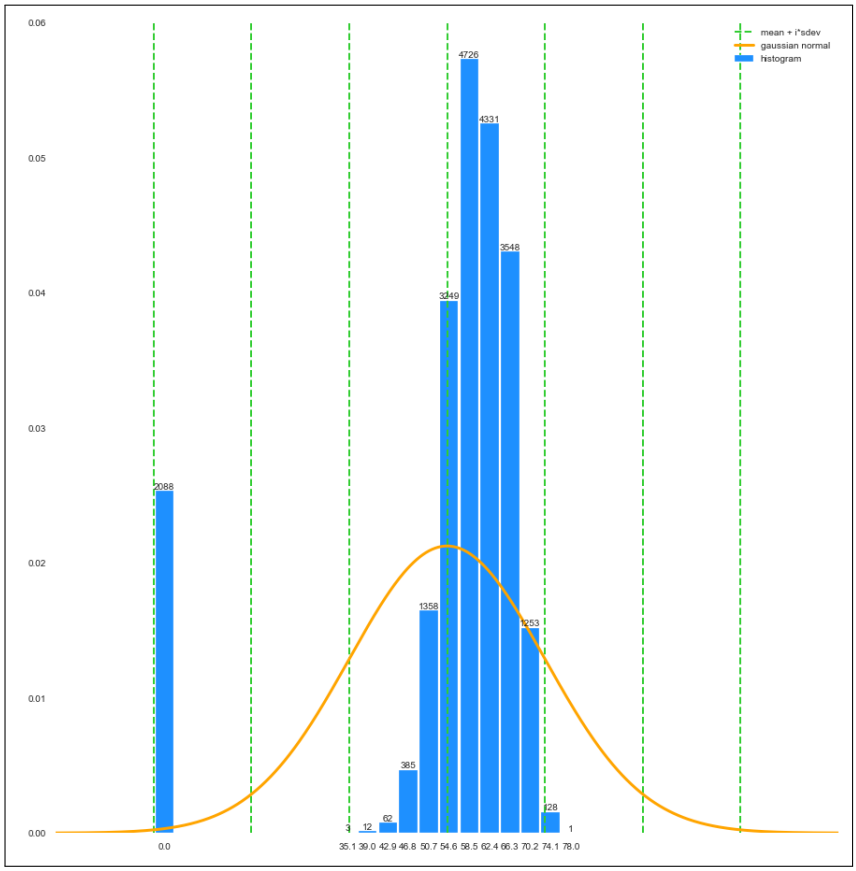

Figure 4.1: Eye width distribution with outlier

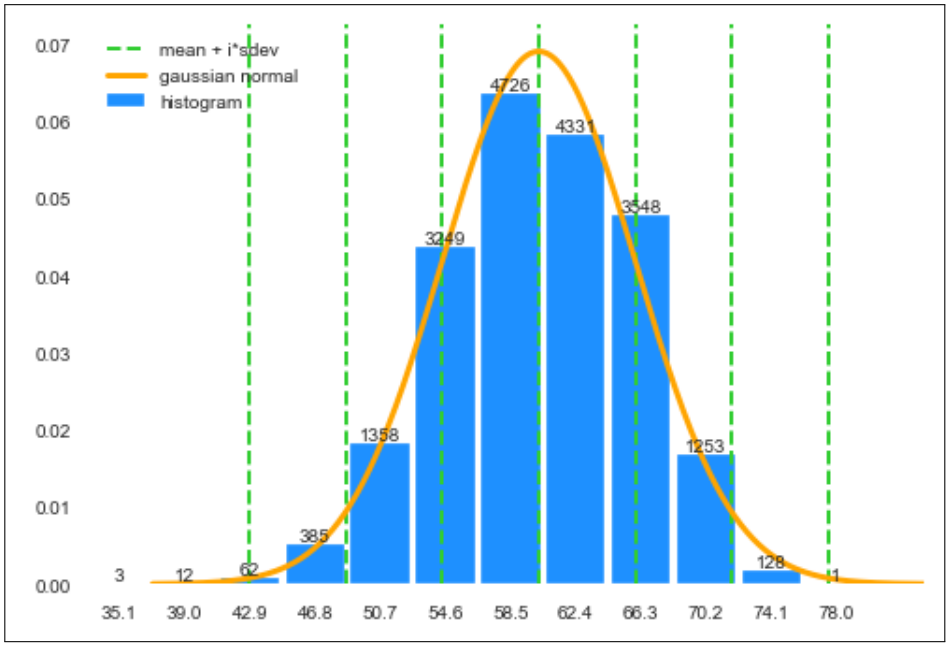

Figure 4.2: Eye width distribution with outlier removed

In Fig 4.3 the distribution of CRPO_VDDN is drawn along the smooth curve. The data is not normally distributed. But the figure give us useful information and suggest that the more component IDs are associated with higher values of CRPO_VDDN, since the graph possess growing trend.

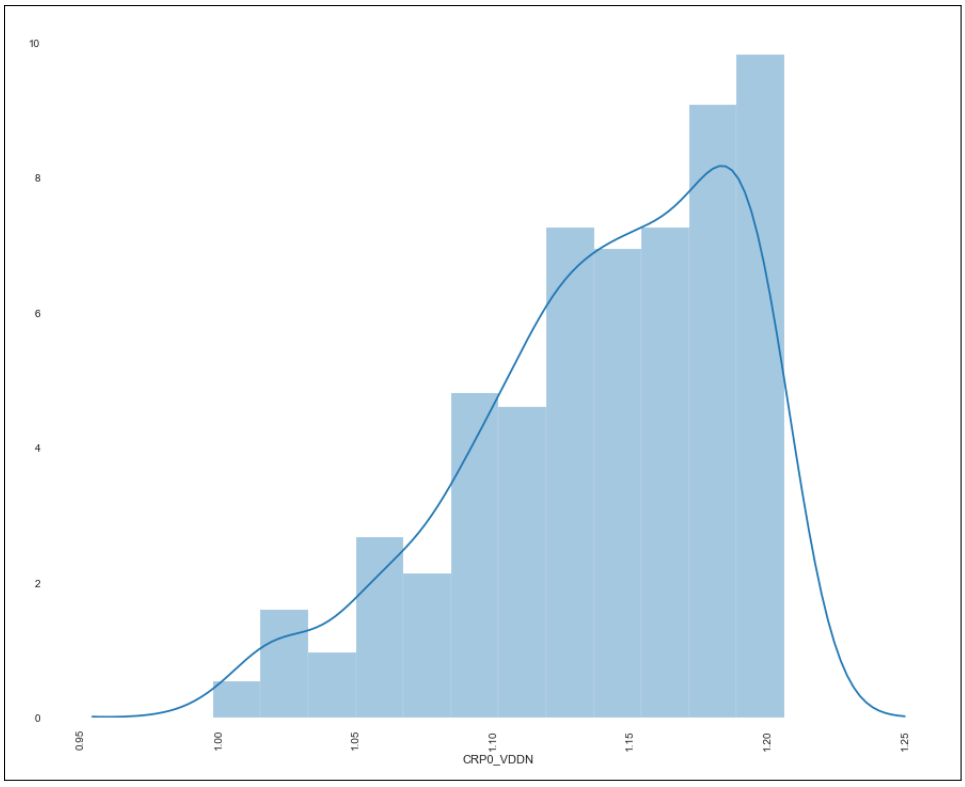

Figure 4.3: Distribution of CRPO_VDDN(voltage) data

In Fig 4.4 the distribution of PSRO is drawn along the smooth curve. The data is not normally distributed. The distribution is random. In Fig 4.5 the distribution of Etch_length6 is drawn along the smooth curve. Just like PSRO data the etch length is also not normally distributed. The distribution is random. As the etch lengths and the variables like delay and loss are highly correlated, it is expected that the distribution of those vairables also approximately follow the same distribution as etch length6. **A notable point here in the given data is that the target variable (Eye width) is normally distributed even when the independent features are randomly distributed. So there is a possibility that even when the inputs are randomly given the ouput would follow a normally distributed values.**

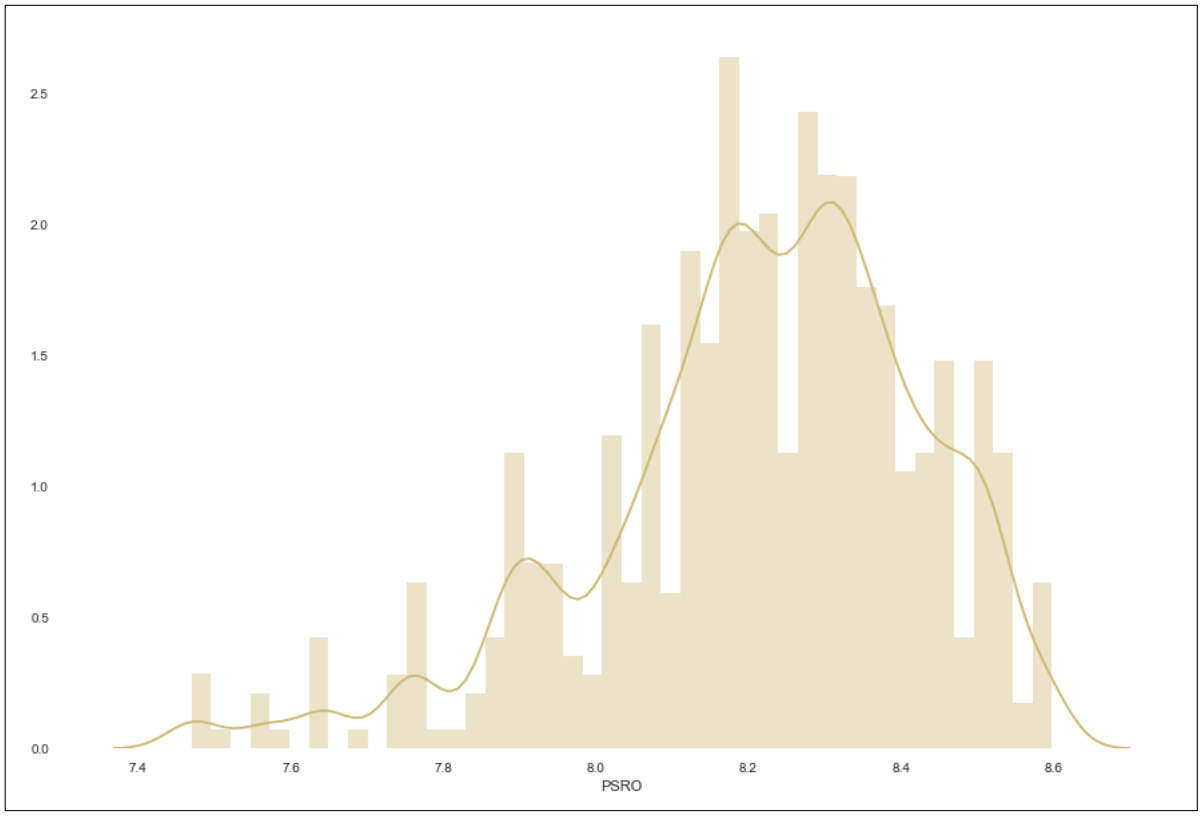

Figure 4.4: Distribution of PSRO data

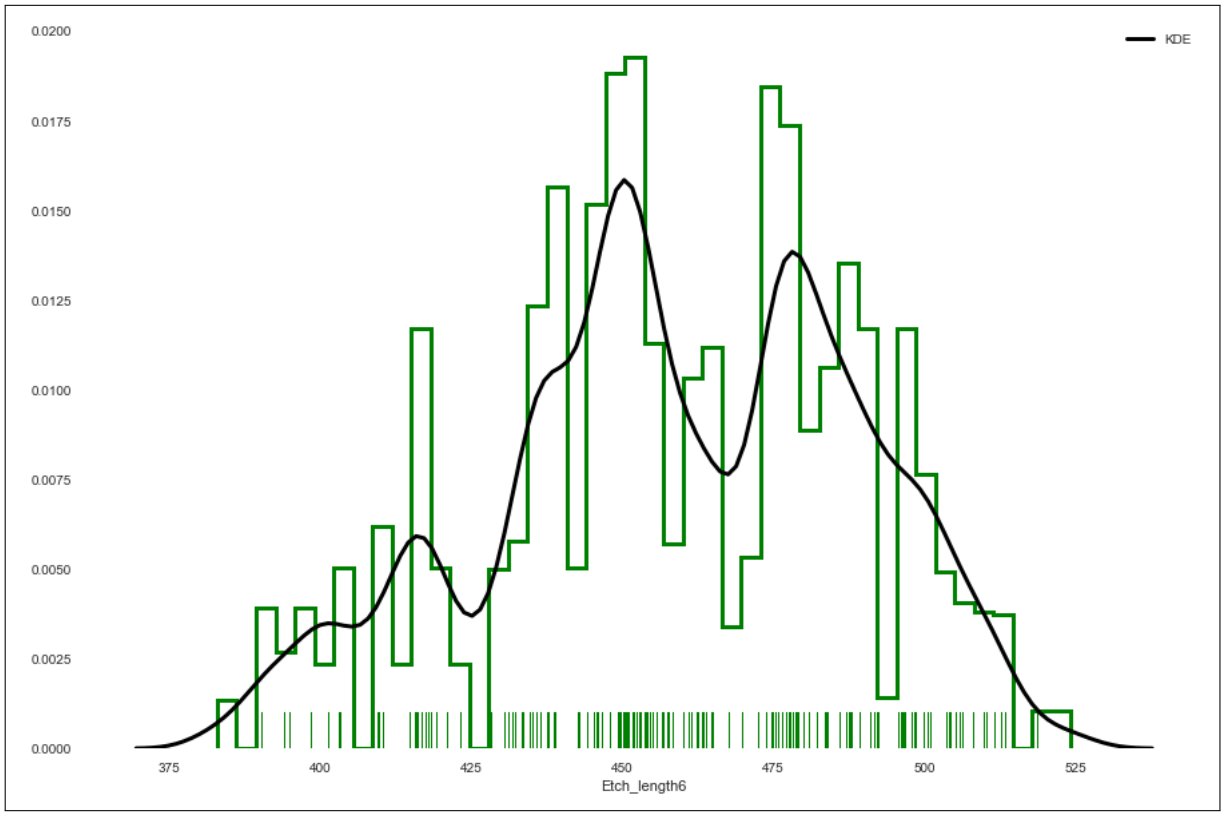

Figure 4.5: Distribution of Etch_length6

4.2 Extracting useful information from existing feature to form a new variable

The purpose of extracting an information from a variable is that, the feature as a whole may not be having a correlation with the target variable but when the part of the feature is extracted, there could be possibility that a correlation exist between the extracted information and the target variable. This experiment is carried out for our data set as shown below:

