**Milestone 3: Preliminary Analysis**

Lincoln Brown and James Mulvihill

Bellevue University

Predictive Analytics - DSC630-T302

Professor Hua

**Will I be able to answer the question I want to answer with the data I have?**

So far, it looks like the data we have will be sufficient to at least shed light on our main question. One variable that was not included in the original data that would be helpful is if the property is waterfront. With Melbourne being a coastal city, we suspect that houses right next to the ocean are significantly more expensive than houses located farther away from the ocean. Since we have the geographical coordinates for the homes, we will use geospatial libraries to add an additional binary column named IsWaterfront.

One column we were not able to add that would have been helpful was assessed value. Generally, the assessed value of a house is highly correlated with what it sells for in America. However, Australia has a different tax policy (Land Tax) that only cares about what the value of the land is without any reference to the buildings built on the land (*Land tax*[,](https://www.sro.vic.gov.au/) 2024). Because of this, it is likely they do not have or need assessed house values like they do in the States. One other piece of missing information is about housing quality. If the property is in disrepair, that will obviously have an impact on how much it sells for. But with the existing columns from this dataset, we still should be able to build a good model for this dataset.

**What visualizations are especially useful for explaining my data?**

One useful visualization of our data is a histogram showing the distribution of home prices. The histogram enabled us to visualize the skew and kurtosis of the dataset, as well as look for potential outliers in the data. We could see a positive skew in the distribution, as well as outliers on both the left and right side of the distribution. We used a logarithmic transform to normalize the data and then used the Inter-Quartile Range (IQR) to identify and remove outliers 1.5 times the IQR beyond the upper and lower quartiles.

Additionally, we used maps to visualize the regions within the greater Melbourne area. We included the mean sale price as a descriptor when clicking on the point for each region. Not surprisingly, the mean price was higher near the city center and got lower as the distance from the city center increased. This was further evidenced by a scatter plot of sale price vs distance to the central business district (CBD). The highest prices are found where the distance from the CBD is low, and gradually get lower as the distance to CBD increases.

**Do I need to adjust the data and/or driving questions?**

As part of our data preparation stage, we made several changes to the data. We decided to remove the CouncilArea column from the data because it contained many null values and is redundant because we already have features for Suburb, Address, and Postcode. We also decided to delete many records that had null values. The two columns missing the most values were BuildingArea and YearBuilt. While we could backfill these missing values with median datapoint, we decided not to, since that could skew the results. Even with those rows removed, we still have plenty to work with. Also, we got rid of all types of dwellings (like apartments) besides houses and townhouses.

**Do I need to adjust my model/evaluation choices?**

We are proceeding with our originally proposed models. After observing the Price feature, we have decided to implement a logarithmic transformation on that feature. We have also decided to implement a Standard Scaler for the Linear Regression Model. There is not a strong correlation between the dependent and independent variables. Therefore, we expect the decision tree and random forest models to outperform the simple Linear Regression model.

Currently, we will be using linear regression, decision tree, and random forest models. Of these models, the random forest has performed the best. In addition to these models, we will be implementing a gradient boosting model to see if it performs better than the random forest. After testing cross-validation scores, it appears that the models are overfitting the data. We will be implementing Multiple Correspondence Analysis (MCA) to reduce dimensionality and hopefully resolve some of the overfitting. We will also use hyperparameter tuning to find the optimal parameters for the best performing model.

**Are my original expectations still reasonable?**

Our starting expectations were to predict house prices from our dataset. So far, we are on track for meeting this goal. Checking the cross-value scores for the performance of the models indicates much higher mean squared error values, which can be a sign that the models are overfitting the data. Hopefully, we will be able to rectify this overfitting with MCA and hyperparameter tuning.

**References:**

*Land tax*. Last modified: 4 April 2024. [State of Victoria (State Revenue Office)](https://www.sro.vic.gov.au/). <https://www.sro.vic.gov.au/land-tax>

**Milestone 2: Predicting Real Estate Prices**

Owning a home is a dream that many people have throughout the world. Understanding what a house is worth is an important aspect of attaining that dream. Using predictive analytics, we hope to help people looking for a home in Melbourne Australia determine the estimated value for prospective houses or what their current home’s value may be.

There are several important questions that can be investigated when working with real estate data. These include:

* What features contribute to a house’s price?
* What are good predictors of housing market bubbles?
* How might a home appraiser estimate how much a home is worth?

The first of these questions will be the primary focus of our project. Our goal is to produce a model designed to predict the sale price of a house.

**The Data**

In 2017, Melbourne, Australia was voted for the seventh year in a row the world’s most livable city (Wahlquist, 2020). While it has not done as well since then (mainly because of COVID), we will be looking at data from the housing market as it was in 2017. The data we will be using comes from Kaggle uploaded by user Tony Pino (*Melbourne Housing Snapshot*, 2018). The dataset has over 13,580 records and 21 features that include the suburb the house is in, number of rooms, size, year built and the all-important target variable - price.

**The Models**

We plan to tackle this problem through the production and evaluation of several different models. We will create a baseline model, a linear regression model, decision tree, and random forest. Each of these models has a specific benefit to the analysis of the problem at hand and will contribute to our evaluation of which model has the best performance.

Before we begin model creation, we will take a closer look at the numerical features in the dataset by creating a correlation matrix. A correlation matrix shows the strength and direction of the relationships between individual features within the dataset. We will use this matrix to investigate the relationships between the dependent variable ‘sale price’ and the independent variables in the dataset.

Understanding correlation between the dependent variables and independent variables allows us to focus our efforts on feature selection, transformation, and elimination. Correlation analysis allows us to measure the strength and direction of these relationships. Positively correlated variables will trend in the same direction as the target variable, whereas negatively correlated variables will trend in the opposite direction as the target variable. Using a correlation matrix will allow us to refine our feature selection strategy, which ideally, will improve the robustness of our predictive models.

