SUMMER INTERNSHIP REPORT

(SEM VI)

Submitted in partial fulfilment of the requirements
Of the degree of

BACHELOR OF ENGINEERING

In

INFORMATION TECHNOLOGY

By

Name: Heet Chheda

BE Roll No.: 17

Information Technology Department
Thadomal Shahani Engineering College
University of Mumbai
2020-2021

CERTIFICATE

This is to certify that the “Internship report” submitted by **Heet Chheda, Roll. No: 17** is work done by him/her and submitted during 2020 – 2021 academic year, in partial fulfilment of the requirements for the award of the degree of **“BACHELOR of** ENGINEERING**”** in **“INFORMATION TECHNOLOGY”**, Thadomal Shahani Engineering College.

Department Internship Coordinator

Dr. Shanthi Therese S. Dr. Shachi Natu

INTERNSHIP CERTIFICATE

ACKNOWLEDGMENT

The internship opportunity I had with **The Sparks Foundation** was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

Successfully completion of any type of project requires helps from a number of persons. I have also taken help from different people for the preparation of this report. Now, there is a little effort to show my deep gratitude to that helpful person.

Finally, we are very much grateful to our families who always give us constant support and encouragement. We would like to thank our seniors who helped us greatly to complete this job. In addition, we will mention our friends who also inspired and helped us to finish our work.

EXECUTIVE SUMMARY

- About the company

  The sparks foundation is working to bring parity in education, making sure children have equal opportunity at success, irrespective of the financial background. They provide students with various Mentorship Program, Scholarship programs, Workshops and Corporate programs. The company values are Resilience, Commitment, Integrity, Respect, People, Training, Excellence, Quality, and Professionalism.

- Methodology

  Identifying the problem and the approach to fix the problem as a starting step, look at what you are trying to solve within a business. The first step to that is understanding the business — what is the business dealing with, what is their input, what is the final output given by the business, and what are the other factors that lead to the final output. With this information, you get a clear understanding of the business. Deduce data requirements and collection methods calculate the amount of data needed and how it will be collected. The second element of the process is data requirements and data gathering.

- My Study/Findings

  Some models are superior at solving certain problems. Neural networks, for example, have demonstrated their supremacy in computer vision and natural language processing. However, no one can guarantee that one model will always outperform the others. We don't spend much effort implementing the learning algorithms in practise. That work has previously been done by a number of frameworks and libraries. On the learning phase, the only task will be to test out a few existing architectures and calculate cross-validation to see how much better they are. Pre-processing data entails incorporating complimentary features from other datasets, cleaning it (removing “bad” rows), and extracting new features from others (feature combination or decomposition). Safeguards are required. In comparison to what it has seen throughout the learning phase, a machine learning model strives to offer the best forecast. However, some unavoidable situations do arise. The predictions will not always be correct, for a variety of reasons. As a result, an app that makes predictions for its users shouldn't rely just on the model behind it. On top of the forecasts, safeguards must be introduced to ensure their veracity.

OBJECTIVES

- To get exposure to every part of the data pipeline
  Ingest, validate, store, extract, transfer, load, clean, model, visualise, assess, and deploy data are all tasks that must be completed.

- To have the opportunity to solve a real, big problem.
  A chance to work with ill-defined objectives, jumbled datasets, and open-ended assumptions about how to solve the problem, as well as the opportunity to tackle a real-world problem.

- To get a hands-on experience while using various algorithms
  The algorithms include supervised learning, unsupervised learning, reinforcement learning, and semi-supervised learning and do feature scaling.

- To work on classification with linear and nonlinear models
  Linear models include Logistic regression and support vectoring machines. Non-Linear models include KNN, SVM, Bayes and decision tree classification.

- To get basic idea about deep learning
  It included getting familiar with tensor flow and artificial neural networks.

- To train a model on historical, labelled data (i.e., data for which the outcome is known) in order to predict the value of some quantity on the basis of a new data item for which the target value or classification is unknown.

WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES: WEEK 1

	Date	Day	Name of the topic /Module
Week #1	JUNE 1	TUESDAY	**Task 1 : Student Marks Prediction** To predict the score of a student based on # of hours studied Used Linear regression to univariate regression of independent variable Hours to predict the dependable variable Scores and further used this regression model to predict the score of a student who studies for 9.25 hrs/ day.
	JUNE 3	THURSDAY	The model validation has been evaluated with Goodness of Fitness - R2, MSE. Also tested T-test and F-test statistics to evaluate the model.
	JUNE 4	FRIDAY	**Task 2 : Prediction-usingUnsupervised-ML** Task : From the given 'IRIS' dataset predict the optimal number of clusters and represents it visually
	JUNE 7	MONDAY	Evaluate the results with your team members

Our first task was to identify different libraries we will use for task 1 and we used pandas, numpy for data manipulation, Matplotlib, seaborn module for data visualization and sklearn for modelling. After importing the dataset we proceeded with the first 5 values and printed its shape, columns and data type. For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. We will have to check for the following (i) Missing values (ii) Outliers. For that describe method was used. Further Tasks included the following parts:

1. Finding out missing values using isnull method
2. Plot a heatmap using cbar
3. Find percentile of each predictors
4. To check outliers by plotting the boxplot
5. Visualize target variable distribution
6. Make a Correlation Matrix
7. Use Linear Regression model
8. Spliting of our data into training and testing sets
9. Train our algorithm and check accuracy scores by doing ytest and ypredict.
10. Lastly predict the score for 9.25 hours.

Second task was prediction using Unsupervised ML and we chose Kmeans clustering for this dataset. The tasks were as follows:

1. Find the shape and info of dataset.

2. Find optimum number of clusters using kmeans classification
3. Choose 3 clusters and creating kmeans classifier
4. Visualize the clusters
5. Plot it for the first two columns
6. Evaluate if any outliers
7. Fine tune the results.

WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES: WEEK 2

	Date	**Day**	**Name of the topic /Module**
Week #2	JUNE 8	TUESDAY	**Exploratory Data Analysis** ● Perform 'Exploratory Data Analysis' on dataset 'SampleSuperstore' ● As a business manager, try to find out the weak areas where you can work to make more profit. ● What all business problems you can derive by exploring the data?
	JUNE 10	THURSDAY	The model validation has been evaluated with Goodness of Fitness - R2, MSE. Also tested T-test and F-test statistics to evaluate the model.
	JUNE 12	FRIDAY	**Exploratory Data Analysis-Terrorism** ● Perform 'Exploratory Data Analysis' on dataset 'Global Terrorism' ● As a security/defense analyst, try to find out the hot zone of terrorism. ● What all security issues and insights you can derive by EDA?
	JUNE 15	MONDAY	Evaluate the results with your team members

Task 3 included the following things:

1. Find the shape and info of the dataset
2. Drop the unwanted features such as country and postal code
3. Remove null values
4. Find a correlation
5. Plot a heatmap using annot
6. Analyse the data in 3 ways which include sales, profit, and discount at each level.
7. Analysing product category at all levels.
8. Geographic level analysis.
9. Discoveries include how to maximize profits along with sales improvement and improve inventory.

Task 4 includes:

1. Exploring data analysis terrorism
2. Drop features not needed for analysis
3. Need to change motive null values
4. The analysis is carried out at different levels:
    1. Terrorist Group
    2. Attack types and Claiming modes
    3. Weapons used and Suicide Attacks
    4. Most Affected Target
    5. Most Casualties
    6. Top Countries and cities
    7. Sentimental analysis based on word clouds
5. Plotting the global terrorist activities trend on a time scale
6. Find the most active group.
7. Find the type of attacks
8. Different Claiming modes used to assume responsibility of attacks.
9. Plot a time series
10. Plot number of attacks and number of casualties in a bar chat for every significant country
11. Observations include finding security issues and insights and finding hot zones of terrorism.

WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES: WEEK 3

	Date	Day	Name of the topic /Module
Week #3	JUNE 17	THURSDAY	**Exploratory Data Analysis** ● Perform 'Exploratory Data Analysis' on dataset 'Indian Premier League' ● As a sports analyst, find out the most successful teams, players and factors contributing win or loss of a team. ● Suggest teams or players a company should endorse for its products.
	JUNE 19	SATURDAY	The model validation has been evaluated with Goodness of Fitness - R2, MSE. Also tested T-test and F-test statistics to evaluate the model.
	JUNE 22	TUESDAY	**Prediction using Decision Tree Algorithm** ● Create the Decision Tree classifier and visualize it graphically. ● The purpose is if we feed any new data to this classifier, it would be able to predict the right class accordingly.
	JUNE 24	THURSDAY	Evaluate the results with your team members

The task 5 includes:

1. Exploring data analysis sports
2. Find the most successful teams, players and factors contributing to win or loss of a team.
3. Show the shape of deliveries and shape of matches.
4. Check the null values
5. Analysing the data in three different ways:
    1. Match Analysis
        - 1.1 Matches per season
        - 1.2 Most player of the match awards
        - 1.3 Most wins by team and percentage
        - 1.4 Most hosted venues
        - 1.5 Toss winning and winning chances
    2. Run Analysis
        - 2.1 Analysis of * Wide, Bye runs * leg byes, no ball * extra runs, Dismissed * runs made by batting and bowling teams
        - 2.2 Most Runs by batsman in IPL, Orange Cap
        - 2.3 Most Sixes by batsman in IPL
        - 2.4 Most Fours by batsman in IPL
        - 2.5 Most Run outs by batsman in IPL

3. Wicket Analysis
    3.1 Top Wicket takers in IPL, Purple cap
    3.2 Most catches across IPL
    3.3 Most Run outs by fielder

6. Plotting the graph of:
    1. Number of matches played in each IPL season
    2. Top 10 man of the match
    3. Number of matches won by each IPL team
    4. Top 10 wicket takers
7. Plotting Delivery metrics in Matrix format
8. The information depicted is of 9 seasons of IPL and a clear trend can be seen for the match winning combination of team members and batting strengths.
9. It was seen that Mumbai Indians played the most number of matches.

Task 6 includes:

1. Prediction using Decision Tree Algorithm
2. Importing the libraries pandas, numpy for data manipulation, matplotlib, seaborn for data visualization 3. Reading the data
4. Checking all the attributes for missing values
5. Plot the outliers of SpecialWidthCm by each species
6. Exploring Dependent/Independent variables 7. Create the split test and training data set
8. Build decision tree model:
    1. Make prediction on Test data set
    2. Showing Testing accuracy, accuracy score and classification report
    3. Plot Confusion matrix
9. Optimizing the decision tree performance
10. It shows the increase in the value of the Accuracy and the increase in accuracy with respect to number of estimators. Here the X-axis contains the number of estimators while the Y-axis contains the value for accuracy. At depth=3 we see the accuracy is increasing.

WEEKLY OVERVIEW OF INTERNSHIP ACTIVITIES: WEEK 4

	Date	Day	Name of the topic /Module
Week #4	JUNE 26	SATURDAY	**Stock Market Prediction Using Numerical and Textual Analysis** ● Create a hybrid model for stock price/performance prediction using numerical analysis of historical stock prices, and sentimental analysis of news headlines ● Stock to analyse and predict - SENSEX (S&P BSE SENSEX) ● Download historical stock prices from finance.yahoo.com.
	JUNE 29	TUESDAY	The model validation has been evaluated with Goodness of Fitness - R2, MSE. Also tested T-test and F-test statistics to evaluate the model.
	JULY 2	FRIDAY	**Timeline Analysis : Covid-19** ● Create a storyboard showing spread of Covid-19 cases in your country or any region (Asia, Europe, BRICS etc.) using Tableau, Power BI or SAP ● Use animation, timeline and annotations to create attractive and interactive dashboards and story.
	JULY 5	MONDAY	Evaluate the results with your team members

Task 7 includes:

1. Stock market prediction using numerical and textual analysis
2. Create a hybrid model for stock price/performance prediction using numerical analysis of historical stock prices, and sentimental analysis of news headlines
3. Importing stocks data from web
4. Visualising the data- open price
5. Visualising Stocks returns
6. Performing time series analysis for closing price:
    1. Performing Decomposition of time series
    2. Calculating mean and standard deviation on transformed data
    3. Perform log transform to the dataset again to make the distribution of values more linear and better meet the expectations of this statistical test.
    4. Calculating mean and standard deviation on Differential Log Transformed data
7. Analysing news dataset
8. Performing Sentiment Analysis by assigning polarity to headlines
9. Combining Stocks data and news data
10. Splitting data into Train and Test
11. Training the model

12. Sentimental Analysis helps to predict if the stock close price with either increase or decrease depending on the news on that day.

Task 8 includes:

1. Create a storyboard showing spread of Covid-19 cases in your country or any region (Asia, Europe, BRICS etc.) using Tableau, Power BI or SAP
2. Collate the data of Covid-19 spread worldwide
3. Get the top 5 countries where there are highest cases
4. Collate the Covid-19 data of India
5. Predict no of confirmed cases in India for first week of July
6. Fit a logistic curve for total (cumulative) confirmed cases in India. Then predict for first week
7. Plotting fitting metrics
8. Collate the Covid-19 data state wise
9. Plot the graph of state with highest mortality rate
10. Plot the graph of top5 states where testing is maximum
11. It helps in identifying the interesting patterns and possible reasons helping Covid-19 spread with basic as well as advanced charts.