The variable 'PNP_FILE' contains the information of which machine the test belongs to. For example one instance of PNP_FILE is 'DMIW263F8CCMA4I 2019-01-26-00-28 POK 3906M04.zip.gz' and this contains the information that the test was conducted on machine 04. This can be seen in the string followed by DMIW263F8CCMA4I 2019-01-26-00-28 POK 3906 in the above example. The charcters 'M04' gives the information that the test belongs to MAchine 04.

When this was carried to whole data, it was found that there are total of 5 machines used for the data. The table 4.1 shows the mean and Standard deviations of the Machines. The standard deviations of each machine indicates that within each machine group the standard deviation is more than 3.9, which means that each machine have eye width of different lables on an average.

Machine ID	Total number of machines	Mean	Standard Deviation
M02	12775	57.67	10.84
M03	1375	58.86	10.57
M04	1576	37.38	29.95
M05	704	58.21	13.32
LM5	1920	58.67	12.23

Table 4.1: Number of machines and its standard deviation and mean

The next step was to create bins of Component IDs based on the values of PSRO. The lower the PSRO value, the higher is the speed of component. Hence the median of PSRO was found and it was made as medium speed . And on the either side of that median, at equal distance, further two classes were done on each side. This gave the 'Very high speed','High speed', 'Very low speed' and 'low speed' components. Among 366 components, the fig 4.6 shows the number of components belonging to the classes. According to the figure, large number of components belong to low speed. When this values were

mapped to the whole data set, this number reflect in the table 4.2. The standard deviations of each group indicates that within each machine group the standard deviation is more than 3.9, which means that each machine have eye width of different lables on an average.

PSRO speed class	Total numbers	Mean	Standard Deviation
Very high speed	288	51.21	23.31
High speed	736	54.23	19.67
Medium speed	2230	53.08	20.38
Low speed	4672	56.5	15.82
Very low speed	1832	53.83	19.21

Table 4.2: Binned PSRO classes and its standard deviation and mean

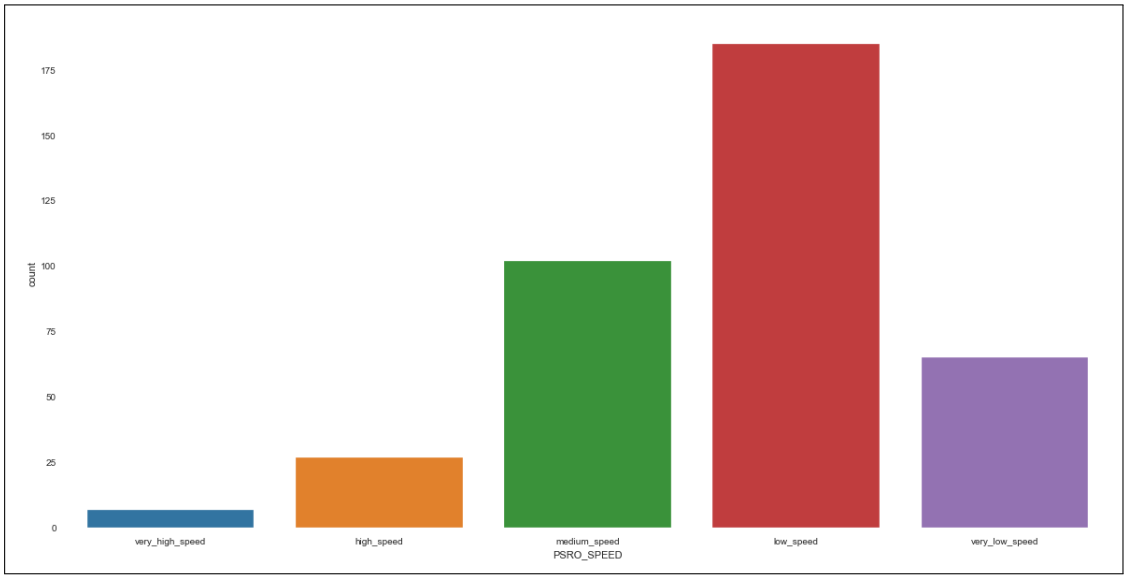

Figure 4.6: Classes of PSRO speed based on binning

Chip Sorting:

The values of the variable Component ID contains the information of the chip class used in the Central Processor. As an example the 4th digit to 10th digit of the component ID '11S02AA435YH1934081365' contains '02AA435' which is the class of Chip used. When the complete dataset was scanned for different Chips, totally there were 13 chip classes available. The table 4.3 shows the different classes of chips and their mean eye width along with standard deviation. As the number of data of each class has varied, the mean and standard deviation has also varied. But the important thing to note here is, the standard deviation varies largely in each class and it shows that the data is widely spread. The mean of most classes also does not vary more than a single rotator shift (3.9) with each other. Hence these classes do not has strong correlation associated with Eye width.

Chip Sort	**Mean**	**Standard deviation**
02AA457	57.99	4.4733
02AA439	60.69	4.95
02AA435	58.03	8.502
02AA475	56.43	9.39
02AA453	57.07	10.02
02CY395	58.89	10.7
02AA471	57.71	12.51
02CY381	57.01	13.37
02CY383	57.40	15.32
02CY389	57.44	15.40
02CY388	54.951	17.94
02CY382	52.72	20.04
02CY394	52.76	21.2

Table 4.3: Chip sorted classes with mean and standard deviation

4.3 Analysing the relationship of variables in various combination

The purpose of this section is to analyse the various variables in combination and then deriving the relationship of them with respect to eye width. This can give some of the important information on how the variables are distributed as a group.

The figure 4.7 shows the visualization of Slot ids present in Machines. A typcial machine suppose to have 20 slot ids but the data provided contains maximum of 19 Slot ids in Machine 04. The LM5 is a low end machine which do not possess slot ids in CP2. This figure along with fig 4.8 suggest that all machines do not contain exact same comparable data. Different machines has different number of data and different slot id numbers. The provided data covers contains most of the data related to M02 as seen in fig 4.7

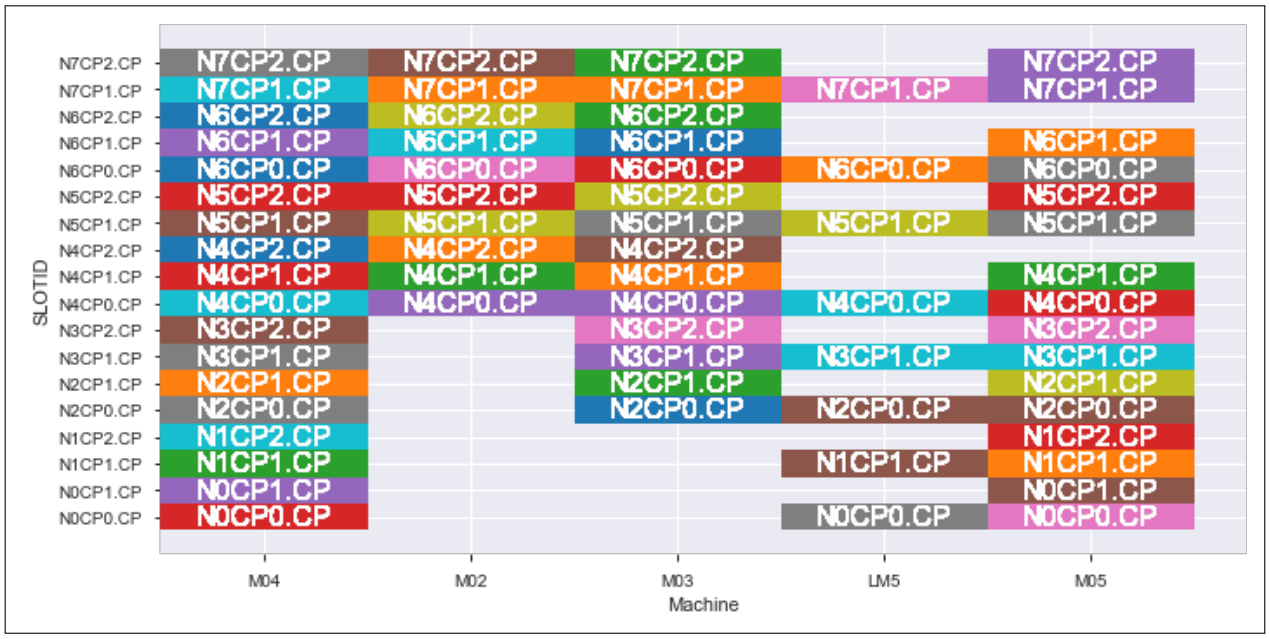

Figure 4.7: Slot IDs present in the Machines

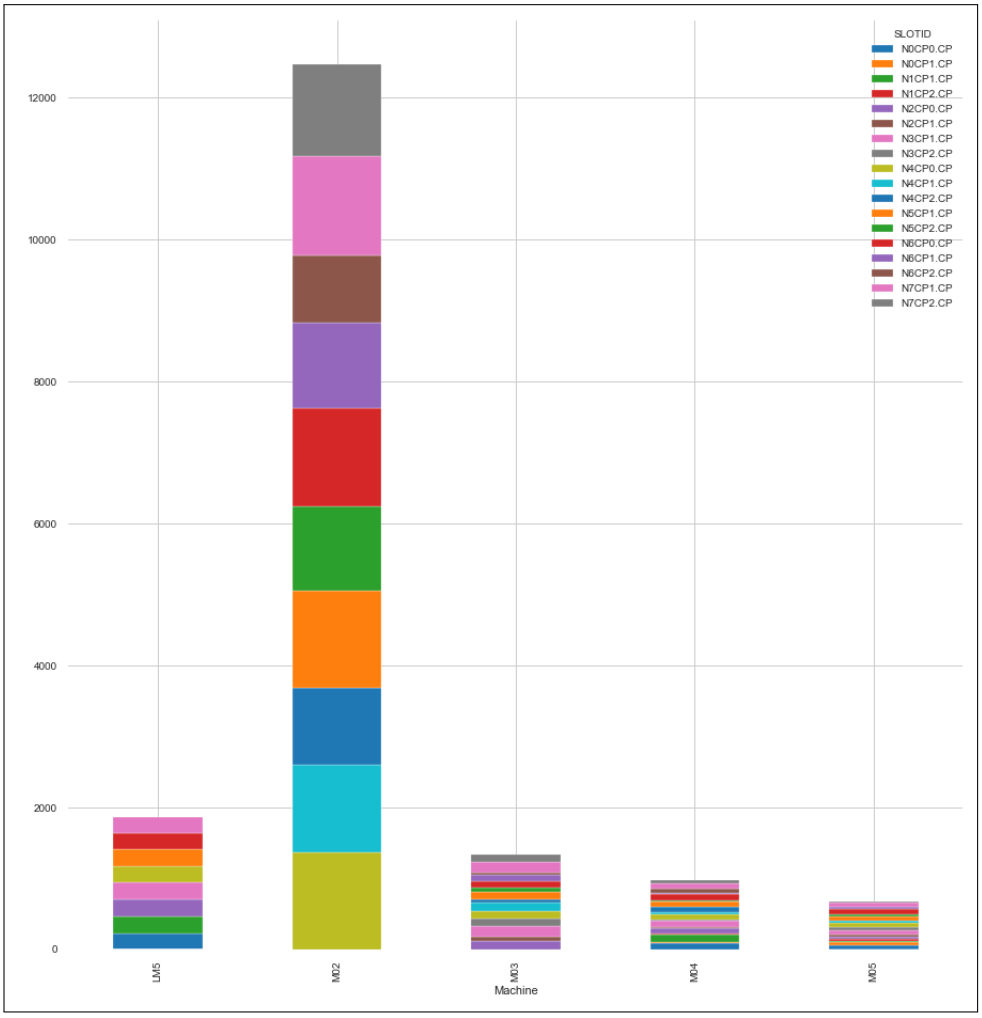

Figure 4.8: Data size of Machines and corresponding Slot IDs

The fig 4.9 shows the visualization of relationship between Delay and etch length 6 where as fig 4.10 shows the visualization of relationship between Delay and Loss. Both the figures suggest a strong linear relationship between the variables. It can also be derived from this figures that there is linear relationship between loss and etch length because the both etch length and loss has linear relationship with delay. By including these three variables to the models can lead to multicollinearity problem because all these variables are highly correlated to each other.