Initially, we will create a linear regression, which will help us identify any potential linear relationships between the target variable (sale price) and the other features. Linear regression is a simple model that will serve well as a baseline for the evaluation of more complex models. The data may not contain linear relationships, so it will be necessary to investigate the dataset further through a decision tree. Since our dataset contains a moderate number of features, there is a potential for overfitting when using a decision tree. We will use a random forest ensemble to help mitigate the chances of overfitting and improve the model's predictive capabilities.

**Why are these Models Used?**

* Linear Regression - Multiple Linear Regression is a statistical model that can analyze the relationships between the target variable (dependent variable) and several independent variables (Pai, 2021). In this model, the target variable ‘sale price’ is predicted using a linear function of the independent variables. Each independent variable in this equation has a coefficient that shows the direction of the relationship and the independent variable’s effect on the dependent variable’s predicted value. We will use these variable coefficients to develop a better understanding of the individual features' importance to the accuracy of the multiple linear regression model. Also, we will use metrics like RMSE, R2 score, and MAE to evaluate the model's overall fit. Through this analysis, we can develop an understanding of the importance of the individual features to the predictive capabilities of the linear regression model.
* Decision Tree - Decision trees are easy to interpret, which make them an excellent candidate for understanding the decision-making process. We will be able to look at the most important features of the model's performance. Decision trees are robust to non-linear relationships between features making them a suitable candidate for complex datasets that might not have clear linear relationships.
* Random Forest - Random forests are an ensemble method that aggregate the predictions of multiple decision trees to improve predictive performance and are robust against overfitting. Using a random forest model is an excellent way to help mitigate the potential impact of overfitting in a decision tree model. Like a decision tree, we can view the features most important to the performance of the model.

**Plan of Evaluation**

Model evaluation will consist of measuring several different metrics including, Root Mean Squared Error (RMSE), R2 score, and Mean Absolute Average (MAE). Used in combination, these metrics provide a well-balanced assessment of model performance.

* Root Mean Squared Error (RMSE) - Mean Squared Error (MSE) is the metric that measures the average of the squared differences between the predicted and actual sale price (Brownlee, 2021). RMSE is the square root of MAE. RMSE penalizes large errors, which makes it sensitive to outliers in the data. Lower RMSE values indicate better performance. RMSE values are interpreted in the context of the dependent variable.
* R2 Score - The R2 score, or Coefficient of Determination is a measure of goodness of fit. The R2 score explains how much of the proportion of variance in the dependent variable is predictable from the independent variables (Rowe, n.d.). R2 scores are measured between 0 and 1, with 1 indicating a perfect fit. Perfect R2 scores are a strong indication of overfitting.
* Mean Absolute Error (MAE) - This metric is a measure of the average absolute difference between the predicted values and the actual values in a dataset (Zach, 2021). MAE values are measured in the same context as the dependent variable, making them easy to interpret. In the case of sale prices, the MAE represents the average absolute difference between the predicted sale price and the actual sale price.

Each of these metrics are individually important for evaluation, but when communicating with our stakeholders, we will primarily use MAE because it can be easily interpreted as the average absolute dollar amount that the prediction deviates from the actual sale price.

**What We Hope to Learn**

The goal is to find a model for predicting house prices given the features of the dataset. Additional questions that we will investigate are:

* Which features have the most significant impact on house prices?
* Are there any differences in price for different neighborhoods?
* Which model performs the best?
* Which features were most important for the model's performance?

**Risks and Ethical Consequences**

One risk is that bad data could skew our results. Inside the dataset, there are a few houses that have huge square meterage. This is unlikely to be the case, especially if the price is around average. The data will need to be carefully cleaned to make sure rows like this do not skew our results. Another risk would be that the data does not contain key factors that would influence house prices. Most notably, the quality of the house is not stated in the dataset. While year built might be correlated with this, we do not have a straightforward way of knowing if the house is in disrepair.

The data is collected from an open-source repository on Kaggle, so there are no ethical concerns with how the data was obtained. If there are any other ethical concerns that emerge, we’ll seek guidance at that time.

**Plan for the Unexpected**

It is possible that we might need to use a different dataset. This could happen if there is confusion about what some of the columns in the Melbourne dataset signify. For example, one column is titled “Rooms,” which might just be bedrooms, but, coming from Australia, might mean something else. As a backup, we will use the Ames Housing Dataset found on Kaggle (*Ames Iowa Housing Data*, 2020).

**References:**

*Ames Iowa Housing Data*. (2020, March 18). Kaggle. <https://www.kaggle.com/datasets/marcopale/housing/data>

Brownlee, J. (2021, February 15). *Regression metrics for machine learning*. MachineLearningMastery.com. <https://machinelearningmastery.com/regression-metrics-for-machine-learning/>

*Melbourne Housing snapshot*. (2018, June 5). Kaggle. <https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot>

Pai, A. (2021, October 6). *Multiple linear regression — machine learning works*. Machine Learning Works. <https://www.machinelearningworks.com/tutorials/multiple-linear-regression>

Rowe, W. (n.d.). *Mean square error & R2 score clearly explained*. BMC Blogs. <https://www.bmc.com/blogs/mean-squared-error-r2-and-variance-in-regression-analysis/>

Wahlquist, C. (2020, February 3). Melbourne “world’s most liveable city” for seventh year running. *The Guardian*. <https://www.theguardian.com/australia-news/2017/aug/16/melbourne-worlds-most-liveable-city-for-seventh-year-running>

Zach. (2021, October 4). *MAE vs. RMSE: Which Metric Should You Use?* Statology. <https://www.statology.org/mae-vs-rmse/>