**INTERNSHIP REPORT**

*A report submitted in partial fulfillment of the requirements for the Award of Degree of*

### Integrated M.Tech

**in**

### ARTIFICIAL INTELLIGENCE

**by**

**BHAVYA MANOJ VOTVAT**

**Regd. No.: 19MIM10112**

**Under Supervision of**

**Dr. Pranshu Pranjal**

**Assistant Professor**

**(Duration: 30.01.2023 to 31.03.2023)**



**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

**VIT BHOPAL UNIVERSITY**

### April -2023

**SCHOOL OF COMPUTING SCIENCE AND ENGINEERING**

VIT BHOPAL UNIVERSITY



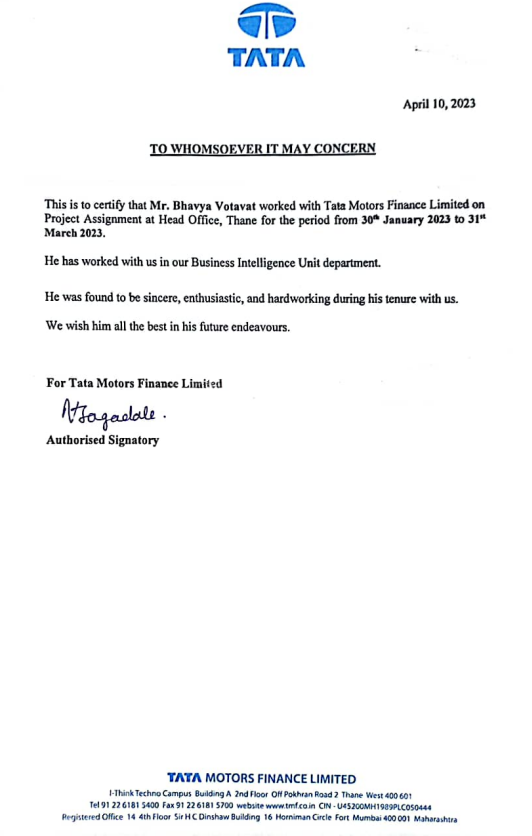
***CERTIFICATE***

 This is to certify that the **“Internship report”** submitted by  **BHAVYA MANOJ VOTAVAT  (Regd. No.: 19MIM10112)** is work done by him and submitted during 2022-2023 academic year, in partial fulfilment of the requirements for the award of the degree of **INTEGRATED M.TECH in ARTIFICIAL INTELLIGENCE** at **Tata Motors Finance Limited, Thane**.

**College Internship Guide Program Chair Dean**

Dr.Pranshu Pranjal Dr.S.Periyanayagi Dr.S.Poonkuntran Assistant Professor Associate Professor Professor

**CERTIFICATION**



# 

# **ACKNOWLEDGEMENT**

## **First, I would like to thank Ms. Pratibha Patil the Director of** Tata Motors Finance Limited, Thane **for giving me the opportunity to do an internship within the organization.**

I also would like all the people that worked along with me **Tata Motors Finance Limited, Thane** with their patience and openness they created an enjoyable working environment.

It is indeed with a great sense of pleasure and immense sense of gratitude that I acknowledge the help of these individuals.

I am highly indebted to Dean **Dr.S.Poonkuntran**, for the facilities provided to accomplish this internship.

I would like to thank my Program Chair **Dr.S.Periyanayagi** , for his constructive criticism throughout my internship.

I would like to thank Dr. **Pranshu Pranjal** for his support and advices to get and complete internship in above said organization.

I am extremely grateful to my department staff members and friends who helped me in successful completion of this internship.