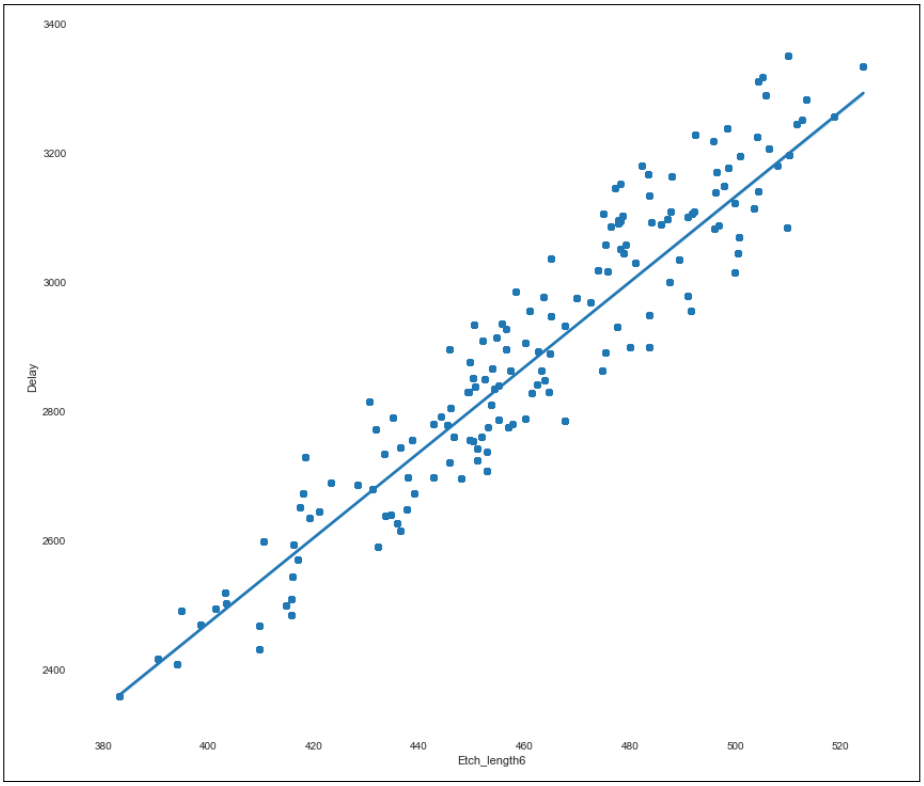

Figure 4.9: Delay and Etch_length6 linear relationship

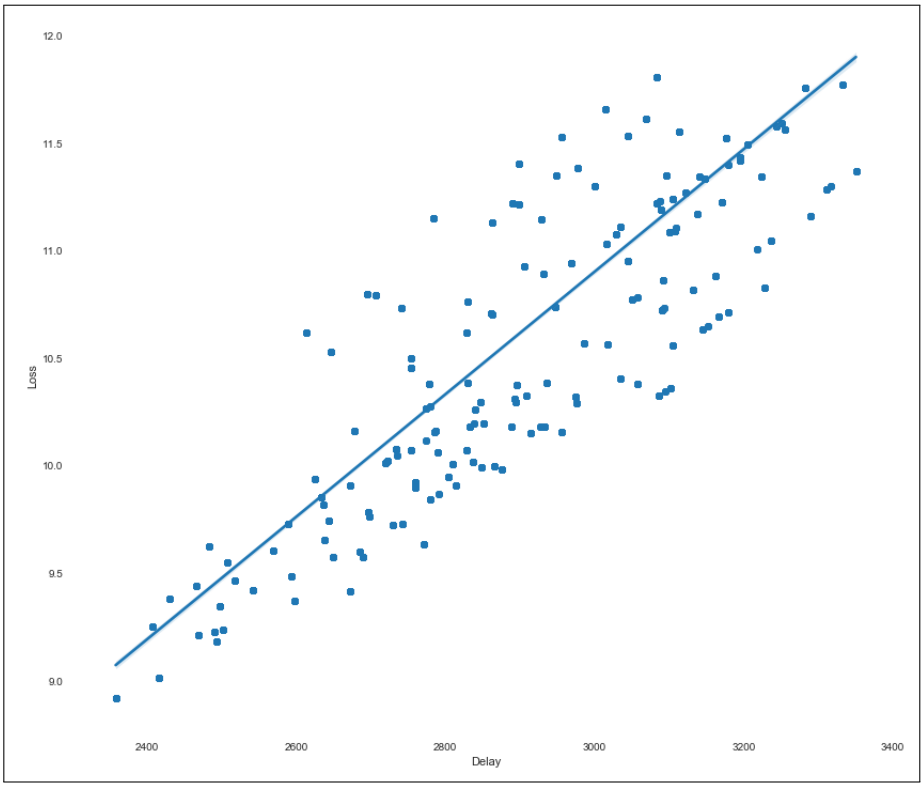

Figure 4.10: Delay and loss linear relationship

The fig 4.11 shows the distribution of number of components and the corresponding eye widths. This follows a normal distribution curve. It can be seen that among 366 components, most of the components read eye widths around mean ateast once. Where as the number of component ids reading particular eye widths become less as moving across each standard deviation. So the out come of this visualization is that there are no certain components favouring a certain eye widths strongly since both eye width and component ids with respect to eye width follow normal distribution.

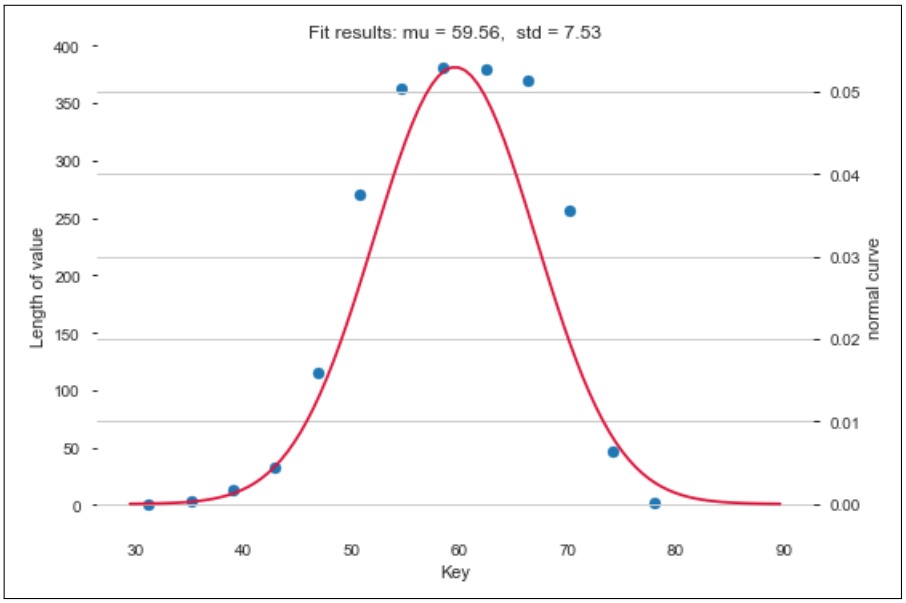

Figure 4.11: Distribution of Number of components having specific Eye width value

The fig 4.12 shows the distribution of lanes with respect to eye widths. There is also visualization of heat map. The heat map suggest that for each lane the population of eye widths peaks at around mean eye width(around 58.5) and then population decrease across mean. One more point is that the peaks around the mean are higher for Lane 0 and the peak decrease as lane goes further to 8 and then the peak raises again until Lane 15 (the dark red color suggests this). The fig 4.13 is a plot of ach lane and corresponding ey widths. This figure suggest that most lanes are asscociated with every possible value of eye width. To derive dependency of eye width on lanes, each lane should have been populated with only certain fixed eye widths. The The fig 4.14 gives the standard deviation with in each lane with respect to eye width. Here with in each lane the standard deviation is more than 1 rotator shift(3.9 eye width units). It suggests that each lane is populated with more distinct eye widths. The fig 4.15 shows the means of each lane with respect to eye width. It is seen that that most of the lanes has means around 60 and most means of lanes are not separated from even 1 rotator shift. These could be the evidences that Lanes do not have an impact on outcome of eye width.

