Logo, company name

Description automatically generated

MMA 867: Predictive Modeling

Individual Assignment 1

For: Professor. Anton Ovchinnikov

By: Derek Luening

Student Number: 20252039

Program: MMA Section 1 Team Bremner

Kaggle Competition: House Prices – Advanced Regression Techniques

Total Teams at time of last submission: 8893

Leaderboard position at the time of the last submission: 2297

Graphical user interface, text, application

Description automatically generated

**Competitions:**

1. [New York City Taxi Fare Prediction](https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data?select=test.csv)
   1. The New York City Taxi Fare Prediction competition required participants to predict the fare amount for a taxi ride in New York City given the pickup and drop-off locations. There have been 1,483 teams that have participated in this competition, allowing for regression techniques to be used. However, I decided not to use this dataset due to computer limitations. The dataset was too large for my computer and took too much time to load, manipulate and export.
2. [PUBG Finish Placement Prediction](https://www.kaggle.com/c/pubg-finish-placement-prediction/data)
   1. The PUBG Finish Placement Prediction competition required participants to predict the final placement/ranking for each player. There have been 1,528 teams that have participated in this competition, allowing for regression techniques to be used. The game PUBG was wildly popular from 2017 – 2018. The game dropped players on an island, and then the mission is to be the last player alive. I decided not to use this dataset due to a lack of interest in the game itself. It is not something that I have actively played or am interested in, and as a result, I decided to pursue a different project.
3. [House Prices – Advanced Regression Techniques](https://www.kaggle.com/c/house-prices-advanced-regression-techniques)
   1. The house prices – advanced regression techniques competition required participants to predict the sale prices of residential homes in Ames, Iowa, from 2006 - 2010. There have been 8,993 teams that have participated in this competition, allowing for regression techniques to be used. I decided to choose this competition mainly out of interest. Over the last few months, my wife and I have sold our home in Ottawa, Ontario, and are in the process of moving to the Greater Toronto Area. Given that we're in the process of selling and buying, I wanted to see how accurately I could predict the sale prices for this competition.

Once I decided that I wanted to pursue the House Prices – Advanced Regression Techniques, I focused on methodically completing the project to ensure that I could stay organized and ensure that I did not forget anything. Below are the steps that I followed to complete the task:

**Data Analysis**

Before I started cleaning and manipulating any data, I wanted to learn more about the dataset and understand how the variables influenced sales prices. To conduct this preliminary analysis, I leveraged Tableau. Given that there were a significant number of variables in the data set, Tableau provided me with a very efficient way to analyze a considerable number of variables. I began my analysis by thinking about the significant factors that influence house prices: Homestyle/layout, square footage, neighborhood, total bathrooms, total bedrooms, and overall quality of the home. My analysis of these variables confirmed my intuition on what would influence the final sale price. However, I also learned that some other variables played a crucial role in influencing the sales price. These variables included how many cars a garage can hold, garage area, and basement square footage. I had not initially considered these variables due to my own biases on what is essential in a home. However, investigating these variables during the first step of the process allowed me to understand what may be relevant to my model. Screenshots of my data analysis can be found below, and my dashboard can be found in my submitted zipped file.

**Data Cleaning**

Once I understood which variables are correlated with the sales price, I began to investigate both the train and test sets for missing values and inconsistencies independently. However, before I began manipulating the dataset, I thoroughly reviewed the data dictionary. The data dictionary is very detailed and provided a firm understanding of what each variable is and the values that populated them. Examining the data dictionary allowed me to begin thinking about how the variables will interact with each other and what variables I may need to create for my model.

Now that I understood each variable, I began identifying the data types of each variable. I uncovered that the variables included in this dataset were only variables with the character/integer/or numeric datatype. This was an essential step in the data cleaning process. My understanding of the variables from the data dictionary allowed me to determine which variables need to be converted into factors.

Given that I conducted an extensive analysis of the data and the variables’ structure, we can now identify any missing data in our dataset. I decided to clean both the train and test sets independently. I found by carrying out the cleaning process this way allowed me to stay more organized. In the training dataset, I determined that there were 6884 missing values in the training set. There were 6998 missing values in our testing set (Refer to figure 1 and figure 2 to understand which variables had missing data). We have determined that most of our "missing" data is categorical and not missing from the analysis that we conducted to get to know our data. By leveraging the data dictionary, we determined that most of the NA values represented that a feature that is not present in the house. As a result, I could impute these values with "None" to indicate that this feature does not exist for this property. However, not all the values were categorical. For some of the variables, they were missing and needed to be replaced. As these missing values were not missing in large quantities, I decided to impute these with the mode of the variable.

Given that this occurred for less than 1% of our data, I do not believe imputing with the mode would skew the results of our model. Finally, there were numerical values that needed to be replaced, like total basement square footage. For these variables, I decided to impute these values with 0’s as they did not have the feature in their house, and they were of a numeric data type.

After I cleaned the data of any missing values, I converted all my variables with a character data type to factors. Correcting these variables to factors allowed me to capture the levels that existed within each variable, assisting and providing flexibility in our modeling process. In addition to converting my character variables to factors, I flipped the variables MSSubClass, YearBuilt, MoSold, GarageYrBuilt, and YrSold, which were variables of the integer data type into factors. Transforming these variables into factors would allow me to capture the difference between the different years in my model.