### BHAVYA MANOJ VOTAVAT

### (19MIM10112)

**ABSTRACT**

Data scientists often develop machine learning models to solve a variety of problems in the industry and academy but not without facing several challenges in terms of Model Development. The problems regarding Machine Learning Development involve the fact that such professionals do not realize that they usually perform ad-hoc practices that could be improved by the adoption of activities presented in the Software Engineering Development Lifecycle. Of course, since machine learning systems are different from traditional Software systems, some differences in their respective development processes are to be expected. In this context, this paper is an effort to investigate the challenges and practices that emerge during the development of ML models from the software engineering perspective by focusing on understanding how software developers could benefit from applying or adapting the traditional software engineering process to the Machine Learning workflow

Supervised machine learning is the construction of algorithms that are able to produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances. Supervised machine learning classification algorithms aim at categorizing data from prior information. Classification is carried out very frequently in data science problems. Various successful techniques have been proposed to solve such problems viz. Rule-based techniques, Logic-based techniques, Instance-based techniques, stochastic techniques. This paper discusses the efficacy of supervised machine learning algorithms in terms of the accuracy, speed of learning, complexity and risk of over fitting measures. The main objective of this paper is to provide a general comparison with state of art machine learning algorithms.

In supervised learning, the algorithm “learns” from the training dataset by iteratively making predictions on the data and adjusting for the correct answer. While supervised learning models tend to be more accurate than unsupervised learning models, they require upfront human intervention to label the data appropriately. For example, a supervised learning model can predict how long your commute will be based on the time of day, weather conditions and so on. But first, you’ll have to train it to know that rainy weather extends the driving time.

Unsupervised learning models, in contrast, work on their own to discover the inherent structure of unlabeled data. Note that they still require some human intervention for validating output variables. For example, an unsupervised learning model can identify that online shoppers often purchase groups of products at the same time. However, a data analyst would need to validate that it makes sense for a recommendation engine to group baby clothes with an order of diapers, applesauce and sippy cups.

The semi-supervised learning has been widely applied in many fields such as medical diagnosis, pattern recognition. The semi supervised learning methods are used to employ unlabelled data in addition to labelled data for better classification of large data sets, where only a small number of labelled examples is available. Ensemble Methods are considered as an effective solution to the problem of dimensionality and can improve the robustness and generalization ability of individual learners. In this paper, we are particularly interested in the overall algorithm Random Forest semi-supervised named Co-Forest for the classification of large biological data. The algorithm is evaluated on its ability to correctly predict the labels of unlabelled examples, and its robustness when the number of labelled examples available decreases.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

A random forest (RF) classifier is an ensemble classifier that produces multiple [decision trees](https://www.sciencedirect.com/topics/computer-science/decision-trees), using a randomly selected subset of training samples and variables. This classifier has become popular within the [remote sensing](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/remote-sensing) community due to the accuracy of its classifications. The overall objective of this work was to review the utilization of RF classifiers in remote sensing. This review has revealed that the RF classifier can successfully handle high data dimensionality and multicollinearity, being both fast and insensitive to overfitting. It is, however, sensitive to the sampling design. The variable importance (VI) measurement provided by the RF classifier has been extensively exploited in different scenarios, for example, to reduce the number of dimensions of [hyperspectral data](https://www.sciencedirect.com/topics/computer-science/hyperspectral-data), to identify the most relevant multisource remote sensing and geographic data, and to select the most suitable season to classify particular target classes. Further investigations are required into less commonly exploited uses of this classifier, such as for sample proximity analysis to detect and remove outliers in the training samples.

**The Company:**

Tata Motors Finances Limited, Thane is based on landing loan to Heavy Vehicle, whether it is new or rented to only Tata Company vehicles. PV, Ace, MHCV, etc are some of their Produces. The company is divided into 3 parts Data Risk Analysis; Collection and Follow-Up Team; Legal Action Team. DRA team has sub 3 teams in which 10-15 employees in each Department complete the work and pass it to the other department on which other department works. The company provides loans to those who do not get loans from the Bank but at higher interest rates

**The Problem or Opportunity:**

As a role of Data Science Engineer in this Company, my role was to perform on this Product by giving Predictive analysis and Building a Model through ML using the Random Forest algorithm (used in this) or Logistics, and SVM Algorithms which in future prediction for the new data. The Company works on the 2 scorecards. Application Scorecard and Collection Scorecard. A model I developed using Random Forest in Application Scorecard for prediction in different months data of PV product. The problem was the outdated Model which the company was using which in turn became an opportunity as building a new optimize model for the company.

**Methodology:**

My Project aims to make an Optimize Model for the Company to give accurate and better results, in this, I used Random Forest as my main Algorithm, other than this I build the model with Logistic and SVM algorithms but Random forest was the most Optimized Model. Cleaning the Data, checking Null values and removing unwanted data, and then training data to our model, this was all the processes done to build the Model. Random forests are great with high-dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features.