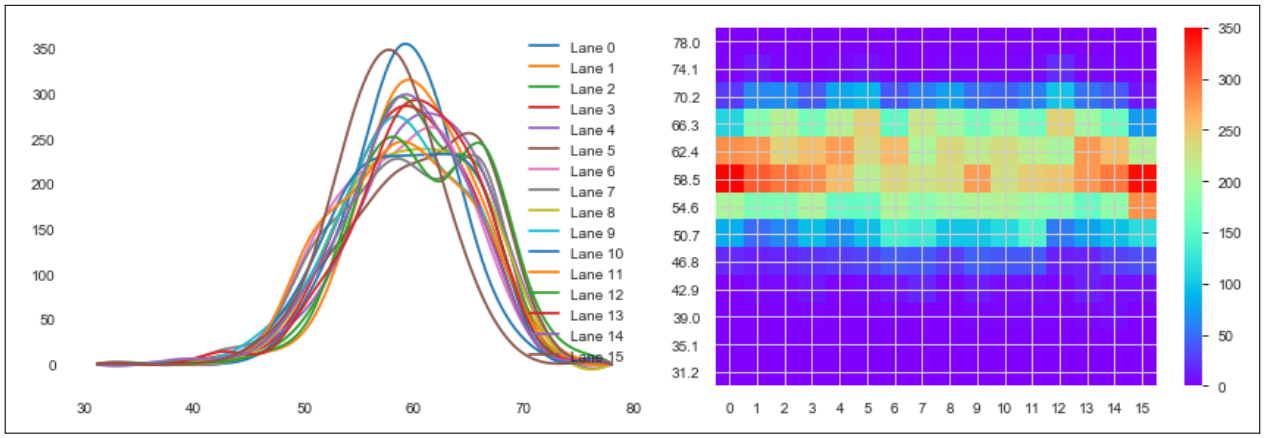

Figure 4.12: Distribution of Lane with respect to Eye width and its heat map

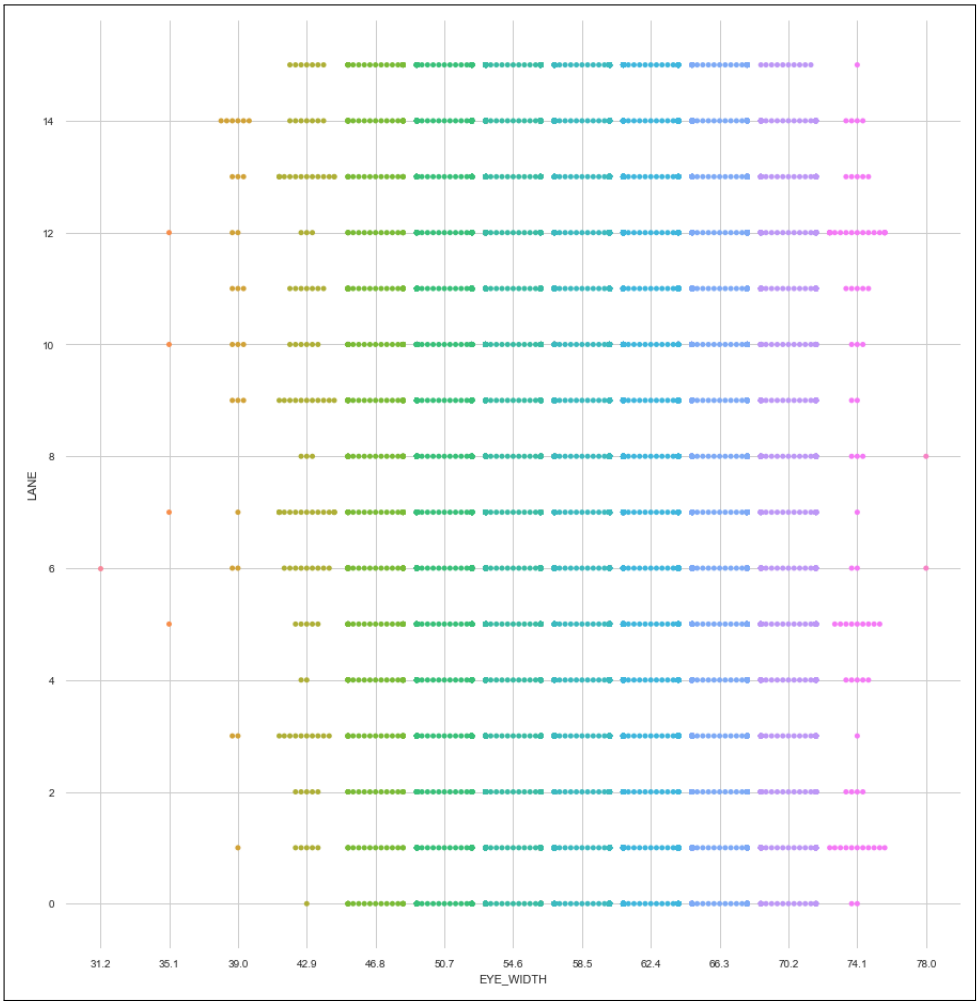

Figure 4.13: Distribution of Lane with respect to Eye width

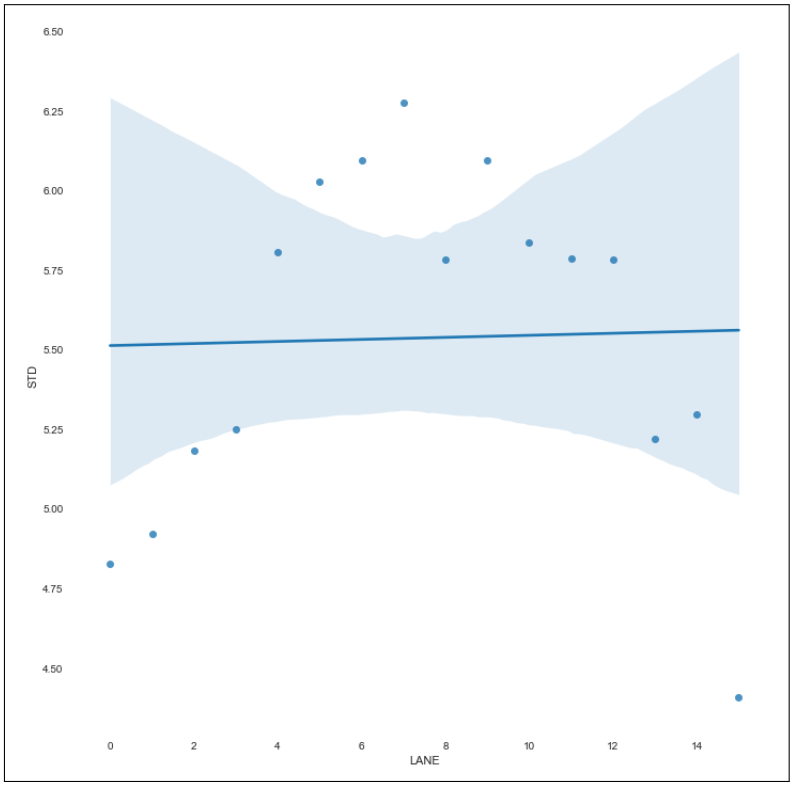

Figure 4.14: Standard deviations of Lanes with respect to eye width

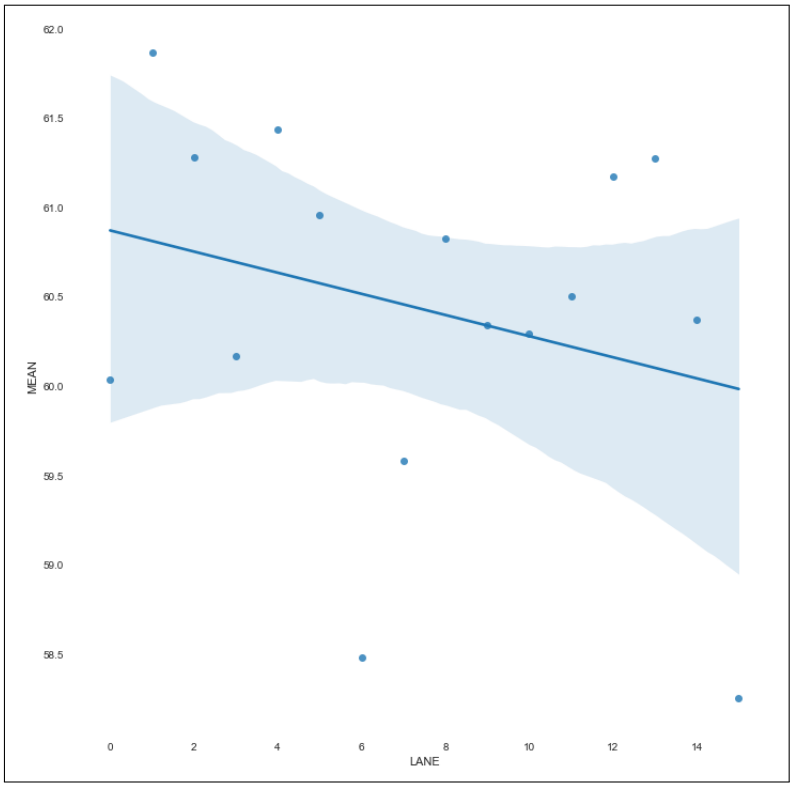

Figure 4.15: Mean of Lanes with respect to eye width

4.4 Grouping of various categorical variables and checking eye width

The purpose of grouping various categorical variables and checking the associated eye width because, if there are any relationship between certain group of unique variables and eye width, then each unique group would be associated with a unique Eye width. This was tested for the whole dataset and thenit was found that each unique group of variables were associated with multiple eye widths which suggested that there is actually no relationship between eye width and categorical variables. In ther words whatever the combination of variables are given, there is chance of any random eye width associated with that data.

Slot ID	Lane	Plug position	Machine	Eye width
N0CP0	0	PB0	LM5	50.7 54.6 58.5 62.4
N0CP0	1	PB0	M02	50.7 54.6 58.5 62.4
N0CP1	2	PB0	M04	66.3 54.6 58.5 62.4
N0CP1	1	PB0	M05	50.7 54.6 58.5 62.4

Table 4.4: Different Variables and Eye width Combination

The table 4.4 shows a part of the analysis, as observed, any combination of variables leads to the same set of eye width values. If there was a relationship between the variables and eye width, the unique combination of variables would have lead to unique eye width value. When tested for whole dataset, it is observed that there is no correlation of eye width and categorical variables.

4.5 TSNE Plot

The motivation behind applying t-distributed stochastic neighbor embedding (t-SNE Plot) is that, if there is a boundary that can be drawn between the data of different labels(Eye width) of the data, the TSNE plot shows the boundary between the different labels data. If there is no boundary that can be derived from them, the TSNE plot overlaps the different labels data.

t-distributed stochastic neighbor embedding (t-SNE Plot)[43] are used to convert multi-dimensional space to 2 dimensional (2D) space in statistical study. The working of t-SNE plot can be comparable to the Principal Component Analysis (PCA), but TSNE works with non-linear functionality and considered as optimal method to consider incase data consists of outliers.

In [11], t-SNE have been implemented along with Tree-Based Algorithms. In the paper, t-SNE used for the visualization and interpretation of multi-dimensional data and used Barnes-Hut algorithm with tree-based algorithm to accelerate the t-SNE working. In the result, conclusion involves that t-SNE had performed better than dual tree algorithm. T-SNE is better than other visualization plot form mapping high-dimensional to low-dimensional plots. In[12], t-SNE is compare with Sammon mapping, Isomap, and Locally Linear Embedding, and the output from t-SNE is considerably better than other techniques.

In this thesis work, t-SNE have been used to visualize the high-dimensional datasets in low-dimensional space. This was beneficial to view datasets output with different groups or does they overlap with all given different groups.

The Eye width variable in the data set consist of about 13 lables varying from 35.1 to 78.0. If each class of label is different from one another, it can be seen in the TSNE plot. First the data belonging to each label is made as a different dataset. Then the TSNE function is applied to reduce the datasets into two dimension. Then this data is scatter plotted. If there is a boundary separating between each labels, then it is an evidence that the data belonging to each label can be distinguished from each other.

For simplicity only the data from labels 50.7 and 62.4 is taken. The fig 4.16 shows the t-sne plot of labels 50.7 and 62.4. The two data is overlapping which is leading to the distortion in clarity of viualization. This is an another evidence that the data of different labels of eye width are not distinguishable from each other. Hence it is suspected that even the machine learning algorithms would lead to low accuracy when predicting the eye widths.

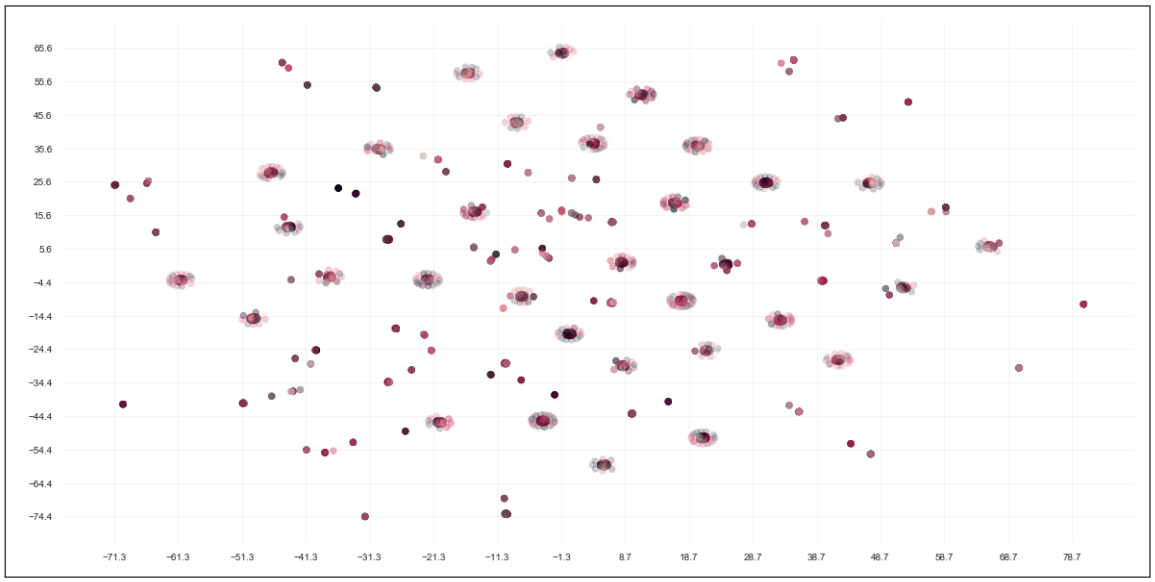

Figure 4.16: TSNE Plot with Eye width labels 50.7 and 62.4

4.6 Summary and Conclusion for the chapter

From this chapter following points can be concluded:

1. Eye width data is approximately normally distributed. While the other variables are randomly distributed.
2. Adding features like PSRO, CRPO_VDDN, Machine, Chip sorting did not derive correlation with Eye width
3. Lanes, Component Ids and Slot ids did not had specific relationship with eye width.
4. When the categorical variables were grouped, there were multiple eye widths associated with each group. And there were overlapping of eye widths. This was an illustration that the categorical variables did not had any correlation with eye width.
5. The visualization of TSNE plot for different labels was overlapped proving that the data of different eye width labels cannot be distinguished.

Chapter 5

Treating problem with Machine Learning Regression Methods

This chapter gives a brief introductory study of theoretical and practical deployment of regression techniques like the OLS model, Random Forest, and Neural Network models. The implementation and result evaluation are analyzed here. Further, the comparisons between tree-based Random Forest and network-based Neural Network model is made with implementation.

The motivation behind experimenting with OLS[53] is that, it is a linear regression model and it determines how well the data is linearly fitted. And the motivation behind experimenting with Random forest is that if there is no linear relationship found from OLS, the non-linear relationship is extracted from Random forest if there is any.

The desirable RMSE is less than 3.9 from the models. Because each eye width label is separated from space of 3.9. And this one space of 3.9 is called one rotator shift. And on average if the model is predicting with in one rotator shift, the model is considered good in our case.

The analysis of various machine learning models and its results are presented in the following sections.

5.1 OLS model and Random forest Implementation

Ordinary Least Square regression is the statistical approach of minimizing the predicted values and original values. This method creates the relationship between input features and output features and build the connection between them. The coefficient of determination which can be obtained from regression model is called as R squared value describes whether the model is fitted better or not. The R square value closer to 1, represents the model is fitted best, as lower the values the model is not fitted well[13]. Outlier detection and finding the extreme points are critical things in data science models. OLS regression are used to find the influencials in the datasets to make the better model tuning[14].

