IST 707 Final Project – Predicting Home Prices

Mateo Arbelaez

2021

Introduction

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses in Boston’s suburbs using The Ames Housing dataset, which was compiled by Dean De Cock for use in data science education. Once a model is identified, we will use it to predict the price of homes. A model like this would be very valuable for a Real Estate agent who could make use of the information provided in a daily basis.

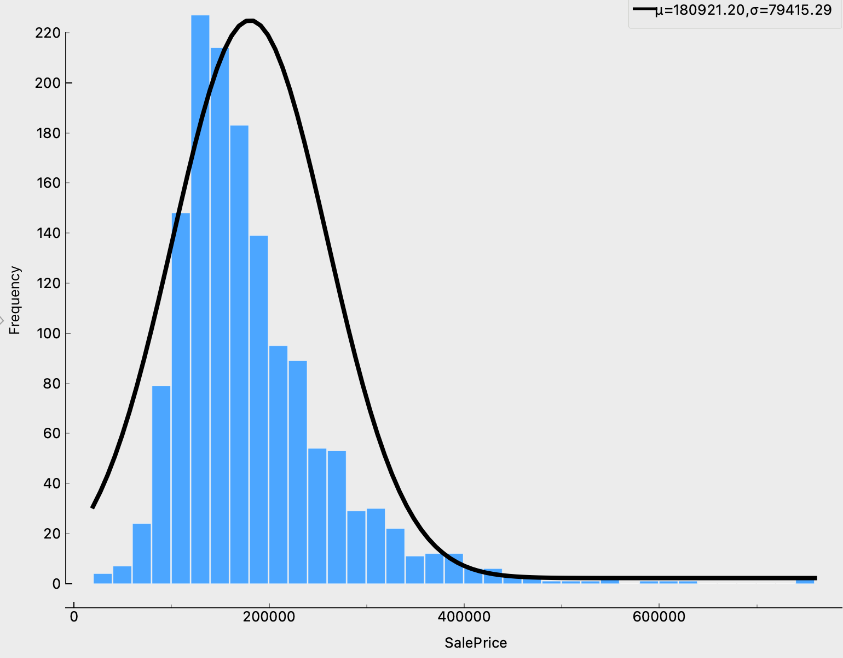
# Data

Our data comes from a Kaggle competition named “House Prices: Advanced Regression Techniques”. It contains 1460 training data points and 79 features that might help us predict the selling price of a house.

## Understanding the Data

In addition to the data files, Kaggle also provides a very helpful data description file that will help us what each feature represent. This will be very important as we will identify correlations and will need to understand what could be driving certain correlations.

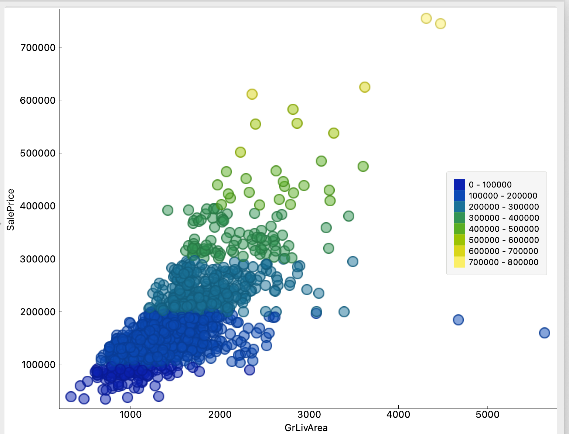
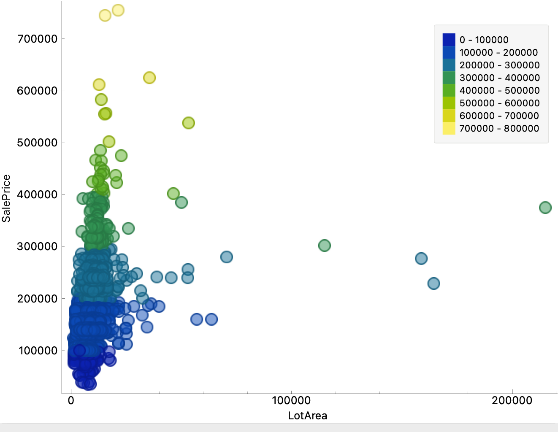
Let’s begin our data discovery by looking at what we are going to predict, Sales Price. We can see in the distribution of the ‘Saleprice’ int eh below chart. The charts shows that most of the concentration is between $100,000 and $250,000, with some outliers closer to $700,000.



Next, we will take a look at a few scatter plot to determine if features such as the greater living area (square feet), lot area, garage area and year built have any correlation to the SalesPrice.