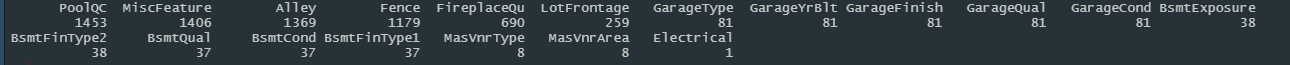


Figure 1. Missing Training Data

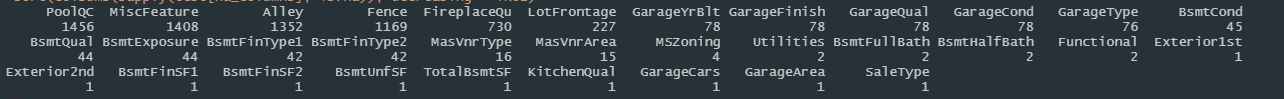


Figure 2. Missing Testing Data

**Feature Engineering and Combining Datasets**

To ensure that the feature engineering I conducted was captured in both the training and testing datasets, I combined them into one master dataset. While I believe that the initial dataset captures many variables considered when buying a home, I felt that three key variables were missing. I created an additional three columns that captured the house's total square footage, the total number of bathrooms in the house and made a binary variable that identifies if a remodel took place or not.

The dataset does an excellent job of identifying the square footage of specific areas in the house. Still, it does not provide an accurate representation of the total square footage in a home. As someone who is currently actively looking for a home, this is one of the considerations that I believe is extremely important to capture. The total square footage of your home could be one of the critical difference makers when buyers determine what they feel the value of your home is as they look at comparable properties when purchasing.

Similarly, this dataset does an excellent job of identifying the number of bathrooms within specific house areas. Still, it does not provide a value for the total bathrooms in the house. Like total square footage, the number of bathrooms and precisely full bathrooms that a home can offer is significant to buyers. By capturing the total number of bathrooms, I believe that my model will identify a critical characteristic that buyers genuinely value in a home. Also, in this dataset, half baths are indicated with whole numbers. To compensate for the fact that a half bath is not equal to a full bathroom, I penalized the half bathroom variables by 0.5.

Finally, I created a dummy variable that indicates if a remodel took place or not. There is no way to tell what was remodeled from the initial data set if a remodel occurred. However, while we cannot determine what happened, I still felt that it was essential to capture that one that took place. Depending on what the homeowner decided to update, it could add significant value to their home, as Professor. Ovchinnikov displayed in our first class, humans focus on visuals first. If your home is not updated, most people have a hard time looking past the dated kitchen and bathrooms and, as a result, are not likely to want to pay over asking for your home. Thus, I felt it was essential to not only understand what year a remodel may have occurred in but to identify if one happened or not.

**Model**

Once I had successfully organized and cleaned the data, I sought to build a simple model that could accurately predict housing prices. However, before I began modeling, I further researched factors that strongly influence house prices. The consistent variables that came back were square footage, total bathrooms, total bedrooms, neighborhood, and upgrades. With this information coupled with the information I had learned in my data analysis, I believed I had enough information to build a model.

I initially began modeling with a multiple linear regression model that included a limited number of variables. While building a model with a limited number of variables to ensure that it was simplistic, I found it very difficult to achieve an acceptable Mean Absolute Percentage Error (MAPE). I was regularly achieving a MAPE larger than 15%. As a result, it informed me that my model was not good enough, and as a result, it performed very poorly on Kaggle. By experiencing this from the onset, I determined quickly that predicting housing prices may be more complicated than I had initially thought. As a result, I believed that I would need a more robust model and decided to focus my efforts on building a model with a LASSO regression.

With my focus now on building a model with a LASSO regression, I decided to try a different strategy to determine which variables belong in the model. I first created a stepwise regression with all 84 variables to help determine what may or may not be essential to include. Once I received this output, I had the significant variables in the LASSO regression. The first model I ran using this strategy led to more promising results than the model using multiple linear regression. However, the results of the model were still generating MAPE’s between 13% - 15%. I continued to create interactions, add and remove variables one by one to see how this impacted the MAPE. This process was not generating any significant differences from what I had experienced from running the stepwise and then immediately putting that output into a LASSO regression. After many iterations, I decided to change my strategy again.

I decided to go back to my intuition and what my data analysis had informed me. Include variables related to the location, total square footage, total bathrooms, and total bedrooms. When I included these variables and the granular square footage (above ground living area, basement square footage), bathroom (full baths, half baths, and basement half baths, and full baths), and bedroom variables, in a LASSO regression model, I achieved a MAPE between 11% - 12%. To further reduce the MAPE, one by one, I logged the numeric variables that were able to be logged and already included in the model. I evaluated how each change impacted the MAPE and decided if that transformation should be kept in the model or not. Next, I started adding additional variables into the model one by one, assessing how the addition of that specific variable impacted the model's performance. Finally, I started creating interaction variables for this version of my model. To make these interactions, I kept in mind what I learned from my research and data analysis; I only created interaction variables that I felt would be of business value and decided not to include random interactions between variables. The interactions that I made were built around square footage, home quality and condition, type of home, location, size of property, if a remodel occurred, the year built, and year sold. Once I included these variables in my model, I reduced my MAPE to 8.43%, with a lambda penalty of ~-5.5 (figure 3 and figure 4 display my crossval plot and my LASSO plot, respectively). This allowed me to score 2297 out of 8893 participants, which put me in the top 25.8% of the House Prices - Advanced Regression Techniques competition.

Chart

Description automatically generatedShape

Description automatically generated with medium confidence

Figure 4. LASSO plot for the final model

Figure 3. Crossval plot for the final model

**Conclusion**

Overall, I found this to be a very challenging yet rewarding project. I ran into many roadblocks with data cleaning and modeling. I encountered a significant number of errors that I was forced to work through, which was an exceptional learning experience. This was also the first time that I have built a model from start to finish, and after having completed the project, it has given me a sense of the challenges that come up when modeling. It also displayed how our own biases can interfere in creating a model and what implications we could face. I also learned that while a topic like the sales price of a home may not seem very complex, there are other factors at play that have more influence and impact than you may initially believe.