**Key parts of the report & your findings and solutions provided in the report:**

While working on my Project I found the Model previously used by Company was less efficient, So I proposed a new Model with greater accuracy and more efficiency which helps Company for better results on other Data Also found out and made changes in the hyperparameter and made changes with Hyperparameter tuning which lead to better modification in the Model for better results.

**Benefits to the company/institution through your report:**

The Institute combines pioneering research with top-class education. A friendly environment that allows connecting with other capable engineers to learn different skill sets and mindsets for approaching different problems. The company also benefits from my different ideas of me for the project and learning and results with help them to get more accurate results

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sl.No** | **Contents** | **Pg.No** |
| 1 | Learning Objectives / Internship Objectives | 1 |
| 2 | Weekly overview of Internship activities | 2 |
| 3 | Introduction | 6 |
| 4 | Internship Discussion | 8 |
| 5 | Conclusion | 22 |
|  | Bibliography | 23 |

1. **LEARNING OBJECTIVES / INTERNSHIP OBJECTIVES**

🡪Internships are generally thought of to be reserved for college students looking to gain experience in a particular field. However, a wide array of people can benefit from Training Internships in order to receive real world experience and develop their skills.

🡪An objective for this position should emphasize the skills you already possess in the area and your interest in learning more

🡪Internships are utilized in a number of different career fields, including architecture, engineering, healthcare, economics, advertising and many more.

🡪Some internships are used to allow individuals to perform scientific research while others are specifically designed to allow people to gain first-hand experience working.

🡪Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

1. **WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES**

|  |  |  |  |
| --- | --- | --- | --- |
| **1st**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 30/01/23 | Monday | Reporting at office with all Photocopies of Documents.  Overview to Company Profile & Total Internship Schedule |
| 31/01/23 | Tuesday | Introduction to their product and understanding |
| 01/02/23 | Wednesday | Understanding and studying their product |
| 02/02/23 | Thursday | Understanding and studying their product |
| 03/02/23 | Friday | Understanding and studying their product |
| 04/02/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **2nd**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 06/02/23 | Monday | Introduction to new learning and work to do in the internship |
| 07/02/23 | Tuesday | Introduction to new learning and work to do in the internship |
| 08/02/23 | Wednesday | Introduction to new learning and work to do in the internship |
| 09/02/23 | Thursday | Getting a laptop and setting all important tools and downloads  In the laptop and setting all sites access. |
| 10/02/23 | Friday | Getting a laptop and setting all important tools and downloads  In the laptop and setting all sites access |
| 11/02/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **3rd**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 13/02/23 | Monday | Introduction to some work with learning |
| 14/02/23 | Tuesday | Introduction to some work with learning |
| 15/02/23 | Wednesday | Introduction to some work with learning |
| 16/02/23 | Thursday | Introduction to some work with learning |
| 17/02/23 | Friday | Introduction to some work with learning |
| 18/02/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **4th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 20/02/23 | Monday | Data is provided and work is assigned |
| 21/02/23 | Tuesday | Learning about Data and how to work on it |
| 22/02/23 | Wednesday | Learning about Data and how to work on it |
| 23/02/23 | Thursday | Learning about Data and how to work on it |
| 24/02/23 | Friday | Learning about Data and how to work on it |
| 25/02/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **5th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 27/02/23 | Monday | Cleaning Data |
| 28/02/23 | Tuesday | Analysing Data |
| 01/03/23 | Wednesday | Appling different concepts on Data to find different  Outputs |
| 02/03/23 | Thursday | Appling different concepts on Data to find different  Outputs |
| 03/03/23 | Friday | Appling different concepts on Data to find different  Outputs |
| 04/03/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **6th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 06/03/23 | Monday | Data Processing |
| 07/03/23 | Tuesday | Holi Holiday |
| 08/03/23 | Wednesday | Holi Holiday |
| 09/03/23 | Thursday | Data Processing |
| 10/03/23 | Friday | Data Processing |
| 11/03/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **7th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 13/03/23 | Monday | Model Building |
| 14/03/23 | Tuesday | Model Building |
| 15/03/23 | Wednesday | Model Building |
| 16/03/23 | Thursday | Finding all required results |
| 17/03/23 | Friday | Finding all required results |
| 18/03/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **8th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 20/03/23 | Monday | First cut results |
| 21/03/23 | Tuesday | First cut results |
| 22/03/23 | Wednesday | Optimized Results |
| 23/03/23 | Thursday | Optimized Results |
| 24/03/23 | Friday | Optimized Results |
| 25/03/23 | Saturday | Holiday |

