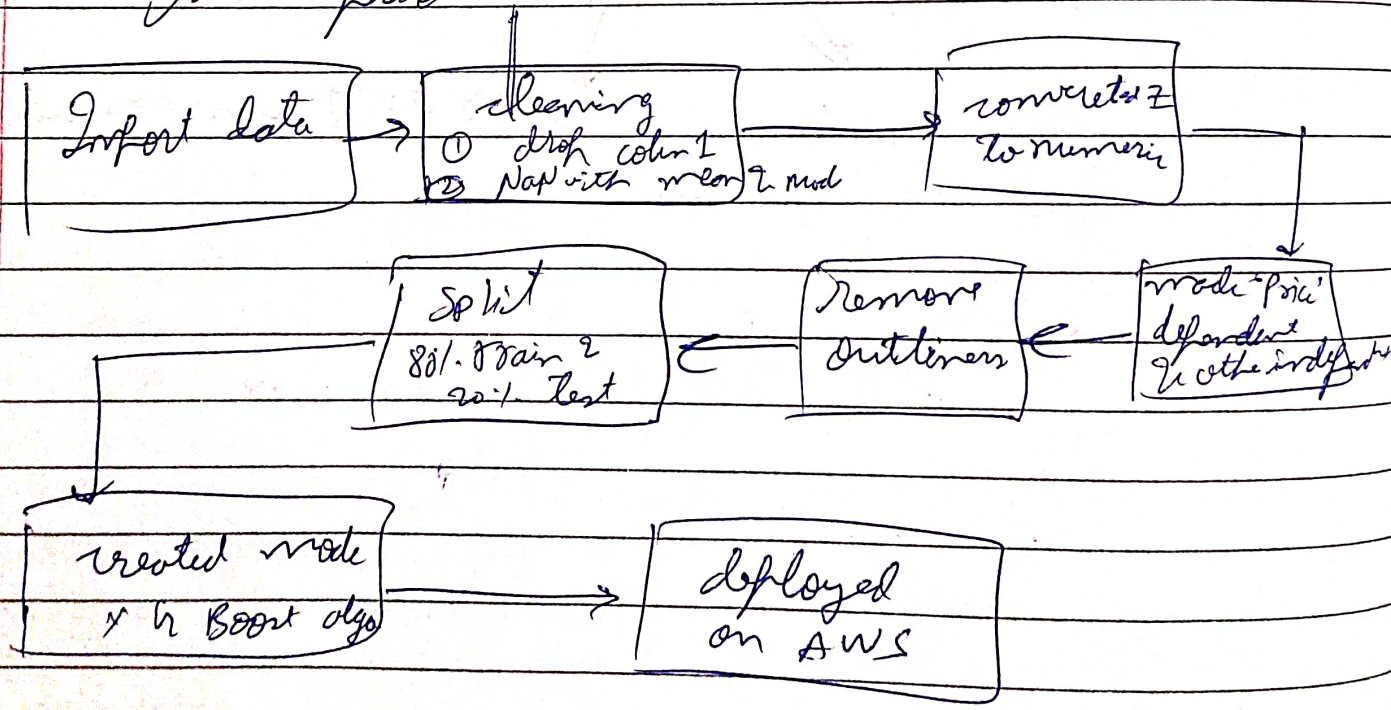
B1 - Siddesh - 56602

Page No.
Date.

Q] Can you explain your C-DAC project? Draw block diagram as appropriate.

→ The project is designed specially for diamond merchants to approximate the cost of a diamond in order to avoid imprecise valuation preventing deceitful transaction. It implements ML model (XGBoost) which predict price using cut, clarity, carat, color and other dimensions. The will also help average consumer to get an ~~average~~ rough idea about the price



B1 - Siddesh - 56602

Page No.
Date.

Q) ~~What's new in your project?~~

Q) Explain why did you select this technology & framework for this project?

→ We have used supervised Machine learning for model creation as it learn from the labeled data and using this model we can predict the outcome. The model is ~~based~~ on XG Boost algorithm. We have used flask for creation of front end which ~~used~~ internally uses rest API for handling HTTP requests like get, post. The project is deployed on AWS EC2 instance. We have used API gateway service of AWS to make the link secure.

BI - Siddesh - 56602

Page No.
Date.

8) How will you deploy your project (on web server / client machine)

→

① Create and launch an AWS EC2 instance.

② Connect to the created instance using SSH & .pem file.

③ install all the required libraries on the instance.

④ copy all the project related files to that instance through SCP

⑤ Run server.py file inside

⑥ get the public I.P and append the port number for the particular .py file running on . Using this URL test whether working or not.

BJ - Siddesh - 56602

Page No.
Date.

Q) What are the limitation of your project?

→ (1) Old data set; so predicted price is not according to the current market price

(2) Not considered aging factor to the time factor.

(3) More number of features required to predict the price.

Q) what are the difficulties you have faced during this project and How you have overcome it?

→

While cleaning data set we ~~found that~~ faced a lot of difficulties. 2 column was object type which should be numberical type. We tried converting it but it was difficult as it was giving errors. & We then ~~got~~ found that some of the row had string type data which we then dropped it.

We were trying to implement API gateway over a secure link to our project's website but we were not able to route the requests from our main page, some research taught us the steps and sucessfully implemented the gateway.

B1 - Siddesh - 56602

Page No.
Date.

Q) Mention sources of your dataset?

→ - Data. world (google dataset search)

Q) Application of your project?

→ - It can be used by diamond merchants & ~~avge~~ average consumer to get an rough idea about how much their diamond could cost.

Q) Name some feature extraction techniques used for ~~dim~~ dimentionality reduction in your project?

→ - Principal component Analysis technique is used for dimentionality reduction and using this PCA technique we have reduced our dimentionality to 1, so that we can visualize our data to compare between actual & predicted values.

BI - Siddesh - 56602

Page No.
Date.

7) How much data you have allocated training, validation & testing in your project?

→ We have splitted our data into 80% & 20%, 80% for training & 20% for testing. As data is not so big so validation data set is not req. We were also getting good accuracy

8) Mention which steps you followed in your Analytics project.

→
1. Imported req. packages
2. created data frame by reading CSV
3. perfor EDA & data cleaning
4. converted categorical variable to numeric using label encoder
5. Converted wrong data types to correct / valid ones
6. check correlation & remove outliers from highly co-related variables
7. created a pipeline of various algorithms to find out the most accurate algorithm
8. Created model using XGBoost

B1 - Siddesh - 56602

Page No.
Date.

Q) Which data cleaning and validation methods used in your project?

→
1. Dropped redundant column.
2. Replacing NaN values with mode & mean accordingly
3. Encoded categorical variables
4. ~~Removed~~ ~~outliers~~ wrong data type conversion
5. Removed outliers from the correlated variables

Q) Which tool you have used for data analysis? Why?

→
- Various libraries in python for ML like numpy, pandas, seaborn, matplotlib, sklearn, XGBoost for visualisation and processing the data
- Or, IDE used is jupyter Notebook
- Apart from this flask is used for front end development
- We have also used PowerBI for data visualization.
- We have deployed our model on AWS