

IntimeTec VisionSoft Pvt. Ltd

PREPARED BY

Samyak Bhagat (2018BTECHCSE030)

FACULTY GUIDE

Prof. Devendra bhavsar



**Department of Computer Science & Engineering
Institute of Engineering and Technology (IET)
JK Lakshmipat University, Jaipur**

May-2022

IntimeTec VisionSoft Pvt. Ltd

Practice School-2

Submitted in partial fulfilment of the requirements

for the degree of

Bachelor of Technology in Computer Science and Engineering

By:

Samyak Bhagat (2018BTECHCSE030)

FACULTY GUIDE

Prof. Devendra bhavsar



**Department of Computer Science & Engineering
Institute of Engineering and Technology (IET)
JK Lakshmipat University, Jaipur**

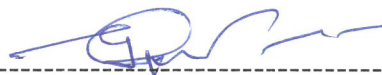
May-2022

CERTIFICATE

This is to certify that the Practice School-2 project work entitled **IntimeTec Visionsoft Pvt. Ltd, Jaipur** submitted by **Samyak Bhagat (2018BTechCSE030)**, towards the partial fulfillment of the requirements for the degree of **Bachelor of Technology in Computer Science and Engineering** of JK Lakshmipat University Jaipur is the record of work carried out by them under my supervision and guidance. In my opinion, the submitted work has reached a level required for being accepted for the Practice School-II examination.

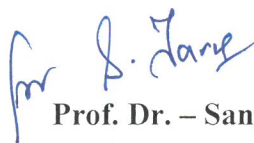


Prof. Devendra bhavsar
Associate Professor,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur



Dr. Gireesh Kaushik
Ps-2 Co-ordinator,
Department of Computer Science
Engineering
Institute of Engineering & Technology
(IET)
JK Lakshmipat University, Jaipur

Date of Submission: 27-05-2022



Prof. Dr. – Sanjay Goel
Director
Institute of Engineering & Technology
JK Lakshmipat University Jaipur

ACKNOWLEDGEMENTS

The internship opportunity I had with **IntimeTec, Jaipur** was a great chance for learning and professional development. Therefore, I consider myself a very lucky individual as I was provided with an opportunity to be a part of IntimeTec. I am also grateful for having a chance to meet so many wonderful people and professionals who led me through this internship period.

I perceive as this opportunity as a big milestone in my career development. I have taken efforts in completing this internship. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to express heartfelt gratitude to **Prof Dr.Sanjay Goel**, Director, Institute of Engineering & Technology for providing this opportunity to me. Also, I would like to thank my faculty guide **Prof. Devendra Bhavsar** for his constant support and guidance.

I am highly indebted to **InTimeTec** for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards the members of IntimeTec for their kind co-operation and encouragement which helped me in the completion of this project. I would like to express my gratitude and thanks to industry mentors **Mr. Rakesh P, Mr. Prasun, and Mr. Tarun Khandelwal** for giving me such attention and time.



Sincerely yours,

Samyak Bhagat
(2018BTechCSE030)

ABSTRACT

Data engineering is a multidisciplinary field that comprises learning, statistics, database, visualization, optimization, and information theory. It is slightly younger than its sibling, data science. It is a broad field, where a data engineer transforms data into a useful format for analysis. This paper provides a brief introduction to data engineering.

Data is everywhere and in varied forms, and *it needs to be perfected to derive actionable insights*. As analytics professionals, data engineers are responsible for generating, cleaning, processing, and storing data in a way that makes it ready for analysis.

Their ultimate goal is *to make data accessible so that organizations can use it to evaluate and optimize their performance*.

According to DICE's 2020 Tech Job Report, *Data Engineer is the fastest-growing job in 2019, growing by 50% YoY*. Data Scientist is also up there on the list, growing by 32% YoY.

TABLE OF CONTENT

Contents

ACKNOWLEDGEMENTS	2
ABSTRACT	3
CHAPTER 1: INTRODUCTION.....	1
CHAPTER 2: WORK FLOW	3
CHAPTER 3: Hackathon for Trainees.....	4
CHAPTER 4: Internal Project	12
CHAPTER 5: DATA ENGINEERING TRAINING	13
CHAPTER 6: PROJECT DEPLOYMENT	20
CHAPTER 6: CONCLUSION	22

CHAPTER 1: INTRODUCTION

1.1 About the organization



In Time Tec provides progressive software development services, enabling its clients to keep their brightest and most valuable talent focused on innovation. In Time Tec has a leadership team averaging 15 years in software/firmware R&D, and 20 years building onshore/offshore R&D teams. Every company has core and context work. As defined by author and speaker Geoffrey Moore, core work is your differentiator, your 'secret sauce'. Context work is non-core work that must get done. In Time Tec helps you focus on core work by taking on your context work. By using a blended model of onshore and offshore talent, we offer the following software development services:

It takes a special kind of person to stand with us on our platform of abundance. Our mission and our way of being drives everything we do. We want people to live their fullest potential while serving our clients and communities. It delivers a wide range of technical security services, from reviewing existing processes and documentation, and assessing the security of current environments, to the design of new secure IT systems to meet modern security requirements. It helps organizations meet their cyber compliance obligations by assisting in the development of risk compliance reviews governance activities, policies, and action plans

It can assist your organization to assess and improve your current cyber security resilience, through the use of mature security architecture, design, development, deployment, vulnerability testing, and management techniques. As cyber security threats evolve and new vulnerabilities are discovered, through an ongoing series of publications and presentations,

Intimetec Visionsoft Private Limited is a Private incorporated on 13 February 2009. It is classified as a non-govt company and is registered at the Registrar of Companies, Jaipur. Its authorized share capital is Rs. 50,000,000 and its paid-up capital is Rs. 37,730,720. It is involved in Software publishing, consultancy, and supply [Software publishing includes production, supply, and documentation of ready-made (non-customized) software, operating systems software, business & other applications software, and computer games software for all platforms. Consultancy includes providing the best solution in the form of custom software after analyzing the user's needs and problems. Custom software also includes made-to-order software based on orders from specific users. Also, included are writing of software of any kind following directives of the users; software maintenance, web-page.