|  |  |  |  |
| --- | --- | --- | --- |
| **9th**  **WEEK** | **DATE** | **DAY** | **NAME OF THE TOPIC/MODULE COMPLETED** |
| 27/03/23 | Monday | Submitting my final work & wait for approval |
| 28/03/23 | Tuesday | Making Project Report |
| 29/03/23 | Wednesday | Making Project Report & Submitting |
| 30/03/23 | Thursday | Waiting for Experience Letter with all project review  and approval |
| 31/03/23 | Friday | Submitting Laptop of Company and other things and  Received Experience Letter for the Internship and  Last Day of internship. |
| 01/04/23 | Saturday | Internship Overs |

1. **INTRODUCTION**

Tata Motors Finances Limited, Thane is based on landing loan to Heavy Vehicle, whether it is new or rented to only Tata Company vehicles. PV, Ace, MHCV, etc are some of their Produces. The company is divided into 3 parts Data Risk Analysis; Collection and Follow-Up Team; Legal Action Team. DRA team has sub 3 teams in which 10-15 employees in each Department complete the work and pass it to the other department on which other department works. The company provides loans to those who do not get loans from the Bank but at higher interest rates

As a role of Data Science Engineer in this Company, my role was to perform on this Product by giving Predictive analysis and Building a Model through ML using the Random Forest algorithm (used in this) or Logistics, and SVM Algorithms which in future prediction for the new data. The Company works on the 2 scorecards. Application Scorecard and Collection Scorecard. A model I developed using Random Forest in Application Scorecard for prediction in different months data of PV product. The problem was the outdated Model which the company was using which in turn became an opportunity as building a new optimize model for the company.

Here I used 3 steps to prepare Data for my internship from raw data:

🡪Exploratory Data Analysis (EDA)

🡪Data preprocessing

🡪Data splitting

# 1. Exploratory Data Analysis (EDA)

# Exploratory data analysis, or EDA for short, is exactly what it sounds like, exploring your data. In this step, you’re simply getting an understanding of the data that you’re working with. In the real world, datasets are not as clean or intuitive as Kaggle datasets. Determine what the feature (input) variables are and what the target variable is. Don’t worry about determining what the **final** input variables are, but make sure you can identify both types of variables.

# 2. Data Preprocessing

Once you understand your data, a majority of your time spent as a data scientist is on this step, data preprocessing. This is when you spend your time manipulating the data so that it can be modeled properly. Like I said before, there is no universal way to go about this. Feature Imputation is the process of filling missing values. This is important because most machine learning models don’t work when there are missing data in the dataset.

# 3. Data Splitting

Last comes splitting your data. I’m just going to give a very generic framework that you can use here, that is generally agreed upon.

Typically you’ll want to split your data into three sets:

1. **Training Set**(70–80%): this is what the model learns on
2. **Validation Set**(10–15%): the model’s hyperparameters are tuned on this set
3. **Test set** (10–15%): finally, the model’s final performance is evaluated on this. If you’ve prepared the data correctly, the results from the test set **should** give a good indication of how the model will perform in the real world.

Our Model used all these steps to prepare clean Data and then fit our Model using Random Forest Algorithm. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare favorably but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression.

1. **INTERNSHIP DISCUSSION**

**MODEL USED**

**🡪RANDOM FOREST**

* **What is Random Forest?**

Random Forest is a powerful and versatile**supervised machine learning algorithm** that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems.

