CS 452 – Data Science with Python Project Report

1st Deniz Özbakır
Bachelor of Science, Computer Science
Ozyegin University)
Istanbul, Turkey
deniz.ozbakir@ozu.edu.tr

2nd Semanur Yaşar
Bachelor of Science, Computer Science
Ozyegin University
Istanbul, Turkey
semanur.yasar@ozu.edu.tr

***Abstract*—This paper focuses on analyzing the Melbourne housing market by analyzing various factors that affect house prices. To ensure data quality, the project includes extensive data cleaning, preprocessing, and feature engineering and to gain understanding of the dataset, descriptive statistics and exploratory data analysis approaches are applied. These approaches reveal correlations between the target variable(price) and categorical and numerical variables. Finally, various regression models are implemented to predict house prices, with assessment metrics showing the effectiveness of the models. These models include Multi-Linear Regression, Multi-Linear Regression with categorical variables, and Random Forest Regression.**

I. INTRODUCTION

The housing market is a complex dynamic system that is shaped by a variety of factors which determine house prices. For those who are involved in the real estate sector, understanding these elements is essential. In order to study the Melbourne housing market, this article will look at the correlations between various factors and home prices. In order to gain knowledge and make predictions, this study uses data cleaning, preprocessing, feature engineering, descriptive statistics, exploratory data analysis, and regression modeling approaches.

Data cleaning, preprocessing, and feature engineering are all steps in the initial phase of the project. These steps are used to make sure the dataset is properly prepared for analysis. Outliers are handled, missing data are resolved, and variables are converted and normalized to ensure data integrity.To fully understand the dataset, descriptive statistics and exploratory data analysis approaches are used. The correlations between the price, and the categorical and numerical variables are explored. Box plots, scatter plots, and correlation analysis provide useful insights into the data's relationships, distributions, and trends.

Next, multiple regression models are applied to predict housing prices. Models such as multi-linear regression, multi-linear regression with categorical variables, and random forest regression are used. These models represent the relationships between independent variables and the price, allowing for housing price predictions. To evaluate the accuracy of the models, evaluation metrics such as R-squared, mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are used.

The study is divided into these four sections: data cleaning, preprocessing, and feature engineering; descriptive statistics and exploratory data analysis; feature selection, modeling, model assessment, and validation; and interpretation and insights. Each section discusses the analyses' methodology, procedures, and findings.

The research results provide important insight into the Melbourne housing market. Significant factors influencing house prices include the number of rooms, bathrooms, car spaces, building area, property count, distance from center, and property age. Property type, region, suburb, and council area are other important categorical variables. These findings have significance for parties inside the Melbourne housing market, allowing them to make more informed judgments and develop better strategies.

II. BACKGROUND

A. Data Cleaning

The process of finding flaws, inconsistencies, and inaccuracies in a dataset and repairing or deleting them. It includes duties such as managing missing numbers, addressing outliers, standardizing formats, and guaranteeing data integrity.

B. Preprocessing

The preparation of data before analysis or modeling is referred to as preprocessing. To enhance the quality and usefulness of the data for analysis, it includes tasks such as data normalization, scaling, encoding categorical variables, addressing missing values, and variable transformation.

C. Feature Engineering

The practice of adding new features or modifying current features to improve the performance of machine learning models. It includes selecting relevant features, generating new characteristics from old ones, encoding categorical variables, dealing with outliers, and scaling variables.

D. Categorical Variables

Variables that reflect different categories or groups are referred to as categorical variables. They do not have a numerical significance and only have a finite set of possible values.

E. Regression Analysis

A statistical modeling technique used to determine how one or more independent variables and a dependent variable are related. By using the values of the independent variables, it assists in predicting the value of the dependent variable.

F. Multi-linear Regression

When there are numerous independent variables, multi-linear regression is one of the regression techniques applied. It assumes that the independent variables and the dependent variable have a linear relationship.