Website	http:// www.intimetec.com
Industry	Software Development
Company size	500-1000 employees
Headquarters	Meridian, ID
Type	Private
Founded	2009
Specialties	Web Development, Mobile Development, Cloud Development, Salesforce Development, Embedded Development, Quality Assurance, Automated Testing, Technical Writing, DevOps, and Continuous Integration/Continuous Development

CHAPTER 2: WORK FLOW

In this chapter I will tell you about the different parts on which I worked during my time at IntimeTec

2.1 Hackathon For Trainees

During the first week the company announced a internal hackathon with the aim to select the candidates for the direct deployment to project.

The hackathon was in 2 phases: -

- Phase 1- coding Round
In this round we must solve two coding problems on HackerRank around 35 candidates were selected out of 150 trainees in this round
- Phase 2- Development Round
The Second phase the was the development round in which we must choose our tech stack according to which problem statement was provided. In this Round out of 35 around 20 candidates were selected and they were intitled

2.2 Internal Project

I was deployed to an internal project of 'Resume Shortlisting' after the hackathon in which I have to work on the existing code base and make functions. We have to talk to HR and get insights into how the process is done manually and then we have to automate the process.

2.3 Data Engineering Training

After the review of the resume shortlisting project, I was provided with the data engineering training where we must learn the new technologies so that I can be deployed to the project

2.4 Project Deployment

I am not currently working with HP Ink and deployed to the HP DataOS team who maintains the data of HP Ink business and work for the new business opportunities by providing help to market their products.

2.5 Internal Hackathon Techfoo

There is an internal hackathon named Techfoo for all employees of Intimetec. My team got selected for round 2 of that Hackathon which will be on 30-31st May 2022.

2.6 Data Science Initiative

There is a research team as well in Intimetec and we can volunteer to join that team. I have to work on convex optimization and learn it.

CHAPTER 3: Hackathon for Trainees

As I previously mentioned that the company provided us with an opportunity for hackathon so that we can directly be deployed on the project. I will talk about the development phase in the section of the report.

3.1 The Problem Statement

- Develop a Model, to predict Cost of Annual Maintenance Contract for any given printer - provided some information about the said printer is available
- Development of a Multi-linear Regression model for prediction of MaintenanceQuote price for each of the devices uniquely identified by the attribute, IdentificationId, in printer_data and MaintenanceQuote files.

3.2 Data Attributes

Maintenance Quote

- IdentificationId
- DeviceProductNumber
- MaintenanceQuote

Printer Data

- IdentificationId
- Make
- DeviceType
- MonoSpeed
- ColorSpeed
- HasA3
- HasA4
- HasColor
- HasScan
- HasFax
- PrinterPrice

3.3 Initial Challenges Faced

- Data Set was very inconsistent and many missing values
- There were many categorical data therefore it was tough to fit it into a linear regression model and that I why I have low accuracy

3.4 Regression Model

Multiple linear regression refers to a statistical technique that uses two or more independent variables to predict the outcome of a dependent variable.

$$y = f(x)$$

$$f(x_1, \dots, x_m) = W_1 x_1 + W_2 x_2 + \dots + W_m x_m + b$$

y = Dependent Variable

x = Independent Variable

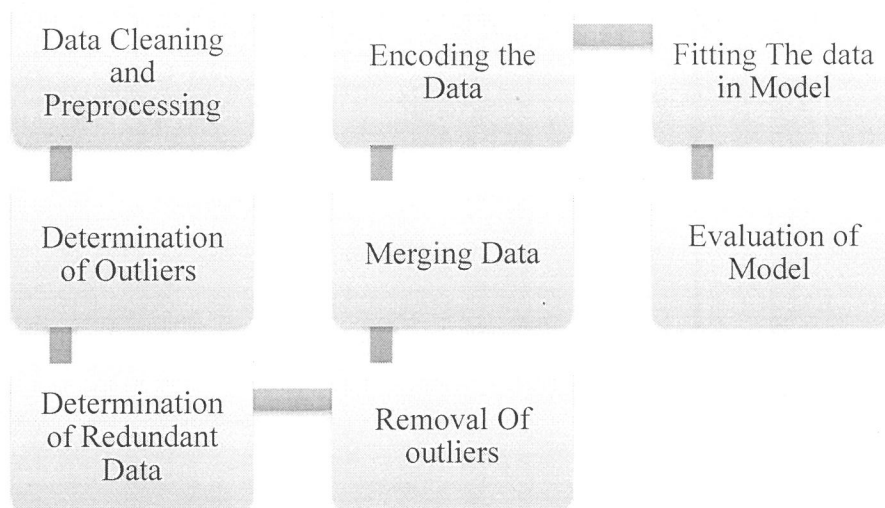
b = bias (intercept)

W = Weights (slope of x)

