House Pricing Estimation

Saurabh Annadate

Mid Point Check-In

Northwestern

Highlights

- The training data has a lot of missing values. However, a deep dive into the data generation process reveals that the data is structurally missing and can be correctly imputed
- Looking at the data summaries, the price of a house seems to be strongly dependent on the total above ground square footage as well as the total basement square footage
- Recency of construction impacts the price of the house
- Specific areas have a higher average house price as compared to others
- Areas with low population density have on an average higher house prices as compared to areas with lower density

Review Progress

Tasks Completed:

- ✓ Set up a S3 instance
- ✓ Initialize RDS database
- ✓ Exploratory Data Analysis
- ✓ Data Cleaning and missing value imputation
- ✓ Environment Setup : requirement.txt files

Demo / Analysis

- Refer to the EDA md for details: Link
- Refer to the project README.md for all configured functionality: Link

Lessons Learned

Importance of logging and Documentation

- Logging helps pin point the exact point of failure in case of an error which helps debug and resolve fast
- Logging is useful to track the usage and performance of the app/module
- Documentation helps maintain readable code which also aids in reproducibility

Reproducibility

- The model needs to be fully reproduceable by following the steps outlined in the README
- The app should be robust to different operating systems, or such details should be highlighted explicitly

Managing Dependencies

- It is important to provide details on all packages with the required versions in the requirements.txt file to prevent dependency issues

Modularity

- Writing modular code helps readability and easy scalability (Adding more features)
- Change implementation becomes very easy
- Having a single source of all configurations helps fast change implementation and gives easy control to the end user

Recommendations

Tasks tracked for next Sprint:

- ❑ Feature Generation
- ❑ Testing different model architectures and parameter tuning
- ❑ Model performance tests
- ❑ Deploy model using Flask
- ❑ Development of unit tests and integrated tests
- ❑ Develop a basic form to input data and output results
- ❑ Setup usage logs
- ❑ Solution reproducibility tests