INTRODUCTION

The Sparks Foundation (TSF) is a non-profit organization registered in India and Singapore. We envision a world of enabled and connected little minds, building the future. We aim to inspire students, help them innovate and let them integrate to build the next generation of humankind. We help the students to integrate and help each other, learn from each other, and do well together.

The Graduate Rotational Internship Program (GRIP) is the flagship program of TSF in which students, recent graduates and professionals focus on technical skills development as well as professional profile improvement on LinkedIn. The program offers a platform to connect with students and professionals from varied diversity, background, skills and countries. During Covid-19 pandemic, the format of GRIP is 1-month, unpaid and virtual internship. It is open for students and professionals from all expertise levels and backgrounds. The tasks are of beginner as well as advanced levels, and you can select which one you want to work on.

Our responsibilities included:

1. Writing codes to collect, crunch and analyze data from internal and external sources.
2. Building machines and tune learning models using R, Python and or any language/tool we were comfortable with.
3. Using BI tools such as Tableau, PowerBI to analyze data, find important patterns and design visualization dashboards.
4. Building a strong professional profile, presenting given tasks and submissions; and improving skills through various activities as part of the internship.

The mentor will share with you a Task Submission form through which you can submit your tasks. Task Submission form will be shared before the deadline giving you ample time to submit your responses. For any unforeseen reason, if there is a delay in sharing the Task Submission form, your deadline will be adjusted accordingly. However, there has never been such a situation before.

Usually, the 1st task (Linkedin profile) is mandatory for all interns, to make them have a better online presence and credibility during job search. Tips given in the PDF shared on the task slide and try to create an awesome and attractive Linkedin profile.

The internship helped us to:

1. Gain valuable work experience

The hands-on work experience interns receive is invaluable and cannot be obtained in a classroom setting, making this one of the most important benefits of internships.

2. Explore a career path

Exploring is an important part of the college experience, and internships are a great way for students to acquaint themselves with the field they are interested in.

3. Develop and refine skills

We learnt a lot about your strengths and weaknesses during an internship. Internships allow for feedback from supervisors and others who are established in the field, and offer a unique learning opportunity that you may not have again as a working adult.

4. Network with professionals in the field

In the working world, it’s all about who you know. As an intern, we will be surrounded by professionals in the industry. Internships are more than just about earning credit, getting a grade, or making money; internships provide an opportunity to learn from the people around you, ask questions, and impress.

\

INTERNSHIP DISCUSSION

The Task #1 and #2 consisted of Supervised and Unsupervised Learning:

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.
In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.
Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to find a mapping function to map the input variable(x) with the output variable(y).
In the real-world, supervised learning can be used for Risk Assessment, Image classification, Fraud Detection, spam filtering, etc.

Unsupervised learning is a type of machine learning in which models are trained using an unlabeled dataset and are allowed to act on that data without any supervision.
Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to find the underlying structure of the dataset, group that data according to similarities, and represent that dataset in a compressed format.

1. Unsupervised learning is helpful for finding useful insights from the data.
2. Unsupervised learning is much similar to how a human learns to think by their own experiences, which makes it closer to the real AI.
3. Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
4. In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

The Task #3 consisted of Exploratory Data Analysis Retail-SuperStore

(i) As a business manager, try to find out the weak areas where you can work to make more profit.

(ii) What business problems can you derive by exploring the data?

(iii) Dash Boards - explaining the charts and interpretations.

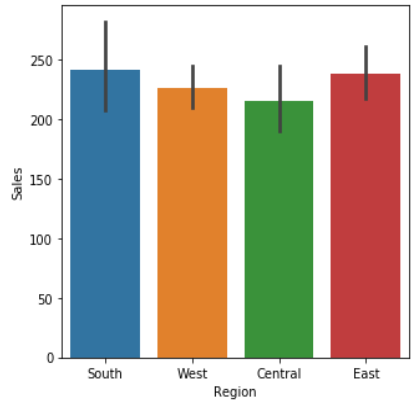
Reading Data

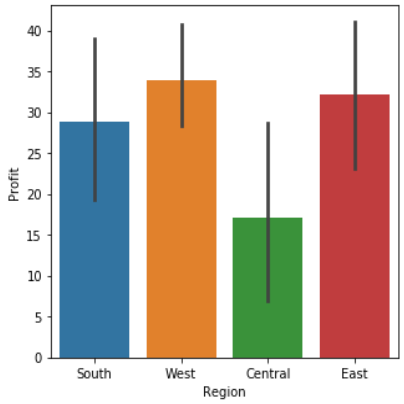
In [2]:
```
data=pd.read_csv('C:\\Users\\Keerthi\\Desktop\\04-Keerthi\\00-Spark Foundation\\SampleSuperstore.csv')
data.head()
```