$\hat{y}$ → predicted Value

$y - \hat{y}$ → Error

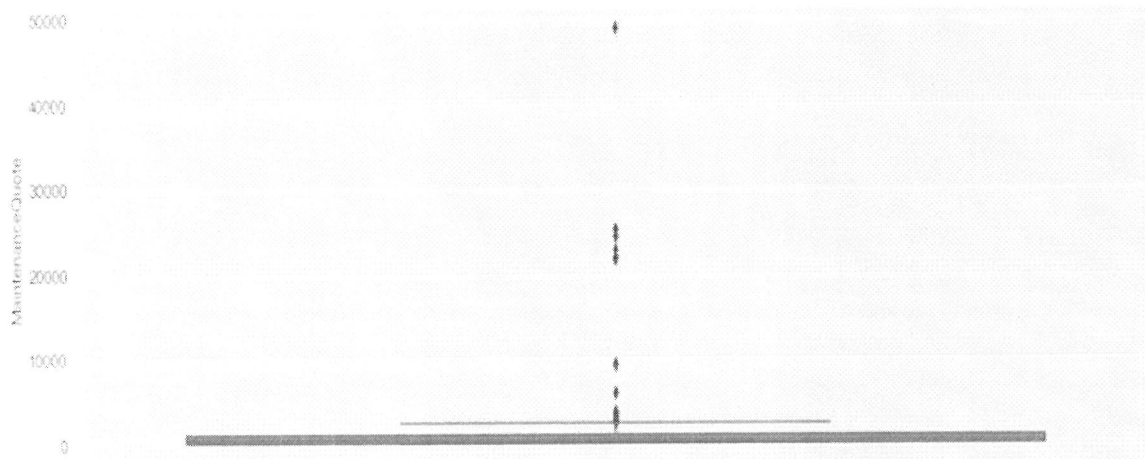
3.5 Data Pipeline and Process



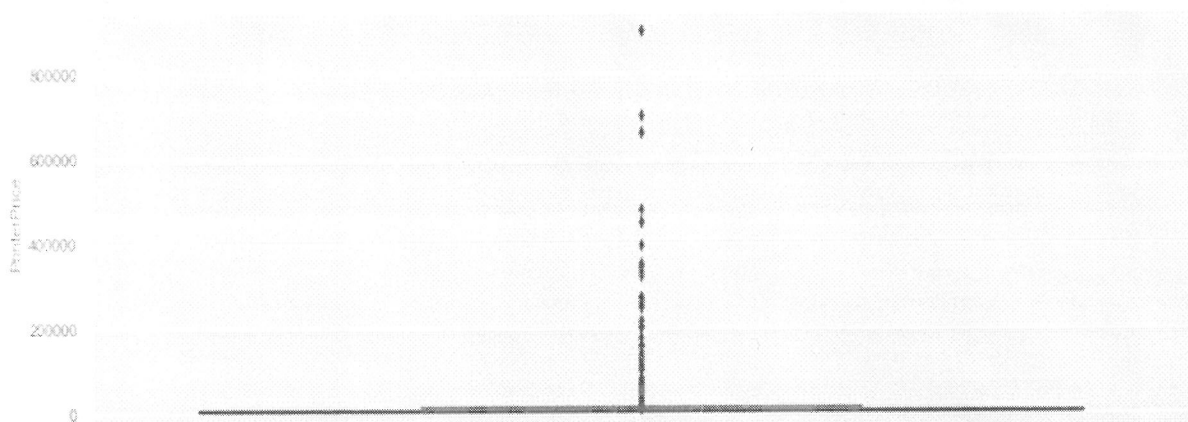
3.6 Data Cleaning and Preprocessing

- Removal of Non_available values by converting them to 0.0 and then removing them
 - maintenance_quote -> 43 removals
 - Printer_data -> 3029 removals
- Finding redundant data and then replacing them by mean
 - maintenance_quote have 29 redundant values
(maintainace quote was different for same Identification id)

3.7 Determination of Outliers



Outliers in Maintenance Quote data

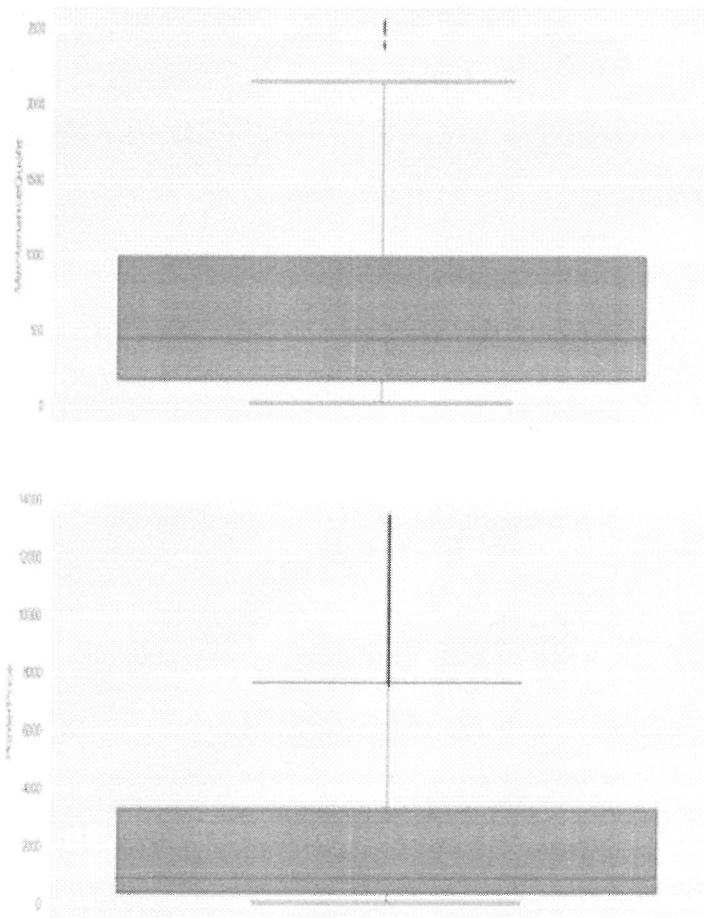


Outliers in printer price data

3.8 Removal of Outliers

- we have to remove outliers from the data so I have chosen Interquartile range method where I am finding where the bulk of the values lie.
- the interquartile range is the region between the 75th and 25th percentile (75 – 25 = 50%) of the data)
- $IQR = Q3 - Q1$
- Upper Limit = $(IQR * 1.5) + 75\% \text{ quantile}$
- We have chosen values that are more than 75 % quantile plus IQR so that all outliers are removed