Random Forest are tree-based methods in machine learning which works based on the decision of trees. Each tree makes its own decision and combined to take maximum vote in case of classification whereas mean values are considered in case of regression trees. The main advantage of random-forest are the flexibility of bagging and bootstrapping approach, through which overfitting of the model can be reduced significantly. The random forest is used for both classification and regression calculations. The features like Bagging, Boosting, Out of Bag error features in the random forest trees makes the model standout in machine learning models. Both classification and regression problems can be solved by using this algorithm. In case of classification, maximum vote of tress and in regression mean value of all trees are considered.

The fig 5.1 shows the visualization of the random forest for the data set containing only 10 Component IDs. And only two variables are considered to make the visualization better. It is observed that for all the value of PSRO and CRPO_VDDN, a branch is built. This is an indication that there is no boundary that could be drawn between the variable values and Eye width. This is a testament for the result we got from TSNE plot. The random forest is a model which is able to derive the non linear relationship between the target variable and independent variables. The data which is loaded to random forest algorithm need not be standardized(put reference) since the technique involved here is building tree type structure.

The visualization of the random forest for the whole data set was attempted. But since there are lot of variables and huge data, and since for this particular data the random forest splits for every available value due to lack of relationship between target and independent variables, the tree was too complex to read.

Further analysis of the results are discussed in next section.

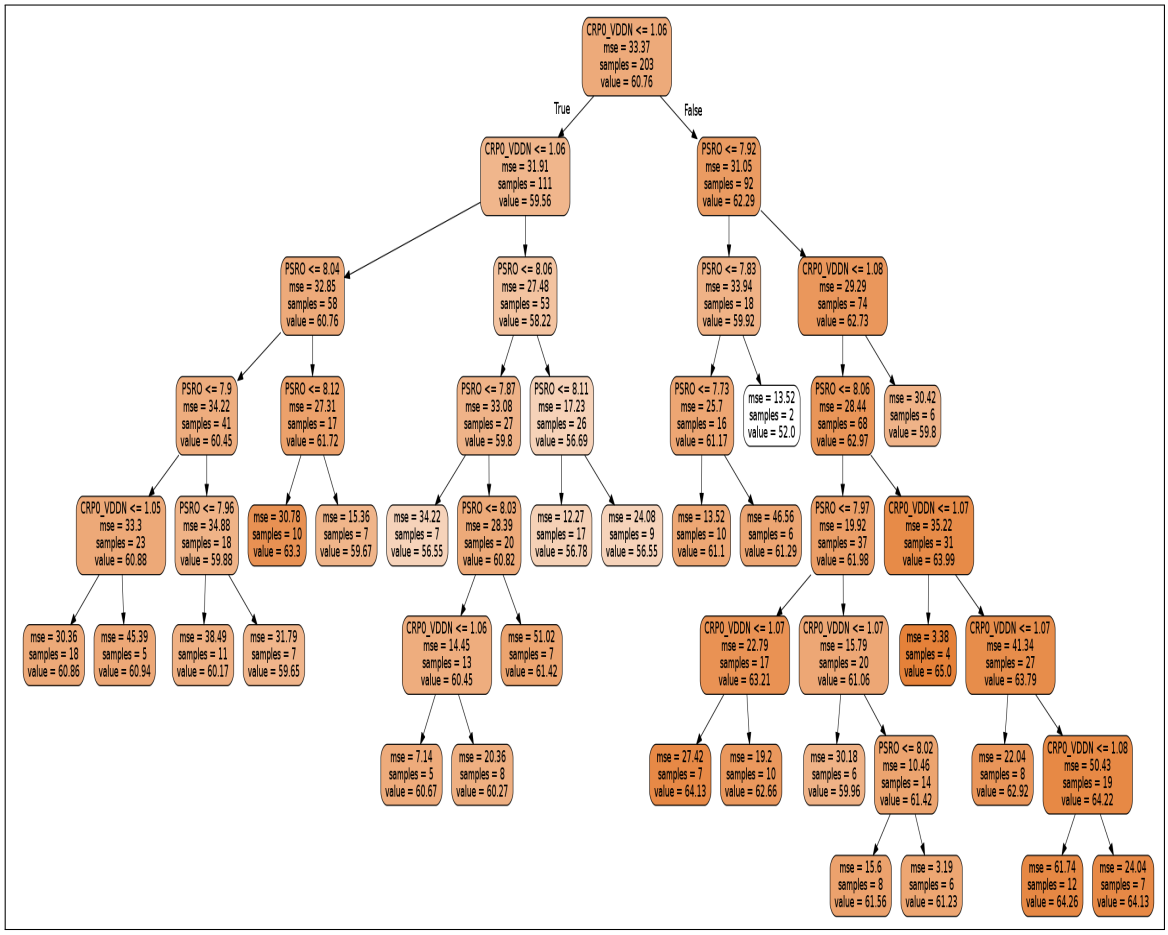

Figure 5.1: Random forest visualization with reduced data set

5.2 Result Anaysis

1. **OLS Model Results:**

   The fig 5.2 shows the OLS model results which was constructed by feeding all the numerical variables and the Eye width variable into the OLS model. The **R-squared value of 0.030** indicates that only 3% of data is linearly fitted and hence the model results are not reliable. The table of P-values below also indicates that there are no variables(other than CRPO_VDDN and PSRO) that have prominent linear relationship with Eye width.The threshold P-value is generally considered to be 0.05 or 5%. The variables below 0.05 P-value is considered to be important variable. However since the R-squared value is very low, the P value can also be not reliable.

OLS Regression Results

Dep. Variable:	EYE_WIDTH	R-squared:	0.030
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	24.53
Date:	Wed, 01 Jul 2020	Prob (F-statistic):	1.28e-50
Time:	21:05:39	Log-Likelihood:	-27762.
No. Observations:	8790	AIC:	5.555e+04
Df Residuals:	8778	BIC:	5.563e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-148.7055	113.742	-1.307	0.191	-371.666	74.255
LANE	-0.0842	0.061	-1.375	0.169	-0.204	0.036
Etch_length1	1.137e+05	1.43e+05	0.796	0.426	-1.66e+05	3.94e+05
Etch_length6	-3.797e+04	4.76e+04	-0.797	0.425	-1.31e+05	5.54e+04
Etch_length2	1.136e+05	1.43e+05	0.795	0.427	-1.67e+05	3.94e+05
Etch_length3	1.136e+05	1.43e+05	0.795	0.427	-1.66e+05	3.94e+05
Etch_length4	1.136e+05	1.43e+05	0.795	0.427	-1.66e+05	3.94e+05
Etch_length5	-7.582e+04	9.53e+04	-0.796	0.426	-2.63e+05	1.11e+05
pin_length	3.785e+04	4.76e+04	0.795	0.427	-5.55e+04	1.31e+05
PSRO	-0.7055	0.334	-2.113	0.035	-1.360	-0.051
CRP0_VDDN	-5.2830	1.576	-3.353	0.001	-8.371	-2.195
Delay	4.3162	1.679	2.571	0.010	1.025	7.608
Loss	5658.4189	2475.131	2.286	0.022	806.582	1.05e+04

Omnibus:	108.165	Durbin-Watson:	1.779
Prob(Omnibus):	0.000	Jarque-Bera (JB):	111.397
Skew:	-0.271	Prob(JB):	6.46e-25
Kurtosis:	2.895	Cond. No.	7.08e+16

Figure 5.2: OLS model results with R-squared and P-values

2. **Random forest Model Results:**

   The Random forest model was constructed by feeding the categorical variables after applying target mean encoding, and also the numerical variables. The data was not standardized since the random forest model results does not alter with standarized and non standardized data. The fig 5.3 shows the results obtained from random

forest model without tuning. The RMSE is about 5.33, which is more than one rotator shift. The parameters of this model is shown in fig 5.4. Since the model without tuning gave a result which is not desired, the model is further tuned using Random search tuning method. The results of this model are shown in fig 5.5. And the tuned parameters are shown in 5.6. But the result has not improved much. This is a further testament to our previous analysis that, there is no useful relationship between the eye width and other variables.

```
Mean Absolute Error: 4.240767223386173
Mean Squared Error: 28.503624147683738
Root Mean Squared Error: 5.3388785477554865
```

Figure 5.3: Random forest results without tuning

```
Parameters currently in use:

{'bootstrap': True,
 'ccp_alpha': 0.0,
 'criterion': 'mse',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': 0,
 'verbose': 0,
 'warm start': False}
```

Figure 5.4: Random forest parameters before tuning

```
Mean Absolute Error: 4.217555390045774
Mean Squared Error: 27.71933574304343
Root Mean Squared Error: 5.264915549469282
```

Figure 5.5: Random forest results with tuning

```
{'n_estimators': 94,
 'min_samples_split': 15,
 'min_samples_leaf': 1,
 'max_features': 'sqrt',
 'max_depth': 20,
 'bootstrap': False}
```

Figure 5.6: Random forest parameters changed after tuning

3. **Feature improtance from Random forest Model and the drawbacks:**

   Random forest model from python provides a feature in the form of a Random forest variable known as feature_importances_. This feature enables us to know which variables were important in building the trees based on Mean decrease impurity and Mean decrease accuracy(put reference).

   The fig 5.7 shows the ranking on features of our data set. It is seen that Component ID is among the top of ranking. This result leads us to one of the drawbacks of random forest. There are total of 366 component IDs in the dataset. This makes it to be 366 distinct values. Among all the features Component ID has the most number of distince values compared to other features. Other than component ID, the next feature with highest number of ditinct values is Etch_lngth6 with 160 unique values.

   The Random forest is found to split the tree with the feature having outstanding number of distinct values. Hence this variable comes out as important feature automatically from the random forest[59]. Since the result from random forest gives evidence that there is no non linear relation existing between eye width and other features, we were also exposed to the drawback of random forest by this study.

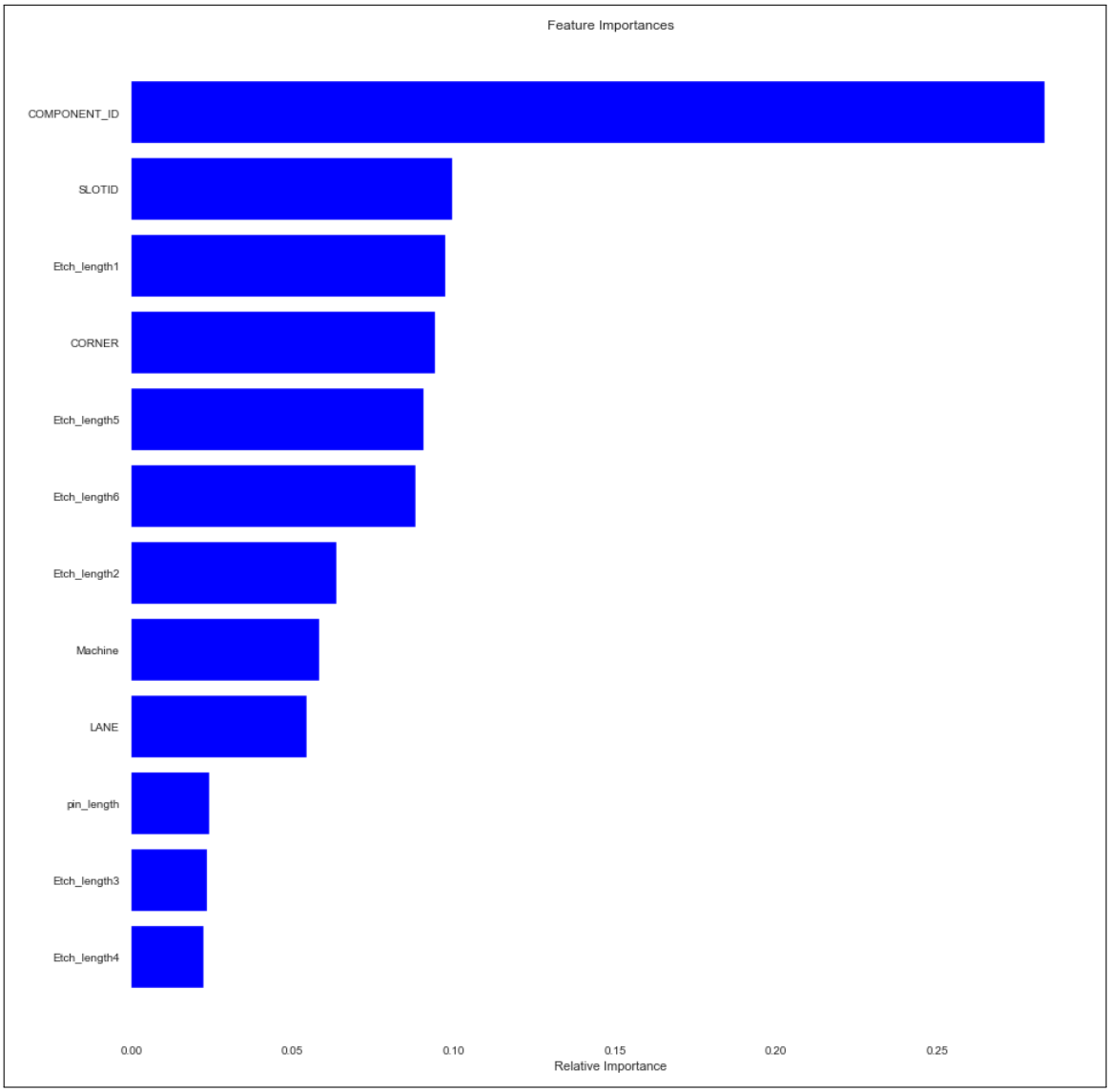

Figure 5.7: Important features from Random Forest perspective

5.2.1 Neural Network Model Implementation

The purpose of applying Neural networks to our data set is that, it is a modern modern machine learning architecture which can derive linear or non linear relationship between variables.

