Project Journal

Soumya Madhav x23332018

Dataset Name**:** Melbourne Housing Data set

1. Summary of Contribution

Actively involved in the discussion and suggestion on the use case and data selection, since we could directly relate to the higher housing prices here in Ireland owing to the demand supply gaps, we decided to use this project to study the housing datasets worldwide to understand and analyze the data for meaningful insights and try to create a predictive model for housing prices. We worked on datasets from 3 different continents separately and collectively analyzed the results to come at the conclusion.

I worked with the Melbourne dataset from Kaggle, converted the CSV to JSON and stored the data to MongoDB, further conducted exploratory analysis on the raw data, cleaned the data and stored the processed data into PostgreSQL database. Also standardized, refined the data and analysed performance of machine learning models like Linear regression, Random Forest and Decision Tree. To ensure easy access to the data and workflows, created a streamlit application. All the code is pushed into a github repo here is the link to access the repo.

Repo link : https://github.com/soumadhav/HousingPricePredection.git

2. Tasks and Responsibilities

2.1 Dataset Selection

- Dataset used: Melbourne Housing data.
- Source:https://www.kaggle.com/datasets/anthonypino/melbourne-housing-marketThedataset
- Reason for selection: The Melbourne dataset was chosen for our project due to its rich and diverse features like rooms, distance from the city center, price and suburb. This provided a realistic representation of citywise real estate pricing enabling the study of factors influencing house prices. This data set is a structured data set containing both categorical and numerical features providing opportunities for data preprocessing and feature engineering.

- **2.2 Data Preprocessing and Transformation**

Handling Missing Values

Features with more than 50 % missing values were dropped. Features like rooms and bedrooms hence only of the columns were considered. And remaining features which could not be dropped were imputed with mean, median and mode values using simple imputer from scikitlearn accordingly to achieve better performance. Duplicate rows were dropped to consider unique data.

Handling outliers

Box plot was plotted to study the features with outliers. Price column exhibited majority of outliers using capping method the outliers were handled.

Normalization

As the data was rightly skewed, applied log transformation to achieve uniform distribution of the data in price column.

Tools/technologies used: Python, Pandas, NumPy, Scikitlearn

2.3 Database Integration

Database used: MongoDB, PostgreSQL.

The raw data was stored in mongoDB on MongoAtlas platform and the cleaned processed data in structured format was stored in Postgresql on neon platform for further analysis.

2.4 Visualization

Visualizations created: Box plots, Scatter plots, Bar plots , Heatmaps

2.5 Feature selection and modelling

All the categorical columns were transformed into numerical features using label encoder from scikit learn library. Data was standardised using standard scaler to

achieve uniform scaling. According to correlation map the features with least influence were dropped to avoid the problem of overfitting to enhance the performance of the model.

3. Key Challenges and Solutions

1. Getting a relevant JSON dataset for the use case. After hours of searching for relevant dataset for a housing dataset in JSON format, the appropriate dataset was in CSV, to ensure that I met the learning objective of using JSON dataset, I resorted to converting the CSV data to JSON.
2. Since I planned to also build an app that would be accessible on the cloud, I needed my data to reside on a cloud hosted database. Hence, I went with Mongo Atlas. However, this added additional complexity of ensuring the network whitelisting as Mongo Atlas give limited network access. Also, there was certificate errors from the pymongo connection, This was resolved using 'certifi' and passing it along with the mongo url to MongoClient as tlsCAFile=certifi.where().
3. During EDA the challenges faced were missing values, irrelevant features, the features were in object format had to convert them to numeric to make them suitable for analysis. Outliers were handled and the skewed data was normalized to achieve uniformity.
4. Loading the data into PostgreSQL, kept failing as the dependency for psycopg2 was not satisfied, trying to install the same kept failing due to build issues, I had to resort to using psycopg2-binary instead, this solved the problem and ensure that I could upload to PostgreSQL.
5. Streamlit Application development the application connection with mongoDB atlas kept throwing errors, realized that the streamlit application was using anaconda instead of the python that was setup for the virtual env that ran the upload. I had to ensure that the required dependencies are installed on anaconda or run app using 'python -m streamlit run'.

4. Time Log

Breakdown of time spent on each task

Task	**Hours Spent**
Use case finalization	8 hours
Dataset Selection	8 hours

Data conversion to JSON	1 hours
Mongo Atlas setup	1 hours
Data upload to Mongo DB Atlas	4 hours
Exploratory Data Analysis	8 hours
Predictive model creation	4 hours
PostgreSQL on Neon - setup	1 hours
Data cleanup and upload to PostgreSQL	4 hours
Streamlit application for data exploration	8 hours
Report creation	6 hours
Presentation and video	5 hours

5. Reflection and Lessons Learned

This project was informative as it involved a diverse set of learnings in tech and teamwork, it helped me understand data gathering, data cleaning, data upload and retrieval using SQL and NoSQL, data analysis, predictive modeling and building application for end user to access the processed data. Seeing the result of the app and data being accessible from the cloud using streamlit, MongoAtlas and PostgreSQL on Neon was a good learning experience. I was able to create a public git repo and upload my work for future reference. This project has helped me learn to collaborate and work together for the project as in real time context.