3.9 Box Plot after removal of outliers



Here from the box plot, you can see that we have removed maximum outliers

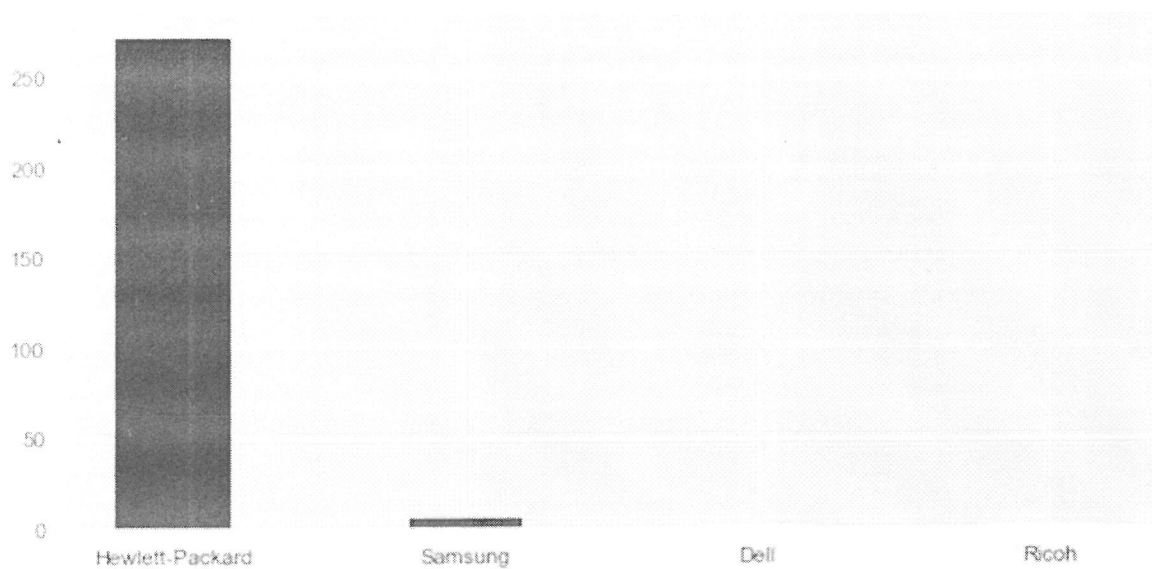
3.10 Merging the Data

- We have used inner join to the maintenance_quote data frame and printer_data dataframe
- After merging the shape of the dataset was 281 X 12
- This Dataframe have continuous as well as categorical data so we had to handle them

```
inner.head()
```

	IdentificationId	MaintenanceQuote	Make	DeviceType	MonoSpeed	ColorSpeed	hasA3	hasA4	hasColor	hasScan	hasFax	PrinterPrice
0	DL0207	167.99	Dell	Type_D	36.0	36.0	False	True	True	True	False	999.99
1	HP0002	216.01	Hewlett-Packard	Type_A	40.0	0.0	False	True	False	False	False	1499.00
2	HP0003	300.99	Hewlett-Packard	Type_A	50.0	0.0	False	True	False	False	False	1800.00
3	HP0004	99.03	Hewlett-Packard	Type_A	33.0	0.0	False	True	False	False	False	528.06
4	HP0007	175.99	Hewlett-Packard	Type_A	45.0	0.0	False	True	False	False	False	1148.72

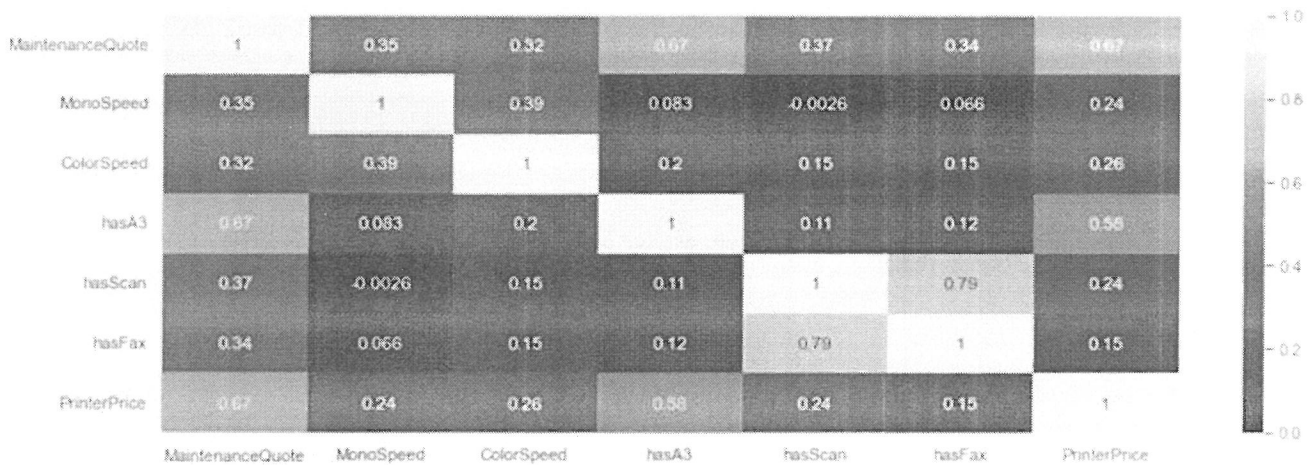
We have dropped 'Make' column because more than 95% data was from a particular company



I have marked the two columns which is highly correlated with Maintenance Quote

MaintenanceQuote	
MaintenanceQuote	1.000000
MonoSpeed	0.350287
ColorSpeed	0.324398
hasA3	0.674557
hasA4	NaN
hasColor	0.127495
hasScan	0.371543
hasFax	0.341050
PrinterPrice	0.671243