Unlike random forest which uses the tree based architecture, neural network uses a different mechanism. It uses densely interconnected cells to learn patterns. This model is applied to check whether Neural network can be capable of deriving hidden non linear relationship which was not derived by random forest.

Neural Network is one of the famous deep learning technique used in AI and ML. The human brain comparable intelligent structure makes the Neural Network stands out from the normal ML algorithms. The memory cells in the network allows to capture the past information of the data by continuous feedforward loop.

Recurrent Neural Network(RNN), is the neural network structure, where data is processed recurrently through out the model and the error rate is reduced by the backpropagation process. LSTM (Long Short Term Memory) and GRU (Gated Recurrent Unit) are the types of NN which comes under RNN. There is modifications in the memory units between these networks.

Since the problem does not demand special purpose neural networks, the neural network used for our analysis is a simple neural network with 3 dense layers.

In the following sections the reults and also the comparision between neural networks and random forests are presented.

5.3 Result Evaluation

The neural network used here is a simple neural network with three dense layers using adam optimizer. Important point to note here is that the neural netwrok must be fed with the standaridized data since the weightage associated to each variable should be evenly balanced. If the non standardized variables are fed, the weightage to the variable with large range becomes automatically more prominent. THe fig 5.8 shows the result from a neural network. And it is no different from the output of random forest. However from the 'PermutationImportance' feature from 'eli5' package, when the important variable was extracted as shown in fig 5.9, based on the weightage associated with each variable, it was different from the important features displayed by random forest. This is due to the fact that neural network does not suffer from the draw back that was mentioned in the previous section. And also the mechanism of both models are very different from each other.

```
Mean Absolute Error: 4.470391166658082
Mean Squared Error: 30.41433164640981
Root Mean Squared Error: 5.514919006332714
```

Figure 5.8: Neural Network results with standardized data

Weight	Feature
61.4046 ± 2.7280	Etch_length4
33.0798 ± 1.0078	Etch_length1
28.8849 ± 0.6049	Etch_length3
14.9031 ± 0.6375	Etch_length5
12.9663 ± 0.5489	Etch_length6
12.7068 ± 0.3608	pin_length
4.4005 ± 0.3422	COMPONENT_ID
3.6972 ± 0.3384	Etch_length2
2.4754 ± 0.1421	LANE
1.5011 ± 0.1402	CORNER
1.1470 ± 0.0683	Machine
0.2987 ± 0.0377	SLOTID

Figure 5.9: Important features from NN perspective in terms of weights associated

5.4 Comparisions of Random Forest and Neural Network Models

Random Forest is the tree-based model, where each tree decides the prediction values, whereas in the neural network model works on the neuron models, where continuous feedforward loop where the network layer predicts the output. The advantage of neural network layer is the backpropagation technique where the error rate is reduced continuously till the error is reduced where such technique is not present in the random forest-based algorithm.

Regarding the tuning of the model, the random forest can be trained with setting number of trees, by adding default hyperparameters or either using hyperparameter optimization results in the better prediction. But tuning of the neural network is tedious, where setting the number of network layers, type of activation function, and initialization of weights, with the type of neural network. Comparing the computational speed, the tree based model is more cost efficient than neural network, where it takes longer time to compute the results.

As per our observations following points can be majorly considered:

1. According to our experiment, the time taken for the running the neural network model took more time than that of Random forest. The results in accuracy, produced by each model was not much different to each other.

2. The tuning of Random forest parameters was a simpler approach using the RandomizedsearchCV tuning. Neural network parameters were tuned manually, which was time consuming.

3. The important features produced by both the models were very different due to their different mechanisms.

4. The random forest can operate with non standardized data where as the neural network needs standardized data[46][47][48].

5.5 Treating problem with classification methods

After the implementation of the regression-based model, in this chapter classification model techniques have been implemented. It elucidates the reason for classification requirements with Random Forest classifier implementation and results analysis.

5.5.1 Reasoning for classifying

The purpose of applying a classification model here is that because there are 13 distinct values of eye width in the data set. Each label here is spaced by 3.9 units. A value of 3.9 units here is called rotator shift. The problem with the regression model is that, they do not account for the spacings, and predict the values in between the spacings.

For example, the valid eye width values are 50.7 and 54.6. The value of 52.3 is not valid here. But the regression models also predicts the values in between 50.7 and 54.6. This can lead to inaccurate results. Hence each eye width value here is treated as a label, and those data are fed to random forest classifier. This is further discussed in next section.

5.5.2 Random forest classifier and reasons to consider it

Random forest is an ensemble tree-based learning algorithm. There are other classiying algorithms like K Nearest Neighbor, Support Vector Machine and Naive Bayes algorithms. Among these Random forest Algorithm was choosen based on the training data size, presence of correlated features,non standardized data and the decision trees technique.

Random forest algorithm is robust even in the presence of highly correlated features. A simple test was conducted and the random forest performed slightly better for our training data set. Hence the result from random forest is presnénted here.

5.5.3 Result analysis

The following points were analysed:

1. The random forest classifier was able to predict 33% of the eye width test labels accurately based on the trained data set. 60% of the test lables were predicted within one lable left or right. 75% of the times test lables were predicted within two lables left or right. This table is shown in table 5.1.

2. This is an indication that the eye widths have random relationship with the other variables. The classifier is unable to classify the data rightly. 2/3 of the times the model classifies data wrongly.

3. It is possible that the eye width is predicted around the mean value most of the times and this results in 33% accuracy. When the results were analysed, the model had predicted varying values.

4. The results are no much different to that of regression models. The regression models had mean error of approximately between 5 and 6. This corresponds to the error between one to two rotator(3.9 to 7.8) shift misclassification when tranformed to classification.

Label Misclassification	**Accuracy in Percentage**
+/- 0 rotator shift	33%
+/- 1 rotator shift	60%
+/- 2 rotator shift	75%

Table 5.1: Random forest Classifier results

5.5.4 Summary and Conclusion for the chapter

The folowing points can be summarized:

1. The OLS model showed that the dataset has very low R-squared value equal to 3%. Hence ruling out the linear relationship.

2. The random forest trees were visualized. The random forest built the tree by splitting each and every value. The margin of RMSE was also high. This is an indication that non-linear relationship does not exist too.

3. It was found that the random forests have the problem of selecting variables having more distinct values. It was seen in feature importance provided by random forest too.

4. In addition to random forests, the neural networks also produced the similar results. The RMSE was high. The important features according to neural network was presented by using Permutation Importance of Eli5 package.

5. Hence with the help of machine learning models, it can be concluded that the variables are not having any kind of relationship with the eye width.

Chapter 6

Analysing the distribution of eye width and Automating Data collection from Data warehouse

The purpose of this chapter serves to the need for laying the foundation to future work. The concepts of Central limit theorem, Probability density of function, normal distribution is introduced which are the necessary terms to be known to analyze the randomness of variables.

Since from previous analysis we already know that the eye width variable from the sample data(1 month) is normally distributed. Now it is necessary to find the distribution of larger population. Hence the data for over 4 years is collected and is needed to be varified whether this data follows normal distribution as well.

The mean and standard deviation in a normally distributed curve gives us an opportunity to find the confidence interval of mean and also to forecast the values through the probability density funtion.

If the data is not distributed normally, then there is a need to sample the data and finding the distribution of its means. According to Central limit theorem, the means are normally distributed. We can then proceed finding mean and standard deviation of that data to find confidence interval and forecasting values.

Since the analysis and its results leads to various conclusions, this creates a need for downloading different datasets for further anaylsis. Repeating the procedure of downloading by verifying the credentials manually is a tedious task. Hence there is a demand for creating a tool of automatic data collection.

The analysis and solutions are presented in the following sections.

6.1 Analysing distribution of eye width

The outcome of the previous chapters have pointed that the eye width is neither linearly nor non-linearly related in a significant manner with other variables given in the data set.

Hence the next step in our approach would be to find the distribution of eye width for the data over the years. If the data is found to be aproximately normally distributed, like the sample data in the fourth chapter, then the probability density and confidence interval could be found using the mean and standad deviation of the data.

As shown in figure 6.1 and 6.2, it is seen that the data is approximately normally distributed.

