Project rapid fire

Q.1 Can you explain your C-DAC Project ? Draw block diagram as appropriate.

→ My CDAC Project titled 'Airbnb Data Analysis' & Price prediction' where me & my project partner worked on a dataset to get some insight of the Airbnb listings & predict the price of new properties using this data.

The dataset is basically used for a kaggle competition. It consists of listings from 2008 till 2018. There were many missing values in the dataset which is why a lot of cleaning was required. We performed feature selection by 3 different ways & selected the best method. Similarly, the choice of algorithm was done based on the performance metrics

We mainly focused on RMSE & R2 score values for determining the performance.

Finally we used Random forest and tuned it for best performance.

Further, a web application is developed using Flask Framework and deployed using aws ec2 instance.

Page No.:
Date. / /

Q.3 Why did you select this technology & framework for this project?

→ We have used many libraries in Python such as pandas, matplotlib & seaborn for multiple advantages.
These libraries have many advantages like easy implementation, readability, openly available resources, the highly efficient data tools, platform independence, and most importantly, the support of a large developer community.

We used flask framework for creating a small web application. Flask does not require a lot of configuration, Hence it is suitable for small to medium projects.

After that, to deploy the application we used Amazon's AWS EC2 instance.

Page No.:
Date. / /

Q.5 Which advanced features have you used in your project?

→ I have used some advanced data cleaning techniques in some features of the dataset like binning, imputing values according to an another feature etc. The main aim during cleaning was to avoid the loss of data and make max use of the available usable data.

For feature selection I have used an external package called featurewiz which uses xgboost for feature selection.

Date: / /

Q.6 What was your role in your project & explain what you did in it?

→ I did this project with my project partner. In the initial phase, we both searched for a suitable dataset. Then after getting one, the flow of the project was decided by me. It includes the steps by step approach, the distribution of tasks and deadlines which we had set for ourself.

Among the technical things, I was responsible for data cleaning and training. Other tasks like visualization & feature selection were done by my partner.

For, whatever part that I did, I first looked for the standard methods to do that, then which python libraries can be used, how to use them etc. then I looked for how to apply it to our project.

Finally the documentation part - like collecting images, writing & editing the report was done by me.

Q.9 How will you deploy your project?

→ I am using Amazon AWS EC2 instance currently for deploying the model because it is free of cost But for bigger Storage and processing requirements, the instance can be scaled up.

Q.10 What are the limitations of your project?

→ I think, In our project, we did cleaning & visualization at a good level. But the accuracy of our model can be improved. And this can be done using deep learning algorithms.

Another limitation which I faced is that I wanted to perform grid search for selecting optimum hyper parameters But couldn't do it because of the configuration of our machines.

Page No.:
Date. / /

Q.11 What are the difficulties that you have faced in your project? How you overcame it?

→ The biggest difficulty which I & my partner found was ~~handling the missing values in~~ feature selection.

During feature selection for better results I started to search online the different ways for feature selection. I came across an external package called featurewiz which uses xgboost for feature selection.

Q.12 How will you improve performance of your project (memory related & response time)?

→

Q.15 Mention sources of your project.

→ We used the dataset used for a Kaggle competition. It was in CSV format and included data from 2008 to 2019.

Q.16 Application of your project?

→ The main aim of the project was to build an application which predicts the price of a property based on various features.

Property owners who want to list their properties on Airbnb but are not sure about what price should they assign, can use this app.

Q.18 How much data have you allocated for training, validation & testing in your project?

→ Training : 80 %

Testing : 20 %

Not allocated data for validation of the data was small.

Page No.:
Date: / /

Q.2) Mention which steps you followed in your project.

→ Step (1) Understanding the data.

From various sources, user experiences, host experiences, we understood the features which highly affect the price.

Step (2) Visualization

To get better understanding of the data and to create some dynamic dashboards, visuals were created using PowerBI

Step (3) Data Pre-processing / Cleaning

Some features cannot be directly used for analytics. The have to be modified if they are of different type, contain null values or corrupted values.

It is the most crucial part, which was done keeping in mind the goal to avoid data loss as much as possible.

Step (4) Feature & Algorithm selection.

Out of the available features, only the few important ones are selected to improve accuracy. Further different algorithm are tested & the one with best results is selected.

Step (5) Training & hyperparameter tuning

On the selected algorithm, tuning is performed for obtaining the best hyperparameters.

Step (6) Create & deploy the application.

A web application is created using flask framework in python. this appⁿ is deployed using Amazon EC2 instance on the cloud.

Q.22 Which data cleaning methods used in your project?

→ For missing values ; if no of nulls is less then it can be replaced by mean, mode or median. But for high no of missing values various techniques have been used.

(i) Binning : converting a numerical feature into a categorical one & assigning a new category for null values

(ii) Imputing values according to some other independent variable

Page No.:
Date. / /

For processing the date columns, first the date columns were converted to no. of days till current date & then by binning them into categories, missing values were handled.

From the amenities feature, unique amenities were extracted & important amenities were converted into new columns.