3.11 Correlation Heat Map



3.12 Feature Selection

- we have dropped hasA4 because it have only true values as all the printer will print at least a4 and it is having Nan values in correlation
- We have dropped Make as it has HP printers in abundance and other company printers are very rare
- We have dropped hasColor because colorSpeed is almost same attribute, and it has a very low correlation with Maintenance Quote
- We have dropped identificationId as well as it is a categorical data, and all are unique id's

3.13 Encoding the Data

```
one_hot_encoded_data = pd.get_dummies(inner, columns = ['DeviceType', 'hasA3', 'hasScan', 'hasFax'])
```

```
one_hot_encoded_data.dtypes
```

```
identificationId    object
MaintenanceQuote   float64
Make               object
MonoSpeed          float64
ColorSpeed         float64
hasA4              bool
hasColor           bool
PrinterPrice       float64
DeviceType_Type_A   uint8
DeviceType_Type_B   uint8
DeviceType_Type_C   uint8
DeviceType_Type_D   uint8
hasA3_False         uint8
hasA3_True          uint8
hasScan_False       uint8
hasScan_True        uint8
hasFax_False        uint8
hasFax_True         uint8
```

We have used one hot encoding to encode the categorical data

3.14 Splitting data into 70:30 ratio and Training

```
# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3)
```

3.15 Fitting the Data in Linear Regression Model

```
# Fitting Multiple Linear Regression to the Training set
regressor = LinearRegression()
regressor.fit(X_train, y_train)
```

Test Accuracy	Train Accuracy
<pre># Predicting the Test set results y_pred_test = regressor.predict(X_test) from sklearn.metrics import accuracy_score accuracy=regressor.score(X_test,y_test) print(accuracy) 0.7494597350095564</pre>	<pre>#predicting the Train Result using the model y_pred_train = regressor.predict(X_train) from sklearn.metrics import accuracy_score accuracy=regressor.score(X_train,y_train) print(accuracy) 0.6603452835162604</pre>

Gap between train accuracy and test accuracy should be the least possible and the above results are showing that

3.16 Calculation of Error

$$J(w_1, w_2, \dots, w_m, b)$$

where, $J \rightarrow \text{Cost/Error/Loss function}$

Mean Absolute Error

$$\rightarrow \text{Error} = \frac{1}{m} \sum_{i=1}^m |\hat{y}^{(i)} - y^{(i)}|$$

Mean Square Error

$$\rightarrow \text{Error} = \frac{1}{2m} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2$$

3.17 Evaluation Model

```
#evaluation matrices
```

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
RMSE= float(format(np.sqrt(mean_squared_error(y_test,y_pred_test)),'.3f'))
MSE= mean_squared_error(y_test,y_pred_test)
MAE=mean_absolute_error(y_test,y_pred_test)
r2=r2_score(y_test,y_pred_test)
```

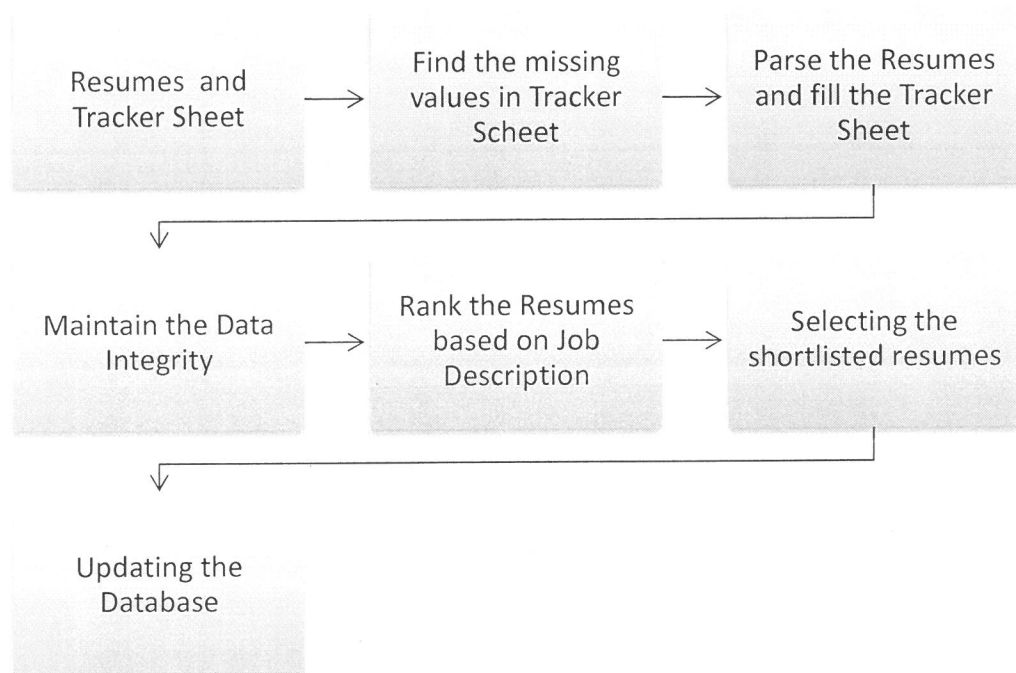
```
print('Root_Mean_Square_Error=',RMSE, '\nMean_Square_Error=',MSE, '\nMean_Absolute_Error=',MAE, '\nR2_Score=', r2)
```