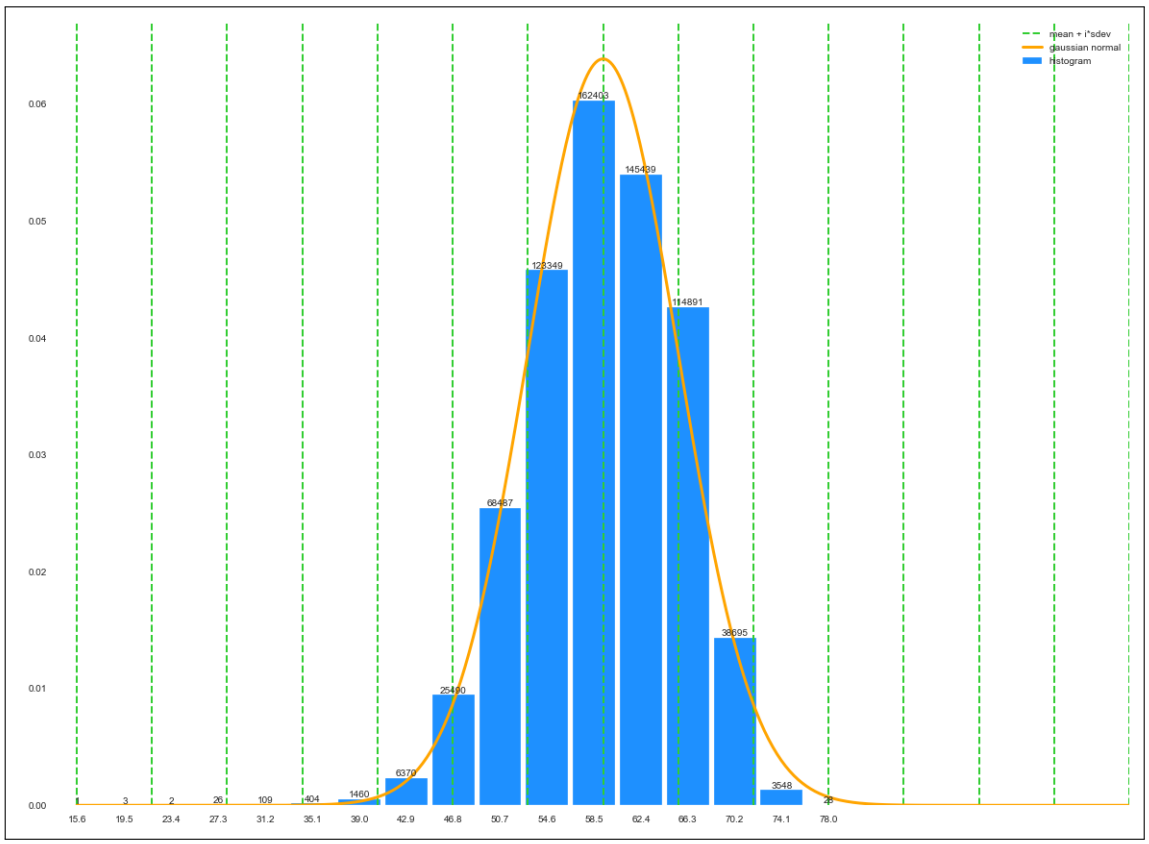

Figure 6.1: Eye width distribution of High end machines over 4 years

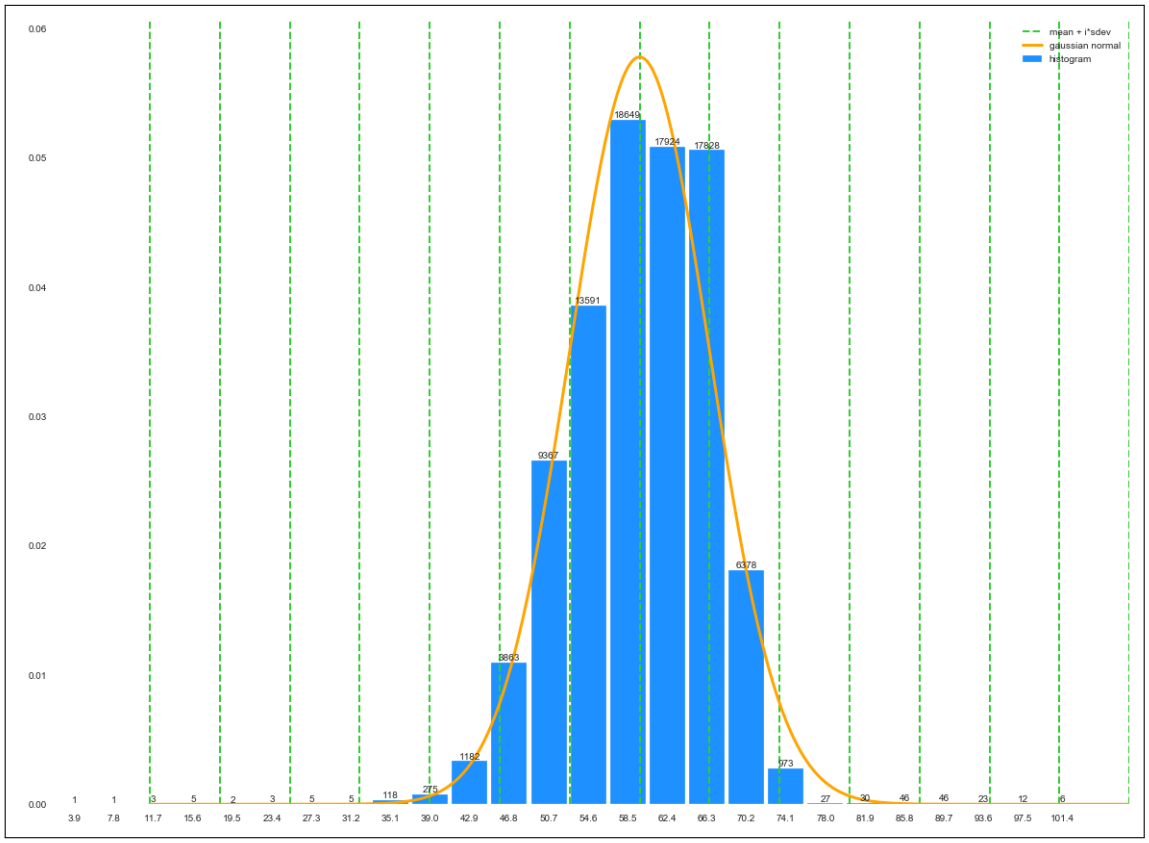

Figure 6.2: Eye width distribution of Mid range machines over 4 years

Dataset class	Duration	Period	Mean	Standard deviation
PCIe High end machines	1 month	January 2019	59.56	7.53
PCIe High end machines	1 month	February 2019	60.18	5.77
PCIe High end machines	1 month	March 2018	81.67	6.38
PCIe High end machines	1 month	July 2018	58.77	6.20
High end machines	4 years	2017-2020	59.24	6.24
Mid range machines	4 years	2017-2020	59.66	6.90

Table 6.1: Summary of Mean and Standard deviations of different Dataset

Confidence Interval:

The fig 6.1 shows that the distribution of eye width across years is approximately normal. This is only one of the sample data. The fig 4.2 is an another sample data. The true population is unknown. The difference in the means of sample data could be estimated as sample error. Since the data is approximately normally distributed, it is possible to define confidence interval for the mean value of eye width using variation within the population of sample data whch can also be called as Standard deviation, mean of sample and also size of sample[55]. The formula for 95% confidence interval states that[62]:

$$C.I = \mu \pm Sm * t \tag{1}$$

$$Sm = \sigma/\sqrt{n} \tag{2}$$

where:

μ = sample mean
t = T statistic determined by confidence interval
Sm = Standard error
n = sample size
σ = Standard deviation

Now by substituting those values from the table 6.2 for the high end machines over 4 years and sample size of 86000, we get confidence interval of population mean between 59.1983 and 59.2817. In other words we can be 95% confident that the real mean of the eye widths of those machines would lie between 59.1983 and 59.2817.

Probability density function(PDF): Now since the eye width is concluded as a result of random event. It is now possible to identify probability density funtion for that variable since the eye width is normally distributed. PDF[53][54] can be used in forecasting the values and know what is he probability of occurance of a value based on mean and standard deviation of the normal curve. This function allows us to get the probability of occurence of each value in the dataset. By applying the formula if we want to know what is the probability of occurence of some extreme value like 78.0 eye width of a high end machine, the probability comes out as 0.000662 which is very low probability of occuring. The probability density function for a normally distributed data is given by[63]:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

where:

μ = sample mean
σ = Standard deviation

Eye Width Value	Probability in Percentage
0	0
3.9	0
7.8	0
11.7	0
15.6	0
19.5	0
23.4	0
27.3	0.000013
31.2	0.000264
35.1	0.003597
39	0.033201
42.9	0.207369
46.8	0.876386
50.7	2.506112
54.6	4.84
58.5	6.34
62.4	5.62
66.3	3.37
70.2	1.36
74.1	0.375
78	0.069
81.9	0.008753
85.8	0.000744
89.7	0.000043
93.6	0.000002
97.5	0

Table 6.2: Summary of Eye width and probability of occurence of high end machines

6.2 Automating the Data collection from Data warehouse

The data was previously downloaded manually from the cloud and loaded into the Jupyter notebook for further analysis. The fig 6.3 shows the flow chart of regular process where the authorised person has to access the cloud storage by giving in the credentials and after verification, the required data set is then need to be searched and then it is downloaded to tranfer the data to juypyter notebook where further analysis is carried.

If an addition data is need to be downloaded, the process is need to be repeated which consumes much time and also makes task cumbersome. Hence there is a need of tool which makes the task automatic which saves time and also makes the task easy.

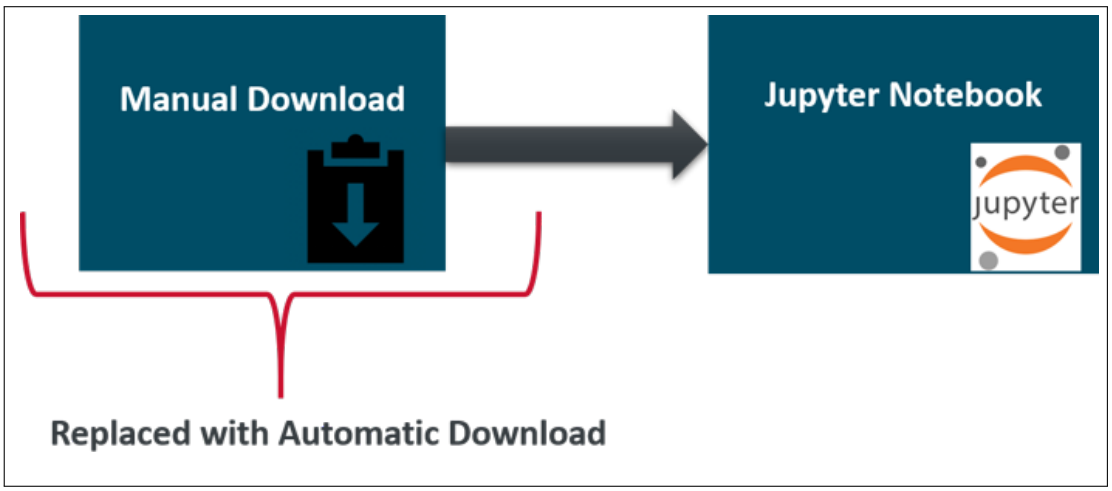

Figure 6.3: Manual downloading of data set into Jupyter

The fig 6.4 shows the flow of the process after the automation of data collection. There is availabilty of ISDW function for accessing the data from IBM cloud which is then need to be altered to provide the credentials and also the variables needed, the information of the database which our data belongs through SQL statements.

The ISDW funtion validates the credentials that we provide in the cloud. If it is valid, then the sql statments that we wrote are fed into the database to acquire the data table. If there is an error by our side, the fuction returns an error from the cloud. This process is much faster and enables us to flexibly analyse the data by varying data size , data variables, time window of data.

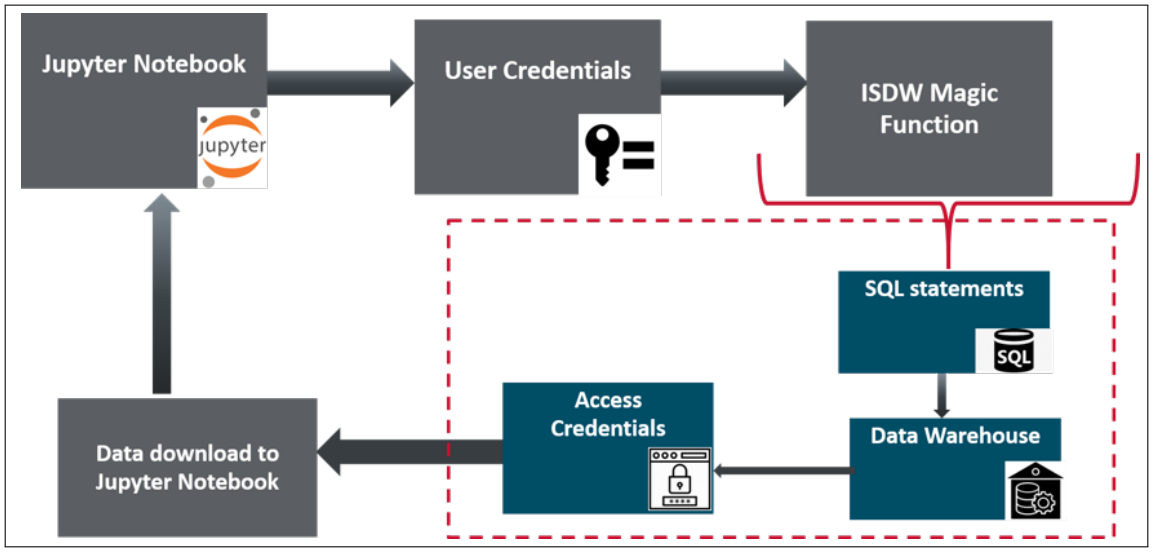

Figure 6.4: Autmated download of data set into Jupyter

6.3 Summary and conclusion for the chapter

The population data was found to be approximately normally distributed. The mean was 59.24 and standard devaition was 6.24. This lead to confidence interval of real mean lying between 59.1983 and 59.2817 if the data is grown in future. Also through probabiity function it was found that eye width of 58.5 which is close to real mean is the most probable value. The table of eye widths and its probability was also presented. This can also be viewed as forecast of values.

The following points were made regarding Automatic data downloading:

1. The automatic downloading of data set was achieved by the aid of ISDW function which autmoatically verifies the user credentials and pulls the data with sql code provided.

2. This analysis forms the scope for future work like applying central limit theorem and confidence interval for all the other variables.

3. Pulling the data automatically and analysing the data until the end result even without the need of single click can also be considered as a future work.

Chapter 7

Summary and Conclusion

The final chapter concludes the work and study accomplished in this thesis work. The first section summarizes the thesis research questions and further conclusion covers the result derived for the research questions. At last ideas on the future work are discussed.

7.1 Summary

The main research question of this thesis was to explore and find whether there is a correlation between the available data and the eye opening of the components. The process started with the collection of the data and performing the data preparation steps like cleaning redundant data, removing non useful data and also remvoing outliers. Once the data was shaped in a way that it can be used by libraries of python like Pandas, the distribution of various variables and also the distribution of eye width was found. Meanwhile the feature correlation was analysed and multicollinearity of group of variables were also checked. Some of the statistical methods like F-test and P values were performed on the variables by keeping the eye width as target variable.

Feature engineering was applied and features like Machine numbers and Chip sorting, along with classes of components based on PSRO speed were derived. After the set of useful variables were ready, with the help of domain knowledge the various combinations of variables were visualized and some relationship between them were observed. In order to distinguish the difference between the data of different eye width labels, a visalization method called TSNE plot was used.

When the direct linear relationship is not to be found between eye width and other variables, the traditional machine learning algorithms like random forest and neural networks were used on the available data. The problem was treated in two different ways. One was as a regression method and the other with classification method. Because the eye width variable could be also treated as a label and also a continuous variable.

After all the methods were tried to extract relationship between the variables, the data dis-

tribution of eye width over the span of 4 years were studied. Since it was approximately normally distributed, probability density function using mean and standard deviation was found.