* **How Random Forest works**

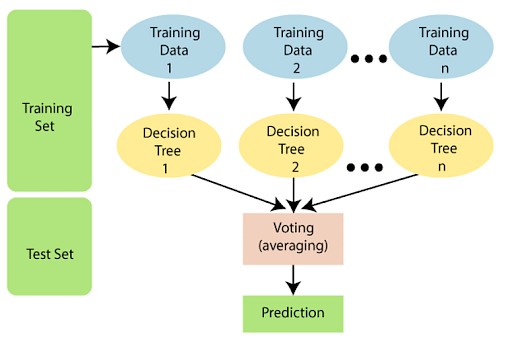


Figure 4.1

Step 1: Select random samples from a given data or training set.

Step 2: This algorithm will construct a decision tree for every training data.

Step 3: Voting will take place by averaging the decision tree.

Step 4: Finally, select the most voted prediction result as the final prediction result.

* **Why Use of Random Forest?**

There are a lot of benefits to using Random Forest Algorithm, but one of the main advantages is that it reduces the risk of overfitting and the required training time. Additionally, it offers a high level of accuracy. Random Forest algorithm runs efficiently in large databases and produces highly accurate predictions by estimating missing data.

* **Why Random Forest Model is used in our Project?**

Random forests are great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features.

We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher it’s accuracy and problem-solving ability.

Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

* **How the objectives achieved?**

Building a More Accurate Model with Random Forest than the Previous one which the company had made me Achieved My Objective.

Based on given Data by the Company, Analyzing the Data, finding missing and unwanted value and variables & eliminating them, cleaning the data, and then Splitting it into Train and Test data to Fit our Random Forest Model used to find the predictive Analysis and Other Results for the same which show that the customer will pay or not in the coming month.

**OBJECTIVE OF PROJECT**

Fitting Random Forest Model to our Data, for finding results as follow:

🡪Confusion Matrix

🡪Recall

🡪Precision

🡪F1-score

🡪Accuracy

For both Train and Test data

* **What skills (scientific and professional) were learned during the internship?**

I learned many skills during this internship, the whole process of Data Scientist or Data Analyst to approach a real-world Problem I learn in this internship.

**SKILLS:**

# Data Preparation

# Exploratory Data Analysis (EDA):

🡪Feature and Target Variables

🡪Data Types

🡪Check for Outliers

1. Z-score/standard deviations

2. Interquartile Range (IQR)

* Data Pre-processing:

🡪Feature Imputation:

Single value imputation, Multiple value imputation, K-Nearest neighbour, Deleting the row, random imputation, moving window, most frequent, etc…

🡪Feature Encoding

Label Encoding, One Hot Encoding (aka. get dummies)

🡪Feature Normalization, Feature standardization

🡪Feature Engineering

🡪Feature Selection

Feature importance, Dimensionality reduction

🡪Dealing with Data Imbalances

* Data Splitting

🡪Training Set

🡪Validation Set

🡪Test set

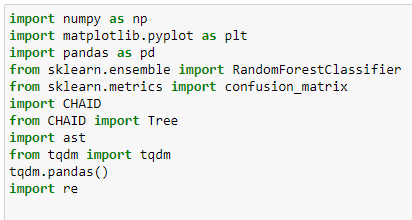
Also, many deep learning skills I learned in this internship, such as importing and reading data or importing and building Models on different and big datasets. Also, I learned how to find many things null points, which data to keep and remove, Confusion Matrix, F1-score, Accuracy, Recall, and Precision. Other than these skills, I learned how to work under pressure, cooperate with seniors and learn from them, work within the given time, and finish the work and make it right in the given time. So overall, I learned many things in this Internship.