```
Root_Mean_Square_Error= 0.5
Mean_Square_Error= 0.2504783116506054
Mean_Absolute_Error= 0.3577381489491203
R2_Score= 0.7494597350095564
```

CHAPTER 4: Internal Project

The internal project was assigned to us just after the results of the hackathon and I was given a project called “Resume Shortlisting”. As the name suggests this project is of internal use to eliminate the manual process done by HR to select the resumes.

4.1 Project Flow



4.2 My Contribution to Project

This was an ongoing project and I have to write some features for this project

- Determination of Carrer Gap if any from the resume
- Calculation of Rank Score for the resume based on the parameters provided by the HR
- Converting the Rank Score of resumes into categories so as to give priority to the candidates

CHAPTER 5: DATA ENGINEERING TRAINING

Data engineers work in a variety of settings to build systems that collect, manage, and convert raw data into usable information for data scientists and business analysts to interpret. Data engineering helps make data more useful and accessible for consumers of data. To do so, data engineering must source, transform, and analyze data from each system. For example, data stored in a relational database is managed as tables, like a Microsoft Excel spreadsheet.

Data engineers add value as they automate and optimize complex systems, transforming data into an accessible and usable business asset.

5.1 Roles of Data Engineer

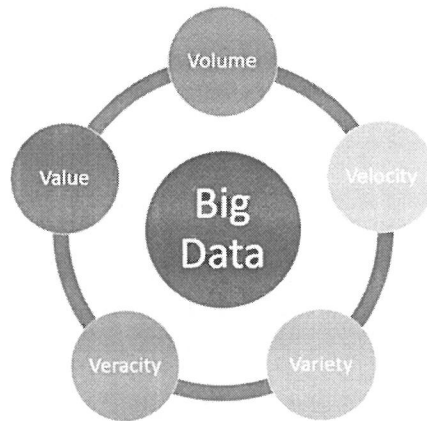
- **Generalist:** Generalists are typically found on small teams or in small companies. In this setting, data engineers wear many hats as one of the few “data-focused” people in the company. Generalists are often responsible for every step of the data process, from managing data to analyzing it. Dataquest says this is a good role for anyone looking to transition from data science to data engineering since smaller businesses won’t need to worry as much about engineering “for scale.”
- **Pipeline-centric:** Often found in midsize companies, pipeline-centric data engineers work alongside data scientists to help make use of the data they collect. Pipeline-centric data engineers need “in-depth knowledge of distributed systems and computer science”.
- **Database-centric:** In larger organizations, where managing the flow of data is a full-time job, data engineers focus on analytics databases. Database-centric data engineers work with data warehouses across multiple databases and are responsible for developing table schemas.

5.2 Responsibility of Data Engineer

- Identify ways to improve data reliability, efficiency, and quality
- Conduct research for industry and business questions
- Use large data sets to address business issues
- Deploy sophisticated analytics programs, machine learning and statistical methods
- Prepare data for predictive and prescriptive modeling
- Find hidden patterns using data
- Use data to discover tasks that can be automated
- Deliver updates to stakeholders based on analytics

5.3 Big Data

Huge amount of data that you cannot process using the traditional methods



Volume

It refers to the size of Big Data. Data can be considered Big Data or not is based on the volume. The rapidly increasing volume data is due to cloud-computing traffic, IoT, mobile traffic etc.

Velocity

It refers to the speed at which the data is getting accumulated. This is mainly due to IoTs, mobile data, social media etc.

Variety

It refers to Structured, Semi-structured and Unstructured data due to different sources of data generated either by humans or by machines.

Veracity

It refers to the assurance of quality/integrity/credibility/accuracy of the data. Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights.

Value

Just because we collected lots of Data, it's of no value unless we garner some insights out of it. Value refers to how useful the data is in decision making. We need to extract the value of the Big Data using proper analytics.

5.4 Challenges with RDBMS

- As the size of the data increases processing will become slow
- It cannot cope with unstructured data processing

5.5 Data Pipeline

A *Data pipeline* is basically a set of tools and processes for moving data from one system to another for storage and further handling. It captures datasets from multiple sources and inserts them into some form of database, another tool, or app, providing quick and reliable access to this combined data for the teams of data scientists, BI engineers, data analysts, etc.

A data pipeline is commonly used for:

- moving data to the cloud or to a data warehouse,
- wrangling the data into a single location for convenience in machine learning projects,
- integrating data from various connected devices and systems in IoT,
- copying databases into a cloud data warehouse, and
- bringing data to one place in BI for informed business decisions.

5.6 ETL (Extract Transform Load) Process

Extract

During data extraction, raw data is copied or exported from source locations to a staging area. Data management teams can extract data from a variety of data sources, which can be structured or unstructured. Those sources include but are not limited to:

- SQL or NoSQL servers
- CRM and ERP systems
- Flat files
- Email
- Web pages

Transform

In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated for its intended analytical use case. This phase can involve the following tasks:

- Filtering, cleansing, de-duplicating, validating, and authenticating the data.
- Performing calculations, translations, or summarizations based on the raw data. This can include changing row and column headers for consistency, converting currencies or other units of measurement, editing text strings, and more.
- Conducting audits to ensure data quality and compliance

- Removing, encrypting, or protecting data governed by industry or governmental regulators
- Formatting the data into tables or joined tables to match the schema of the target data warehouse.

Load

In this last step, the transformed data is moved from the staging area into a target data warehouse. Typically, this involves an initial loading of all data, followed by periodic loading of incremental data changes and, less often, full refreshes to erase and replace data in the warehouse. For most organizations that use ETL, the process is automated, well-defined, continuous, and batch-driven. Typically, ETL takes place during off-hours when traffic on the source systems and the data warehouse is at its lowest.

5.7 Hadoop vs Spark

Hadoop is a framework that manages big data in a distributed way and processes it parallelly

Hadoop HDFS

- Storage Unit of Hadoop

Hadoop Map Reduce

- Processing Unit of Hadoop

Spark is a newer project, initially developed in 2012, at the AMPLab at UC Berkeley. It's a top-level Apache project focused on processing data in parallel across a cluster, but the biggest difference is that it works in memory.

In my company we use spark for the following reasons: -

- Spark is lightning-fast cluster computing technology, which extends the MapReduce model to efficiently use with more types of computations.
- Spark reduces the number of read/write cycles to disk and stores intermediate data in memory, hence faster-processing speed.
- Spark is designed to handle real-time data efficiently.
- Spark is a low latency computing and can process data interactively.
- Spark can process real-time data, from real-time events like twitter, facebook
- Spark requires a lot of RAMS to run in-memory, thus increasing the cluster and hence cost.

5.7 Databricks

The Databricks Lakehouse Platform provides an end-to-end data engineering solution — ingestion, processing and scheduling — that automates the complexity of building and maintaining pipelines and running ETL workloads directly on a data lake so data engineers can focus on quality and reliability to drive valuable insights.



databricks

Streamline data ingestion into your lakehouse

Incrementally process new files as they land on cloud storage — with no need to manage state information — in scheduled or continuous jobs. Efficiently track new files (with the ability to scale to billions of files) without having to list them in a directory. Databricks automatically infers the schema from the source data and evolves it as the data loads into the Delta Lake lakehouse.

Automate data transformation and processing

Once data is ingested into the lakehouse, data engineers need to turn raw data into structured data ready for analytics, data science or machine learning. Simplify data transformation with Delta Live Tables (DLT) — an easy way to build and manage data pipelines for fresh, high-quality data on Delta Lake. DLT helps data engineering teams by simplifying ETL development and management with declarative pipeline development, improved data reliability, and cloud-scale production operations to help build the lakehouse foundation.

Code For Creating Dataframe In Databricks Using Pyspark