At last the data collection step was replaced by an ISDW magic function which enables us to get the data through cloud when provided with relevant SQL statements and also valid credentials. The whole summary of this study is concised in Fig 7.1

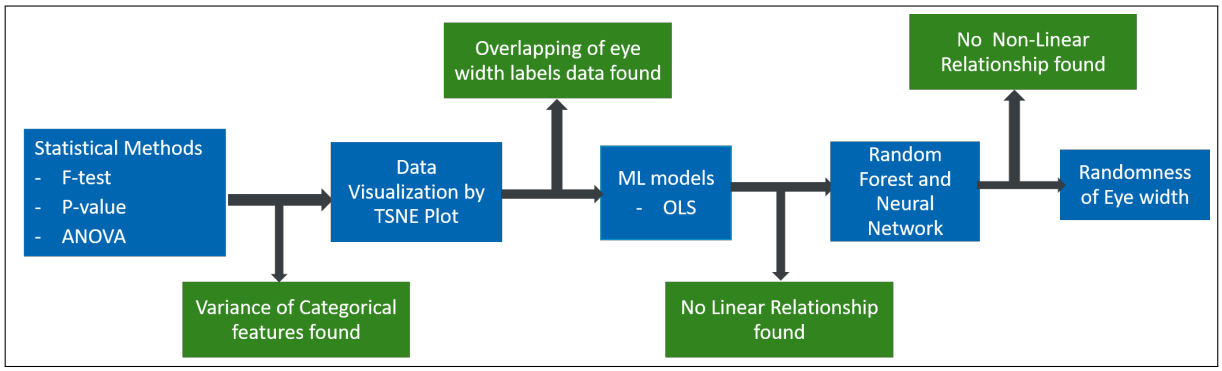

Figure 7.1: Summary of the study

7.2 Conclusion

From the initial steps like feature correlation and F-test, there was no strong relationship found between eye width and other variables. For the categorical variables, the grouping method was used and even that confirmed that each unique group of categorical variables were associated with multiple eye width labels.

Further the points can be summarized as :

1. The comprehensive visualization technique like TSNE plot was introduced, and was very helpful in determining the uncertainity between the variables of the dataset. Methods like Anova value and VIF factor were introduced and its usage in the given dataset were elaborated.

2. The drawbacks of neural network and also random forest, along with few advantages were also found. The random forest gets affected by high cardinal features and neural networks get affected by multicollinearity.

3. The concepts of Central limit theorem, probabibilty density function were discussed based on the normal distribution of eye width data.

4. The ISDW magic function was integrated to the script to automate the data collection process.

7.3 Future work

7.3.1 Central Limit theorem

"The Central Limit Theorem states that, for a large sample of n observations from a population with a finite mean and variance, the sampling distribution of the sum or mean of samples of size n is approximately normal"[61].

It means as per our analysis other than the eye width,the rest of the variables do not follow normal distribution. But according to the Central Limit theorem, when each variable is divided into different samples by boot strapping and when the means are taken for each sample and when the distribution for these means are plotted, it should follow normal distribution. By this we can then derive Probability density function and also Confidence interval for the means.

This analysis leads us to what could be the probability of occurence of each variable values when the data is forecasted.The deviation of the mean from the confidence interval value could be then investigated.

7.3.2 Automatic update of Data

The ISDW function implementation for automatic downloading of data has speeden up the process. By addition of API to this tool makes it possible to update the data in jupyter notebook as soon as the data is added in the cloud database. This makes us to know what changes in the variable has occured, has there are some outliers in the data or has the confidence interval of the data got affected or not.

Bibliography

[1] Andy Liaw and Matthew Wiener (December 2002):Classification and Regression by randomForest

[2] M. Pal (2005): Random forest classifier for remote sensing classification, International Journal of Remote Sensing, 26:1, 217-222, DOI: 10.1080/01431160412331269698

[3] Carolin Strobl*1, Anne-Laure Boulesteix2, Achim Zeileis3 and Torsten Hothorn4 (25 January 2007):Bias in random forest variable importance measures: Illustrations, sources and a solution BMC Bioinformatics 2007, 8:25 doi:10.1186/1471-2105-8-25.

[4] Maria Clara de Lacy · Mirko Reguzzoni · Fernando Sansò · Giovanna Venuti(2008):The Bayesian detection of discontinuities in a polynomial regression and its application to the cycle-slip problem

[5] Orazio Giustolisi,Dragan A. Savic(2006):A symbolic data-driven technique based on evolutionary polynomial regression

[6] Ling Huang, Jinzhu Jia, Bin Yu, Byung-Gon Chun, Petros Maniatis, Mayur Naik(2010):Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression

[7] Ryan O'Donnell , Karl Wimmer(April 2010):Polynomial regression under arbitrary product distributions Eric Blais ·

[8] Mohammad Rezaie-Balf,Ozgur Kisi(2017):New formulation for forecasting streamflow: evolutionary polynomial regression vs. extreme learning machine

[9] David C.Knill, AlexandrePouget(2004):The Bayesian brain: the role of uncertainty in neural coding and computation

[10] Donald F. Specht(NOVEMBER 1991):A General Regression Neural Network IEEE TRANSACTIONS ON NEURAL NETWORKS. VOL. 2. NO. 6.

[11] Laurens van der Maaten(2014):Accelerating t-SNE using Tree-Based Algorithms Journal of Machine Learning Research 15 3221-3245

[12] Laurens van der Maaten, Geoffrey Hinton(2008):Visualizing Data using t-SNE 9(Nov):2579–2605.

[13] Andrew F. Hayes & Jörg Matthes(2009):Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations

[14] Milan Melouna(2001):Detection of single influential points in OLS regression model building

[15] H.A. Rowley , S. Baluja , T. Kanade(1998):Neural network-based face detection

[16] G. S. Novak(Aug. 1995): "Conversion of units of measurement," in IEEE Transactions on Software Engineering, vol. 21, no. 8, pp. 651-661, doi: 10.1109/32.403789.

[17] Guo, H., Zhu, H., Guo, Z., Zhang, X.X., Su, Z (2009): Address standardization with latent semantic association. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1155–1164. ACM

[18] Van den Broeck J, Argeseanu Cunningham S, Eeckels R, Herbst K (2005) Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Med 2(10): e267. https://doi.org/10.1371/journal.pmed.0020267

[19] Hall, M.A. (2000). Correlation-based feature selection of discrete and numeric class machine learning. (Working paper 00/08). Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[20] Yu, Lei, and Huan Liu(2003): "Feature selection for high-dimensional data: A fast correlation-based filter solution." Proceedings of the 20th international conference on machine learning (ICML-03)..

[21] Ahlgren, Per, Bo Jarneving, and Ronald Rousseau.(2003): "Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient." Journal of the American Society for Information Science and Technology 54.6 : 550-560.

[22] Sedgwick, Philip.(2012): "Pearson's correlation coefficient." Bmj 345 : e4483.

[23] Schroeder, Mary Ann, Janice Lander, and Stacey Levine-Silverman.(1990): "Diagnosing and dealing with multicollinearity." Western journal of nursing research 12.2 : 175-187.

[24] Kumari, S. S.(2008): "Multicollinearity: Estimation and elimination." Journal of Contemporary research in Management 3.1 : 87-95.

[25] Alin, Aylin.(2010): "Multicollinearity." Wiley Interdisciplinary Reviews: Computational Statistics 2.3 : 370-374.

[26] Craney, Trevor A., and James G. Surles.(2002): "Model-dependent variance inflation factor cutoff values." Quality Engineering 14.3 391-403.

[27] Thompson, Christopher Glen, et al.(2017): "Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results." Basic and Applied Social Psychology 39.2 81-90.

[28] Van Buuren, Stef.(2018): Flexible imputation of missing data. CRC press,

[29] Graham, John W.(2009): "Missing data analysis: Making it work in the real world." Annual review of psychology 60 549-576.

[30] Furusjö, Erik, et al.(2006): "The importance of outlier detection and training set selection for reliable environmental QSAR predictions." Chemosphere 63.1 99-108.

[31] Hido, Shohei, et al.(2011): "Statistical outlier detection using direct density ratio estimation." Knowledge and information systems 26.2 309-336.

[32] Dunson, David & Xing, Chuanhua. (2012): Nonparametric Bayes Modeling of Multivariate Categorical Data. Journal of the American Statistical Association. 104. 1042–1051. 10.1198/jasa.2009.tm08439.

[33] Micci-Barreca, Daniele. (2001). A Preprocessing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems.. SIGKDD Explorations. 3. 27–32. 10.1145/507533.507538.

[34] Protassov, Rostislav, et al.(2002): "Statistics, handle with care: detecting multiple model components with the likelihood ratio test." The Astrophysical Journal 571.1 545.

[35] Kanioura, Athina, and Paul Turner*.(2005): "Critical values for an F-test for cointegration in a multivariate model." Applied Economics 37.3 265-270.

[36] Anderson, David R., Kenneth P. Burnham, and William L. Thompson.(2000): "Null hypothesis testing: problems, prevalence, and an alternative." The journal of wildlife management 912-923.

[37] Haug, Marie R.(1972): "Deprofessionalization: an alternate hypothesis for the future." The Sociological Review 20.1_suppl 195-211.

[38] Olson, Chester L.(1976): "On choosing a test statistic in multivariate analysis of variance." Psychological bulletin 83.4 579.

[39] Sullivan, Gail M., and Richard Feinn.(2012): "Using effect size—or why the P value is not enough." Journal of graduate medical education 4.3 279-282.

[40] Hung, HM James, et al.(1997): "The behavior of the p-value when the alternative hypothesis is true." Biometrics 11-22.

[41] Hoecker, Andreas, et al. (2007). "TMVA-toolkit for multivariate data analysis." arXiv preprint physics/0703039

[42] St, Lars, and Svante Wold.(1989): "Analysis of variance (ANOVA)." Chemometrics and intelligent laboratory systems 6.4 259-272.

[43] Gelman, Andrew.(2005): "Analysis of variance—why it is more important than ever." The annals of statistics 33.1 1-53.

[44] Wattenberg, Martin, Fernanda Viégas, and Ian Johnson.(2016): "How to use t-SNE effectively." Distill 1.10 e2.

[45] R. Shreyas, D. M. Akshata, B. S. Mahanand, B. Shagun and C. M. Abhishek(2016): "Predicting popularity of online articles using Random Forest regression," 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), Mysore, pp. 1-5, doi: 10.1109/CCIP.2016.7802890.

[46] Shanker, Murali, Michael Y. Hu, and Ming S. Hung.(1996): "Effect of data standardization on neural network training." Omega 24.4 385-397.

[47] Deng, Houtao, and George Runger.(2013): "Gene selection with guided regularized random forest." Pattern Recognition 46.12 3483-3489.

[48] Menze, Bjoern H., et al.(2009): "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data." BMC bioinformatics 10.1 213.

[49] Baker, David P., Christopher F. Chabris, and Stephen M. Kosslyn.(1999): "Encoding categorical and coordinate spatial relations without input-output correlations: New simulation models." Cognitive science 23.1 33-51.

[50] Famili, A., et al.(1997): "Data preprocessing and intelligent data analysis." Intelligent data analysis 1.1 3-23.

[51] Turner, C. Reid, et al.(1999): "A conceptual basis for feature engineering." Journal of Systems and Software 49.1 3-15.

[52] Keim, Daniel A.(2001): "Visual exploration of large data sets." Communications of the ACM 44.8 38-44.

[53] Noreen, Eric.(1988): "An empirical comparison of probit and OLS regression hypothesis tests." Journal of Accounting Research 119-133.

[54] Parzen, Emanuel.(1962): "On estimation of a probability density function and mode." The annals of mathematical statistics 33.3 1065-1076.

[55] Vachaud, G., et al.(1985): "Temporal stability of spatially measured soil water probability density function." Soil Science Society of America Journal 49.4 822-828.

[56] Brookmeyer, Ron, and John Crowley.(1982): "A confidence interval for the median survival time." Biometrics 29-41.

[57] IBM z-15 Overview: https://www.ibm.com/products/z15. Retrieved 26.8.2020

[58] PCI Express: https://en.wikipedia.org/wiki/PCI_Express. Retrieved 26.8.2020

[59] Strobl C, Boulesteix AL, Zeileis A, Hothorn T.(2007): Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics. ;8:25. doi:10.1186/1471-2105-8-25

[60] Garg, Akhil, and Kang Tai.(2013): "Comparison of statistical and machine learning methods in modelling of data with multicollinearity." International Journal of Modelling, Identification and Control 18.4 295-312.

[61] Anderson, C.-J.:(2010): Central limit theorem. The Corsini Encyclopedia of Psychology, John Wiley & Sons, Inc.

[62] "Confidence Intervals". www.stat.yale.edu. Retrieved 26.08.2020

[63] https://sites.nicholas.duke.edu/statsreview/continuous-probability-distributions. Retrieved 26.08.2020

Declaration of Authorship

I hereby declare that this master thesis was independently composed and authored by myself.

All contents, figures, information and ideas drawn directly or indirectly from external sources are acknowledged and fully cited. All sources and materials that have been used are referred to in this thesis.

Place, Date

Signed :Chinmay D Hegde