* **Source Code:**

The Model Used here is Random Forest Modelling

**CODE:**

1. **Libraries imported**

****

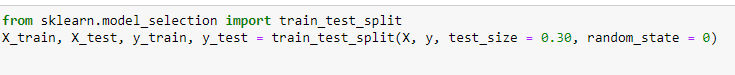
1. **Data Reading**

**Screenshot (71)**

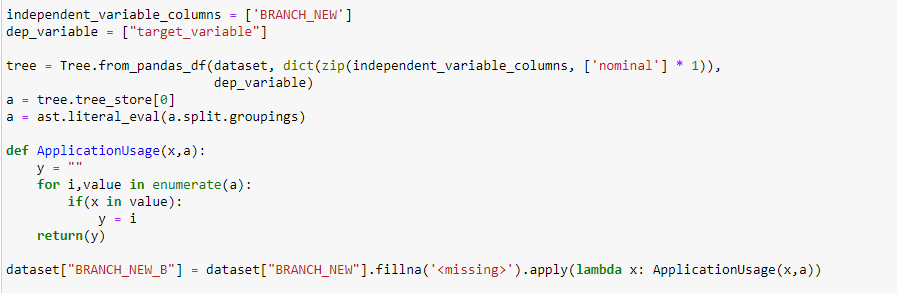
1. **Data Frame**

**Screenshot (59)**

1. **Splitting Train and Test Data**

****(Depends on each individual what hyperparameters value provides and train test splits they want.)

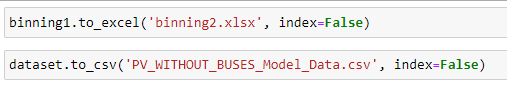
1. **Binning Code**

****

****

****

1. **Saving Bins to Excel and to Data Set**

****

1. **Random Forest Classifier**

**Screenshot (41)**

(Hyper parameters values depend on the individual what to provide, here it’s an example of code used in one of the cases to Classify Random Forest Model.)

1. **Random Forest Model Fitting**

**Train**

**Screenshot (86)**

**Test**

**Screenshot (84)**

1. **Y Prediction Variable assigning**

**Train**

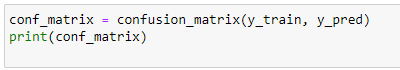
**Screenshot (85)**

**Test**

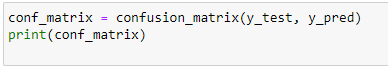
**Screenshot (83)**

1. **Confusion Matrix**

**Train**

****

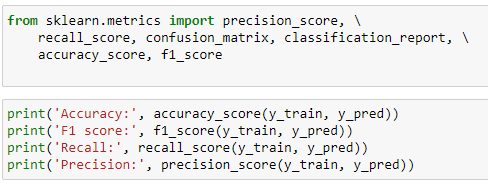
**Test**

****

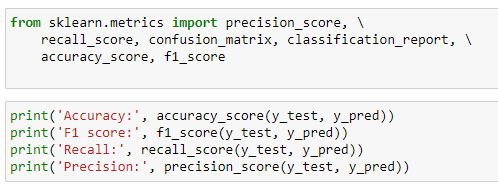
1. **Other Results**

**Library Importing & Code:**

**Train:**

****

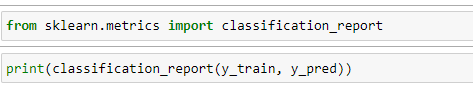
**Test:**

****

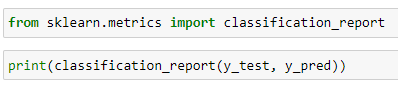
1. **Classification Report**

**Library Importing & Code:**

**Train**

****

**Test**

****

* **Results/output/observations/work experiences get in the internship:**

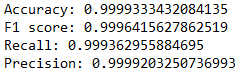
**FIRST CUT RESULTS**

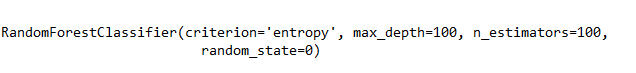
**Train Results:**

1. **Confusion Matrix Results:**

**Screenshot (52)**

1. **Accuracy, F1 score, Recall, Precision**

****

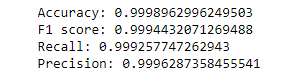
1. **Parameters use in Random Forest:** 

**Test Results:**

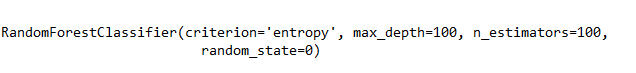
1. **Confusion Matrix:**

**Screenshot (46)**

1. **Accuracy, F1 score, Recall , Precision:**

****

1. **Parameters used in Random Forest:**

****

These were My first cut results, which were not the desired outcome.

After the Changes, I got Optimized and desired outcomes.