G. Multi-linear Regression with Categorical Features

An extension of linear regression that includes categorical variables in the model is called linear regression with categorical features. Categorical variables are transformed into numerical representations.

H. Random Forest Regression

A machine learning approach that predicts the dependent variable by using an ensemble of decision trees. For managing complicated relationships and interactions between variables, it is especially useful.

I. R-squared (R^2)

In a regression model, R-squared measures the percentage of variance in the dependent variable that can be explained by the independent variables. It has a value between 0 and 1, higher values representing a better fit between the model and the data.

J. Mean Absolute Error (MAE)

An accuracy metric for regression models that assesses the average absolute difference between predicted and actual values.

K. Mean Squared Error (MSE)

The average squared difference between the predicted values and the actual values. It is frequently used in regression analysis and gives a measure of the model's accuracy.

L. Root Mean Squared Error (RMSE)

The square root of mean squared error. It displays the typical difference between the predicted values and the actual values in the dataset. A common metric used for assessing the effectiveness of regression models.

III. DATA CLEANING, ANALYSIS AND INTERPRETATION

A. Data Cleaning, Preprocessing, Feature Engineering

The original dataset went through necessary cleaning and preprocessing steps in this phase of the project to ensure it is suitable for analysis and modeling. The first dataset had 19,741 items with 21 columns representing different housing attributes. The first step was to calculate the age of each property by subtracting the 'YearBuilt' value from the current year. It resulted in a more useful representation of the historical importance. The houses were then classified as "Historic" or "Contemporary" based on whether or not they were more than 50 years old. Also, to show that it is as a distinct location id, the 'Postcode' column is used as a categorical variable.

Then, missing values were fixed by removing rows with any missing data, including variables such as 'Price', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'YearBuilt', 'CouncilArea', 'Lattitude', and 'Longitude'. Outliers were further fixed by removing data with a 'BuildingArea' of zero, since such values are unlikely. However, 'Landsize' values which are zero are considered acceptable since some houses were constructed extremely near to the property boundaries. There were 1,015 such occurrences. These extensive data cleaning and preprocessing steps cleaned the dataset, made it ready for analysis and modeling.

B. Descriptive Statistics, Exploratory Data Analysis

In the descriptive statistics part of the project, various analyses were conducted to gain insights into the dataset. First, the 'Price' variable was analyzed using the 'description()' function, which produced statistical summaries such as count, mean, standard deviation, minimum, quartiles, and maximum values. The results showed that the dataset had 6,195 items, with a mean price of around $1,068,865 and a maximum price of $9,000,000. To visualize the distribution of prices, a histogram was plotted. The resulting right-skewed histogram indicated that a majority of properties in the dataset had higher prices.

Then, the relationships between the target variable and categorical features were studied. The category columns 'Type', 'Method', 'Regionname', and 'Historic' were analyzed. To demonstrate the relation between each categorical attribute and the price, boxplots were created. It was observed that median prices for different property types were generally over $800,000, with apartments having the lowest median price and houses having the highest. Selling methods showed relatively consistent price ranges. Median prices in the Metropolitan Region were higher compared to the Victoria Region, with the Southern Metro area having the highest median home price. Furthermore, historic houses(older than 50 years) had significantly higher average prices compared to newer homes, but also had more variation in price.

The relationships between the target variable and numerical features were also studied. Scatter plots were used to visualize the relationships between 'Price' and numerical features such as 'Rooms', 'Distance', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'Age', and 'Propertycount'. These plots provided insights about correlations between these features and the price. Notably, the number of rooms showed a concentration around 4 or 5, and there was a negative correlation between the distance from Melbourne's Central Business District (CBD) and the price, with the most expensive homes typically located within 20km of the CBD.

A correlation heatmap was created to examine the relationships between all variables. The heatmap visualized the correlation coefficients between the numerical characteristics,

helping in the identification of any major correlations. Also, it was revealed that the 'Rooms' and 'Bedroom2' columns contained comparable information with just minor differences. As a result, to reduce repetition, the 'Bedroom2' feature was removed from the dataset.