```
df=spark.read.format("csv")\
    .options(header='true',delimiter=',')\
    .load("/FileStore/tables/walmart_stock.csv")

display(df)
df.schema

# COMMAND -----

from pyspark.sql.types import *
CustomSchema = StructType([
    StructField("Date",DateType(),True),
    StructField("Open",DoubleType(),True),
    StructField("High",DoubleType(),True),
    StructField("Low",DoubleType(),True),
    StructField("Close",DoubleType(),True),
    StructField("Volume",IntegerType(),True),
    StructField("Adj Close",DoubleType(),True)
])

CustomSchema.fieldNames
```

Code For Query or Transforming Data

```
# MAGIC Q.3.Format the 'Adj Close' column to show up to two decimal places only.

from pyspark.sql.functions import *

df \
    .withColumn("Adj Close",col("Adj Close").cast(DecimalType(5,2))) \
    .show()
```

Creating A Cluster For Performing Job Which Is In The Form Of Big Data.

Create Cluster

New Cluster

Cancel

Create Cluster

0 Workers:0 GB Memory, 0 Cores, 0 DBU
1 Driver:15.3 GB Memory, 2 Cores, 1 DBU

Cluster name

datavowel_cluster

Databricks runtime version

Runtime: 10.4 LTS (Scala 2.12, Spark 3.2.1)

Instance

Free 15 GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For more configuration options, please upgrade your Databricks subscription.

Instances

Spark

Availability zone

auto

NOTEBOOK VIEW OF DATABRICKS:

Assignment Python

demoduster

Cell 1

Q.1.Create a spark session in data bricks notebook and attach it to a cluster.

Cell 2

Q.2.Load the given csv file and check its schema. If not correct, then correct the schema and the datatypes of attributes.

Cell 3

```
1 df=spark.read.format("csv")\
2   .options(header='true',delimiter=',')\
3   .load("/FileStore/tables/walmart_stock.csv")
4
5 display(df)
6 df.schema
```

5.8 Apache-Airflow



Airflow is a platform to programmatically author, schedule and monitor workflows. Use Airflow to author workflows as Directed Acyclic Graphs (DAGs) of tasks.

The Airflow scheduler executes your tasks on an array of workers while following the specified dependencies. Rich command line utilities make performing complex surgeries on DAGs a snap. The rich user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed.

When workflows are defined as code, they become more maintainable, versionable, testable, and collaborative.

Airflow is not a data streaming solution. Tasks do not move data from one to the other (though tasks can exchange metadata!). Airflow is not in the Spark Streaming or Storm space; it is more comparable to Oozie or Azkaban.

Workflows are expected to be mostly static or slowly changing. You can think of the structure of the tasks in your workflow as slightly more dynamic than a database structure would be. Airflow workflows are expected to look similar from one run to the next, this allows for clarity around the unit of work and continuity.

Dag Creation To Automate Workflow :

```
from airflow.models import DAG
from airflow.providers.sqlite.operators.sqlite import SqliteOperator
from datetime import datetime
default_args = {
    'start_date' : datetime(2022,1,1)
}