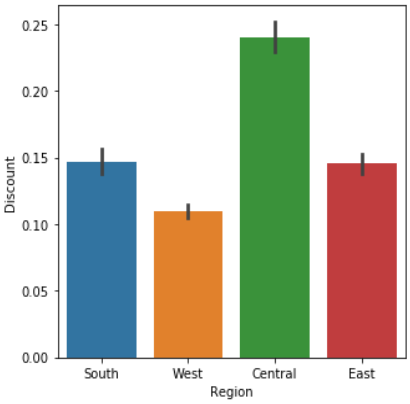
Out[2]:

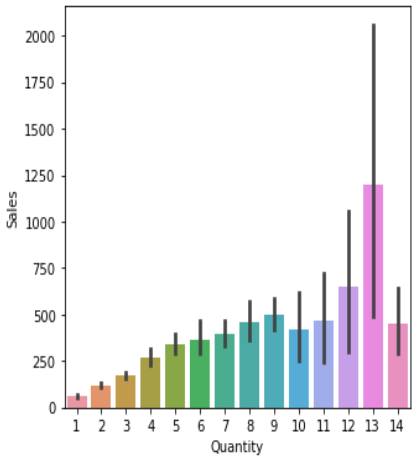
	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub-Category	Sales	Quantity	Discount	Profit
0	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	261.9600	2	0.00	41.9136
1	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	731.9400	3	0.00	219.5820
2	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	14.6200	2	0.00	6.8714
3	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	957.5775	5	0.45	-383.0310
4	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	22.3680	2	0.20	2.5164

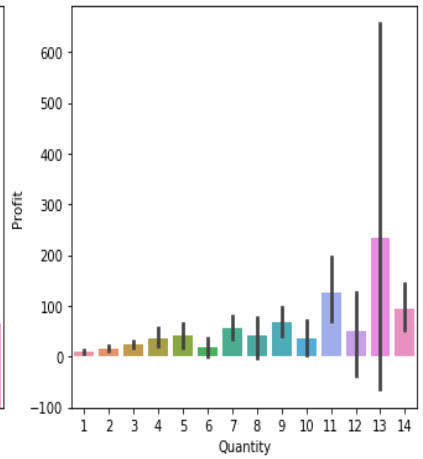
The results were as follows:

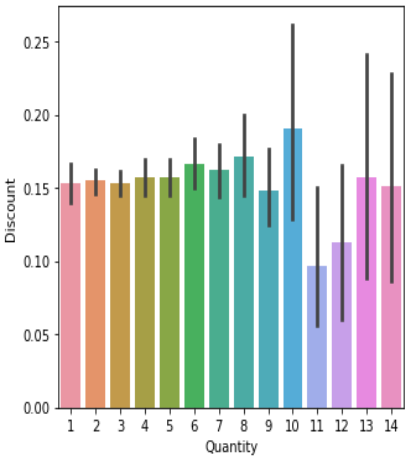

Sales
Region
South
West
Central
East

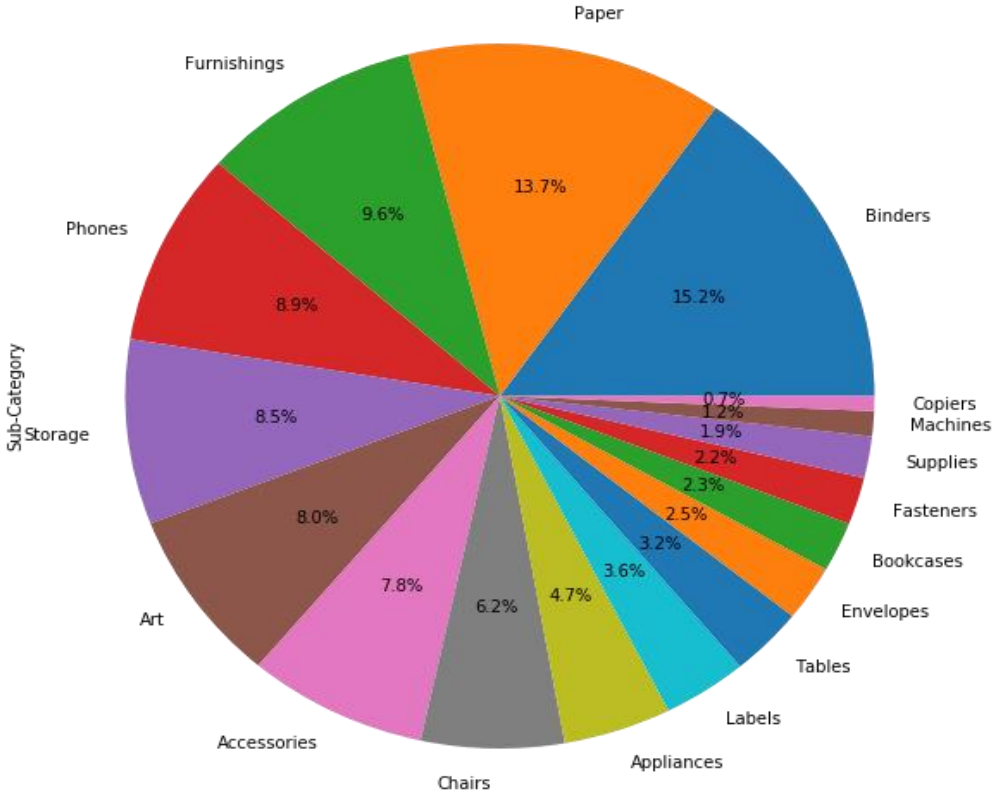

Profit
Region
South
West
Central
East


Discount
Region
South
West
Central
East


Sales
Quantity


Profit
Quantity


Discount
Quantity


Sub-Category
Paper
Furnishings
Phones
Storage
Art
Accessories
Chairs
Appliances
Labels
Tables
Envelopes
Bookcases
Fasteners
Supplies
Machines
Copiers
Binders
13.7%
9.6%
8.9%
8.5%
8.0%
7.8%
6.2%
4.7%
3.6%
3.2%
2.5%
2.3%
2.2%
1.9%
1.2%
0.7%
15.2%

The observations made by such results were:

Weak areas where managers can work to make more profit.

- Profits and Sales are not Linear in most of the states.
- Florida, Texas, Pennsylvania, Illinois, Arizona, Tennesse, Orgeon, Colorado and Ohio are the loss making states.
- Central region needs to be given more attention.
- Machines and supplies are having less profit and are a loss making subcategory.
- Furniture is not providing much profit margin.
- Higher discounts are not of much use, the profit are in negative and even the sales are having downtrend after 60% discount.
- The Office Supplies have a maximum loss at 80% and 0% discount. • Furniture and Technology had maximum loss at lower Discount rates What all business problems you can derive by exploring the data

1) How to maximise the PROFITS along with Sales improvement?

- TECHNOLOGY gives more profit compared to the furniture category.
- Profits can be maximized if the Ship mode is 'Same day'.
- Western region has more profit margins, by analyzing the marketing strategies the profit of other regions can be increased.
- Vermont State has a high profit margin even though the sales are not high, marketing strategies analysis needs to be performed.
- Discount less than or equal to 50% is having the highest sales and profit margin.
- Provide optimal Discount to Technology and Furniture to attract much customers
- The Top 5 Subcategories account to ~50% of the Sales, suggesting need for accelerated marketing Strategies or introducing additional Products in those categories.
- In Segments, Home-office has High Profit & sales, suggest to promote more for higher profits

2) Improve Inventory

- As furniture has low profit margin and have more storage cost, they can be sold on long waiting period, as people dont prefer to buy Tables and Bookcases from Superstore. Hence these departments are in loss.
- The stock of Office Supplies and Technology can be improved for the same price of Furniture storage.

The Task #4 consisted of:

1) Hot zone of terrorism.

The Hot zones of Terrorism are in
Countries :Iraq, Afghanistan ,Pakistan,India, Colombia
Cities : Baghdad, New york , Kabul,Mosul,Karachi
Regions : Middle east, North Africa, South Asia ,Sub-Saharan Africa

2) Security issues and insights:

Most of the targets are innocent people and their property followed by Military and Police who are always in duty to protect the private citizens and their property
The most active group is Taliban which is generally adopting Bombing/Explosion and Armed Assault because they are facilitated with Explosives and Firearms
the most claiming mode is personal announcement or through posting online
Suicide attacks from year 2013 till 2016 which is a worrying factor and it is likely to increase further.
We can see Iraq , Afghanistan and its neighboring countries are constantly topping the chart
By Word Cloud we can see Maosit ,Government, white, Protest are main reasons for the attacks
From sentimental analysis, the most frequent words observed are " Muslims", "Islamic". It is common belief that terrorists are Muslims but in fact the most targeted and suffered by terrorism are Muslims. Iran, Pakistan, Afghanistan are Muslim dominations.