These descriptive statistics provided important information on the price distribution, the correlations between the target and categorical factors, and the correlations between numerical variables. These findings provided the foundation for this project's future research and modeling.

C. Feature Selection, Modelling, Model Evaluation and Validation

In the feature selection phase, various steps were taken to prepare the dataset for regression analysis. Firstly, a new feature called "Price/Rooms" was created, representing the price per room. This feature was introduced to facilitate meaningful comparisons of house values. A set of relevant features was then selected for inclusion in the analysis. These features include 'Rooms', 'Distance', 'Bathroom', 'Car', 'Landsize', 'BuildingArea', 'Propertycount', 'Age', and 'Price/Rooms'.

1) Modeling with Multi-Linear Regression: The dataset was split into training and testing sets using a test size of 20%. A multi-linear regression model was employed, specifically the LinearRegression algorithm, to establish a relationship between the selected independent variables and the dependent variable, which is the house price ('Price'). The model was trained using the training data and subsequently used to predict house prices on the test data. The performance of the linear regression model was assessed using several evaluation metrics. The R-squared value, calculated with the 'explained_variance_score' function from the metrics module, was employed to determine the amount of variance in the dependent variable that can be explained by the model. The R-squared value obtained was 94.2%, indicating a strong correlation between the independent variables and the house prices. Additionally, the Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were calculated. The RMSE value of 168,178.07 indicates the average deviation of the predicted house prices from the actual prices in the dataset.

2) Modeling with Multi-Linear Regression with Categorical Features: To further explore the dataset, LinearRegression model was applied to the dataset after including categorical features. The dataset was preprocessed by removing unnecessary columns ('Date', 'Address', 'Historic', 'Lattitude', 'Longtitude', 'Postcode') and converting categorical variables into numerical features using one-hot encoding. The resulting dataset was then split into training and testing sets (70% and 30% respectively). The linear regression model was trained on the training data and utilized to make predictions on the test data. Similar to the previous model, the performance of the multi-linear regression model with categorical features was evaluated using the R-squared value, MAE, MSE, and RMSE. The R-squared value of 93% indicates that the model explains a significant portion of the variance in the dependent variable. The MAE, MSE, and RMSE values were 123,945.08, 107,846,075,474.96, and 328,399.26, respectively. These metrics provide insights into the accuracy and precision of the predictions made by the model.

3) Modeling with Random Forest Regression: A Random Forest Regression model was employed to explore the dataset further. Similar to previous steps, irrelevant columns were dropped, and categorical features were transformed into numerical representations using one-hot encoding. The dataset was then split into training and testing sets (80% and 20% respectively). To ensure optimal model performance, feature scaling was applied using the StandardScaler. The Random Forest Regression model with 10 estimators was trained on the scaled training data and used to predict house prices on the test data. The performance of the Random Forest Regression model was evaluated using the R-squared value. The R-squared value of 95.6% indicates a strong correlation between the independent variables and the house prices. Additionally, the MAE, MSE, and RMSE values were computed and RMSE found to be 21,968.42.

After conducting a thorough analysis, it was observed that three of the models employed in the linear regression analysis did not exhibit any signs of underfitting or overfitting. The models demonstrated consistent performance across both the training and validation sets, indicating a balanced fit. This suggests that the selected features and the complexity of the models were appropriate for capturing the underlying patterns in the data. The absence of underfitting or overfitting implies that the models have the potential to generalize well to new, unseen data, providing reliable predictions for future observations.

D. Interpretation and Insights

The random forest tree regression model results show that the age, landsize, number of rooms, and whether it is located in suburban areas are the most significant factors that affect housing prices. A larger number of rooms has a positive influence on pricing, implying that larger houses bring in higher market prices. Other factors that influence the price include the number of bathrooms, car spots, building area, and property count. As the visualizations indicate, distance from center and the age of the property, on the other hand, have negative coefficients, suggesting that houses farther from important services and those which are older tend to have lower prices. Furthermore, the price per room, as displayed by the 'Price/Rooms' feature, is taken into account when computing property pricing.