with DAG('User_samyak', schedule_interval = '@daily',
        default_args=default_args,
        catchup = False) as dag:
    #define the tasks/operators
    creating_table = SimpleHttpOperator(
        task_id = 'creating_table',
        sqlite_conn_id='db_sqlite',
        sql='''
        CREATE TABLE users(
            firstname TEXT NOT NULL,
            lastname TEXT NOT NULL,
            country TEXT NOT NULL,
            username TEXT NOT NULL,
            password TEXT NOT NULL,
            email TEXT NOT NULL PRIMARY KEY,
        );
        ''')
```

CHAPTER 6: PROJECT DEPLOYMENT

6.1 Introduction

I am deployed to The project known as HP DataOS.

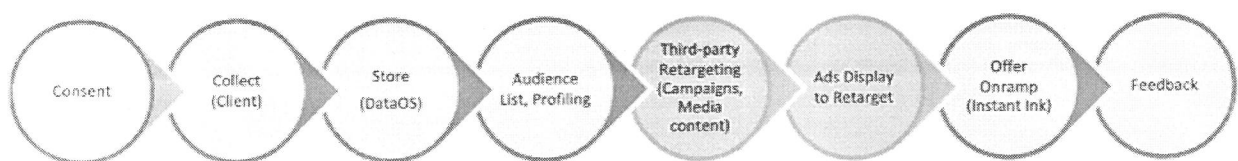
HP DataOS works on P2 opportunities.

6.2 What is P2?

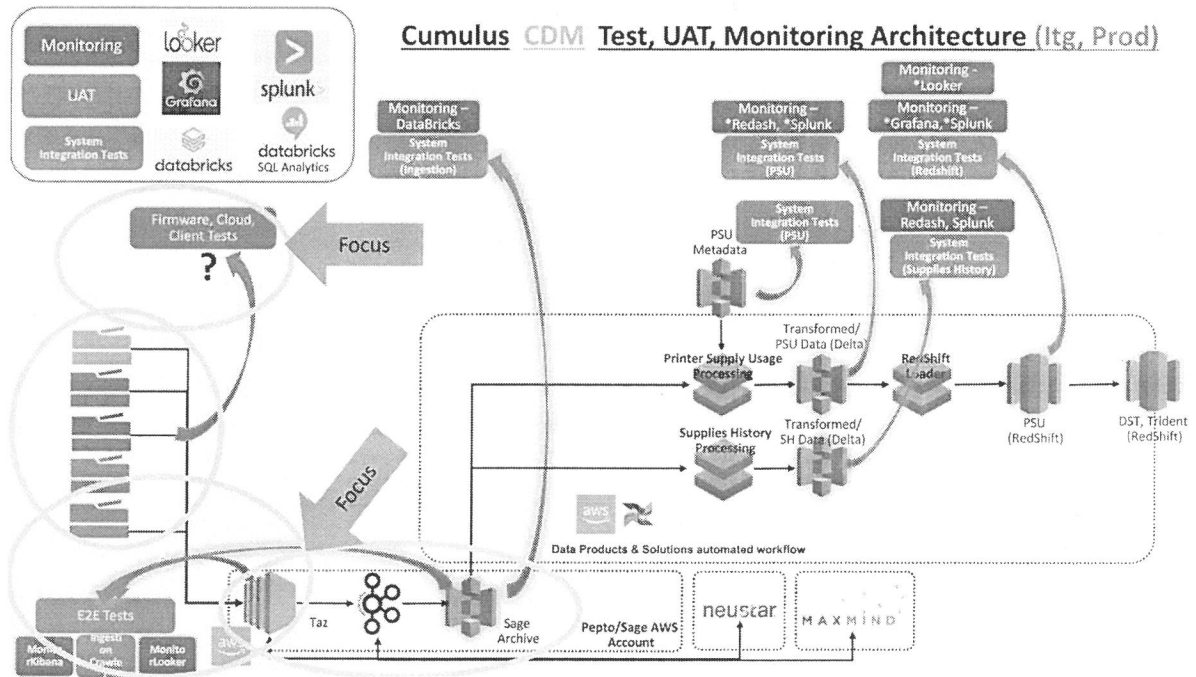
- P2- General term describing 'Post' engagement opportunities with our customers
- P2 Retargeting is to
 - Expand addressable audience to retarget a greater group (bring print customers back to HP)
 - Focus on HP Smart App Users as Initial Key Growth Opportunities
 - Grow instant ink subscribers aggressively in FY21



6.3 Functional Flow



6.4 Project Architecture



6.5 Tools and Technologies Used

Amazon S3 bucket

Amazon Simple Storage Service (Amazon S3) is an object storage service that offers industry-leading scalability, data availability, security, and performance. You can use Amazon S3 to store and retrieve any amount of data at any time, from anywhere.

Jira

Jira is a software application used for issue tracking and project management. The tool, developed by the Australian software company Atlassian, has become widely used by agile development teams to track bugs, stories, epics, and other tasks.

Looker

Bring the power to connect, analyze, and visualize data across multicloud environments. Looker for Google Cloud Platform allows anyone in your business to quickly analyze and find insights in your datasets.

CHAPTER 6: CONCLUSION

At IntimeTec Visionsoft Pvt. Ltd., the Practice School- II opportunity provided a lot of knowledge in Development technologies, DeveOps and DataOps. I can affirm that the professional experience and knowledge I acquired here would really assist my overall career. I have accomplished my internship by learning lots of new tools and technologies.

Working in a professional setting, such as IntimeTec Visionsoft Pvt. Ltd., has also taught me valuable skills such as teamwork and basic professional work ethics. The mentors here are extremely helpful and supportive.

Moving on to the project work. My contributions to the project work at the company were highly appreciated by my mentors and my seniors. Working on a live project improved my knowledge and skillsets. During this internship period.

Lastly, I am very pleased to receive a full-time employment offer from IntimeTec for the role of Software Engineer.

REFERENCES

1. **Data Engineering Course:** <https://moodle.intimetec.americas>
2. **Apache-Airflow:** <https://airflow.apache.org/docs/apache-airflow/stable/>
3. **Databricks:** <https://docs.databricks.com/>
4. **PySpark:** <https://spark.apache.org/docs/latest/>



Plot No. 1/2, Kanakpura Industrial Area, Kanakpura, Sirsi Road, Jaipur-302034
Site No.271, Sri Ganesha Complex, Hosur Main Road, BTM 1st Stage, Bangalore- 560068.

Date: 26th May'22

TO WHOMSOEVER IT MAY CONCERN

Employee Name: Samyak Bhagat

Employee ID: ITTV/EMP/2703

Address: Plot no. 1 / 2, Kanakpura Industrial Area, Kanakpura. Sirsi Road, Jaipur-302034.

This is to certify that **Mr. Samyak Bhagat** is pursuing internship followed by employment with **In Time Tec Visionsoft Pvt. Ltd** from **1st Feb '22** at our **Jaipur** office located at **Plot no. 1/2, Kanakpura Industrial Area, Kanakpura. Sirsi Road, Jaipur-302034.**

During his Internship period, he is entitled to a stipend of Rs. 15,000/- per month, other details are valid as per the offer letter.

Regards,

Neha Dubey

Sr. HR Associate

In Time Tec Visionsoft (p) Ltd.

Confidential

This is a Computer Generated Document. No Signature Required.