And the Task #5 created a storyboard which consisted of:

- The data set contained two csv file
- Matches.csv : Information of all matches played in the IPL from 20082019 providing the below information.
- Delivereris.csv : Information of all balls bowled and runs scored on it .

The objectives of the entire tasks were to: As a sports analyst, find out the most successful teams, players and factors contributing to the win or loss of a team. Suggest teams or players a company should endorse for its products.
The observations were:

The information depicted is of 9 seasons of IPL and a clear trend can be seen for the match winning combination of team members and the batting strengths.
It was seen that Mumbai Indians played the most number of matches. Virat Kohli was the best batsman and has scored against some of the best bowlers The information shown for the opponents of Delhi Daredevils would include the bowlers against whom Virat performed poorly.

The top batsmen have been consistent in their performance.
Mumbai Indians have needed more games for each win while teams like CSK have needed less to win.

The Task #6 consisted of Decision Tree Project:

- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees.

The result and the interpretation of the task were as follows:

The feature importance of the features is as shown : PetalWidthCm 0.589775 PetalLengthCm 0.410225 SepalWidthCm 0.000000 SepalLengthCm 0.000000 The feature importance tells us how much a feature helped to improve the purity of all nodes. The overall importance of a feature in a decision tree can be computed in the following way: 1)Go through all the splits for which the feature was used and measure how much it has reduced the Gini index compared to the parent node. 2) The sum of all importances is scaled to 100. This means that each importance can be interpreted as a share of the overall model importance. In each node we will have information about the feature on which cutoff point the splitting is being done along with following :

1) Evaluation metric: Gini metric for classification model. Here, the variance was used, since predicting petroleum consumption is a regression task.

2) Samples : number of observations falling in the respective node.

3) Value : As this is a classification model, the predictions for the node are made as the class label of the class with maximum frequency in this region. Here we have two classes 0,1,2 . 1)The nodes with purple colour are the nodes having maximum frequency or mode of 'Iris-Versicolor' 2)The nodes with orange colour are the nodes having maximum frequency or mode of 'Iris-Virginica'. 3)The nodes with green colour are the nodes having maximum frequency or mode of 'Iris-Sentosa'. As we can see from the feature importance of "variance" , it forms the root node of the decision tree. We can also observe that the value of gini is reducing with each node(top to down). Starting from the root node, we go to the next nodes and the edges tell us which subsets you are looking at. Once you reach the leaf node, the node tells us the predicted outcome. All the edges are connected by 'AND'.