When categorical characteristics are added to the multiple-linear regression model, the interpretation becomes more detailed. Property type and region both have a big impact on house prices. Specific property kinds, such as homes or townhouses, and particular areas have higher effects on the price. Furthermore, particular suburbs and council areas affect the predictions significantly. These findings show the importance of considering both numerical and categorical characteristics when predicting housing prices.

The random forest regression model provides feature importances, which help identify the elements that influence housing prices. The top price predictors include the number of rooms, distance to center, property count, building area, and land size. These align with our basic knowledge, since larger houses in central locations have higher prices. Further, categorical variables such as suburb and region are important, implying that specific locations and neighborhood characteristics influence house prices.

Overall, our findings highlight the effect of multiple variables on home prices and give important insights for real estate stakeholders. This information may be used by real estate agents and property investors to make informed decisions regarding house values and investment plans. Also, policymakers may use these insights to better understand housing market dynamics and develop effective housing regulations. Stakeholders can understand the complexity of the housing market with more assurance and precision by studying the importance of different factors.

IV. DISCUSSION

It is important to understand this study's limitations. Firstly, the research was based on a specific dataset of Melbourne housing market data, and the results may not be relevant to other areas. Also, the analysis depended on the variables provided in the dataset, and there may be other unexplored elements influencing property prices. Additionally, the models applied in this study have their own assumptions and constraints, and different methods of modeling might provide different results. Considering non-linear correlations and also applying other regression techniques may help to capture more complicated patterns in the data. Finally, while the outcome suggested the models' usefulness, it is crucial to read these metrics carefully and consider different factors when making predictions or choices based on the models.

V. CONCLUSION AND FUTURE WORK

In conclusion, the goal of this study was to analyze the Melbourne housing market and predict house prices using a variety of variables. Data quality and integrity were assured by extensive data cleaning, preprocessing, and feature engineering. Exploratory data analysis and descriptive statistics gave insights into the relationships between the target variable (price) and categorical and numerical variables. For predicting housing prices, multiple regression models were used, including multi-linear regression, multi-linear regression with categorical variables, and random forest regression. Among these models, the random forest regression performed the best, with an R-squared value of 95.6%. The complex relationships and interactions between the independent factors and property prices were more accurately captured by this model. The mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) all validated the precision and accuracy for the random forest regression model. Therefore, the random forest regression model is the most suitable model among the applied models for predicting house prices in the Melbourne housing market.

The study provided valuable insights into the factors influencing house prices, such as the number of rooms, distance to center, property count, building area, land size, and categorical variables such as property type, region, suburb, and council area. These results have practical information for real estate agents, property investors, lawmakers, and urban planners, since they will allow them to make more informed decisions about property value, investment tactics, and housing regulations. However, there is opportunity for improvement in this study. Studying the impact of local characteristics and doing geographical analysis could provide a deeper understanding of housing market dynamics. In conclusion, this study contributes to our understanding of the Melbourne housing market and provides significant insights for a variety of stakeholders, with the random forest regression model appearing to be the most successful in predicting home prices in Melbourne. Future research might build on these findings to obtain deeper insights and improve the models for more accurate forecasts in the dynamic housing market.

REFERENCES

[1] Leila Josefowicz. "Heatmaps: A Visual Representation of Data." Data Science for Business 1.1 (2015): 1-12. Web. 4 June 2023.

[2] Leo Breiman. "Random Forests." Machine Learning 45.1 (2001): 5-32. Web. 4 June 2023.

[3] Melbourne Housing Market Data Set. [Dataset]. Melbourne, Australia: Data.Gov.Au, 2016-2022. Web. 4 June 2023.