**OPTIMIZED RESULTS**

**Here, results are optimized and improved.**

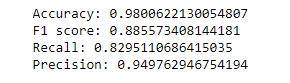
**Hyperparameter tuning is applied and also unwanted and correlated variables are eliminated to get more improved results.**

**Train Results:**

1. **Confusion Matrix Results:**

**Screenshot (43)**

1. **Accuracy, F1 score , Recall , Precision**

****

1. **Parameters use in Random Forest:**

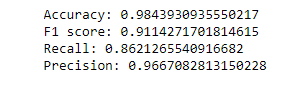
**Screenshot (41)**

**Test Results:**

1. **Confusion Matrix:**

**Screenshot (38)**

1. **Accuracy, F1 score , Recall , Precision:**

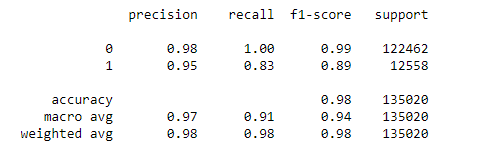


1. **Parameters used in Random Forest:**

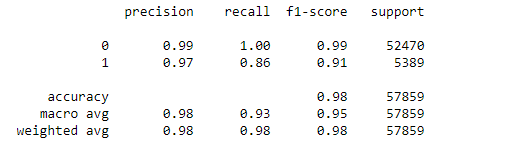
**Screenshot (41)**

**CLASSIFICATION REPORT RESULTS**

**Train Results:**

****

**Test Results:**

****

* **What challenges did you experience during the internship?**

Firstly, as I was new at the internship and I didn’t know how to work on different Big data that was a big challenge for me. Also, as I don’t know others it was difficult for me to learn new things without help and peer pressure was there to perform in the internship.

Also, the system on which I worked was the company’s laptop which has many complications and required different access to the internet with permission, even if no other app or anything can’t be downloaded directly. Another way with administration permission required to do all.

Also, there was a reputation with other seniors and the head thing that I would know everything and work was given like a bunch of pills which myself has to learn from my personal laptop on the internet as there was no such access to other sources and do it all by my own. Seniors many a time complain, I don’t know anything and then show a little which I learn myself to give the results. Asking my friends, internet and other employee works with me and get learn how to do and then perform and find solutions was a big challenge.

After the first Cut result, as can be seen in the Result section, I didn’t get the desired outcome which indeed was a big failure. I would find the solution on 100 different sources and use Hyperparameter tuning and delete unwanted data for model building knowing that the variable is least important to the company after asking seniors and then finding the Optimized solution which was a big challenge that I finally overcome just before the internship period was about to over.

These were all the major challenges that every new intern for the first-time doing internship will face and I did all that, it was a good learning experience.

**5. CONCLUSION**

Random Forest Modelling helps to predict the Outcome of train data and test Data and their correlation which lets us know the Customer to will pay in the coming Month. The Optimized Model is built for the company to predict the outcome of the new dataset.

The Project “**Propensity to Pay – Analysis & Random Forest Modelling**” was successfully built and tested. Integrating features of all the components used in Model building has developed it. The presence of every module has been reasoned out and placed carefully thus contributing to the best working of the Model. Secondly, using highly advanced Random Forest Algorithm and with the help of growing technology the project has been successfully implemented.

**BIBLIOGRAPHY**

The following books are referred during the analysis and execution phase of the project

* **Random Forest – Springer Book of AI** by Leo Breiman
* **The random forest algorithm for statistical learning – SAGE Journal** by Matthias Schonlau and Rosie Yuyan Zou

**REFERENCES:**

* Ing E., Su W., Schonlau M., and Torun N. 2019. Support vector machines and logistic regression to predict temporal artery biopsy outcomes. *Canadian Journal of Ophthalmology* 54: 116–118.

<https://journals.sagepub.com/doi/10.1177/1536867X20909688>

* Wolpert, D. H. & Macready, W. G. (1997). An efficient method to estimate Bagging's generalization error (in press, Machine Learning).

<https://cdn.aaai.org/AAAI/1999/AAAI99-047.pdf>

* Freund, Y. & Schapire, R. (1996). Experiments with a new boosting algorithm, *Machine Learning*: *Proceedings of the Thirteenth International Conference*, 148–156.

<https://www.sciencedirect.com/science/article/pii/S002200009791504X>