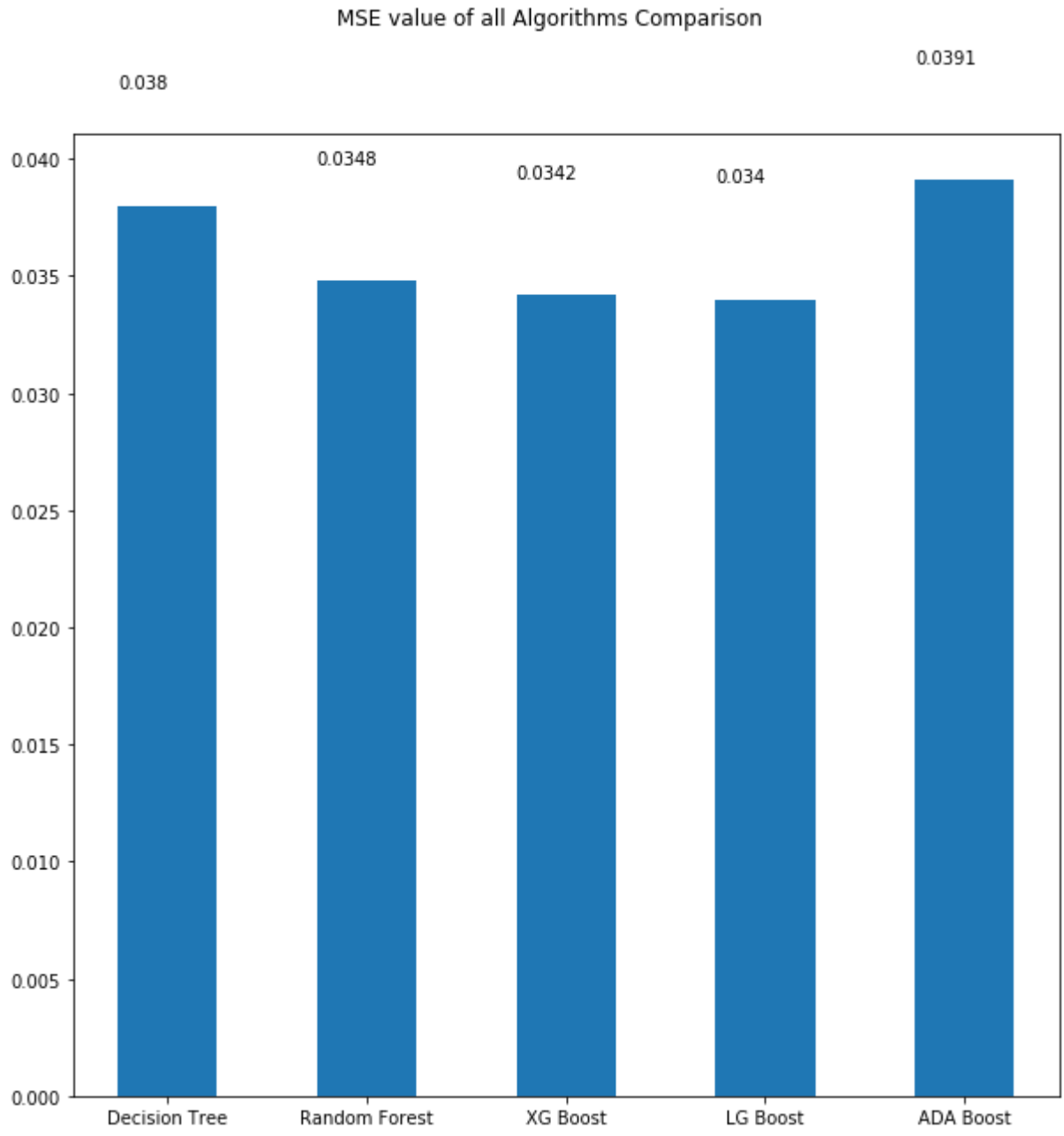
The Task #7 had a stock market analysis using sentiment analysis:

The task was to create a hybrid model for stock price/performance prediction using numerical analysis of historical stock prices, and sentimental analysis of news headlines. In this I have predicted if a company's stock will increase or decrease based on news headlines using sentiment analysis.This model will determine if the price of a stock will increase or decrease based on the sentiment of top news article headlines for the current day using Python and machine learning. I have used both numerical and textual data for this.

(i) Time series analysis is performed on the Stock data.

(ii) Sentiment analysis is performed on the News data.

(iii) An analysis is performed by merging both the data to predict if the Close price of the stock will increase or decrease.



LGBMRegressor has the least MSE and it has performed best for sentimental Analysis to predict if the stock's close price will either increase or decrease depending on the news on that day.

CONCLUSION

The internship training for 4 weeks in Data Science and Analytics was short but it has truly been a good experience for me as an ITstudent. I was given a good chance to expose myself in different types of projects which requires different requirements.I have learnt the process and stages of a project going through. It is great to have the opportunity to participate in the process. Every details in the project need to be in concern as it is in the real world.

Learning Outcomes were:

- Have an understanding of the strengths and weaknesses of many popular machine learning approaches.
- Appreciate the underlying mathematical relationships within and across Machine Learning algorithms and the paradigms of supervised and unsupervised learning.
- Be able to design and implement various machine learning algorithms in a range of real-world applications.
- Ability to integrate machine learning libraries and mathematical and statistical tools with modern technologies
- Ability to understand and apply scaling up machine learning techniques and associated computing techniques and technologies

BIBLIOGRAPHY

✓ https://www.simplilearn.com/

✓ https://www.wikipedia.org

/ ✓ https://towardsdatascience.com/

✓ https://www.expertsystem.com/

✓ https://www.coursera.org/

✓ https://www.edureka.co/

✓ https://subhadipml.tech/

✓ https://www.forbes.com/

✓ https://medium.com/

✓ https://www.analytixlabs.co.in/blog/scope-of-data-science/

✓ Hands-on Machine Learning with Scikit-learn & Tensorflow By Aurelien Geron

✓ Python Machine Learning by Sebastian Raschk