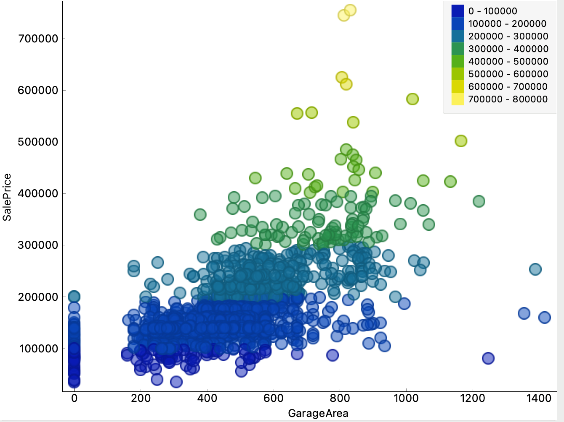
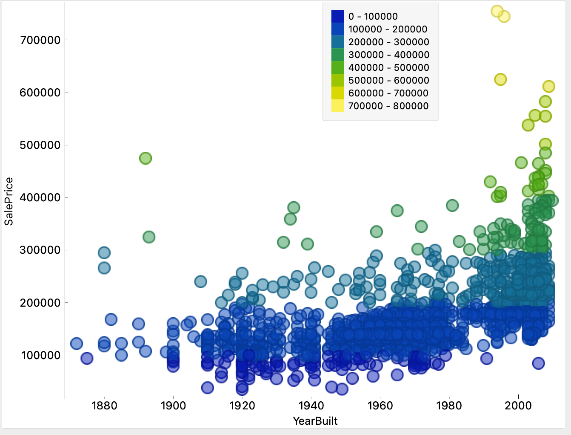
The left plot shows the greater living area (square feet) against the sale price. We could expect that the larger the area the higher the price. Based on the graph, it appears we are mostly correct, however there are some small living rooms at somewhat high prices.

The right plot shows the lot area (square feet) against the sale price. We could expect that the larger the area the higher the price. Based on the graph, it appears we aren’t correct, the size lot does not seem to correlate with the price, as majority of our records are concentrated on the left of the graph.

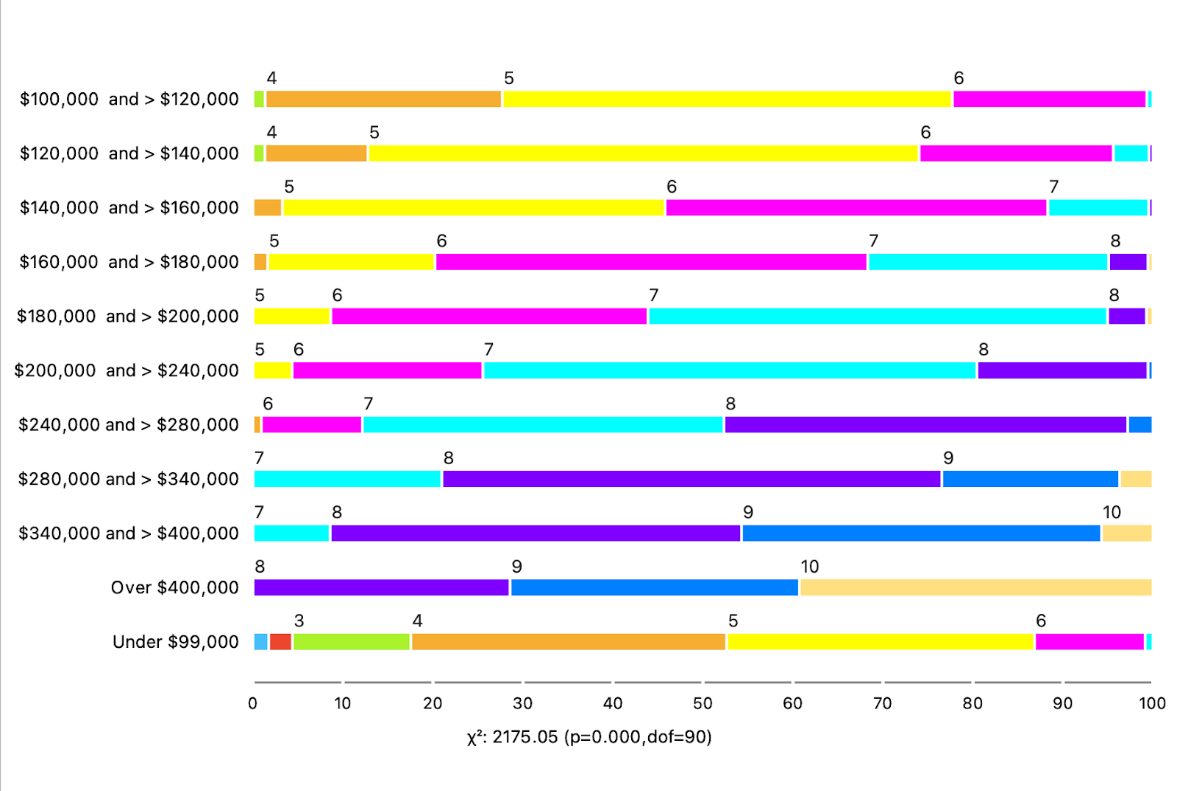
 

The left plot shows the garage area (square feet) against the sale price. We could expect that the larger the area the higher the price. Based on the graph, it appears we are mostly correct, however there are some small garage areas at somewhat high prices.

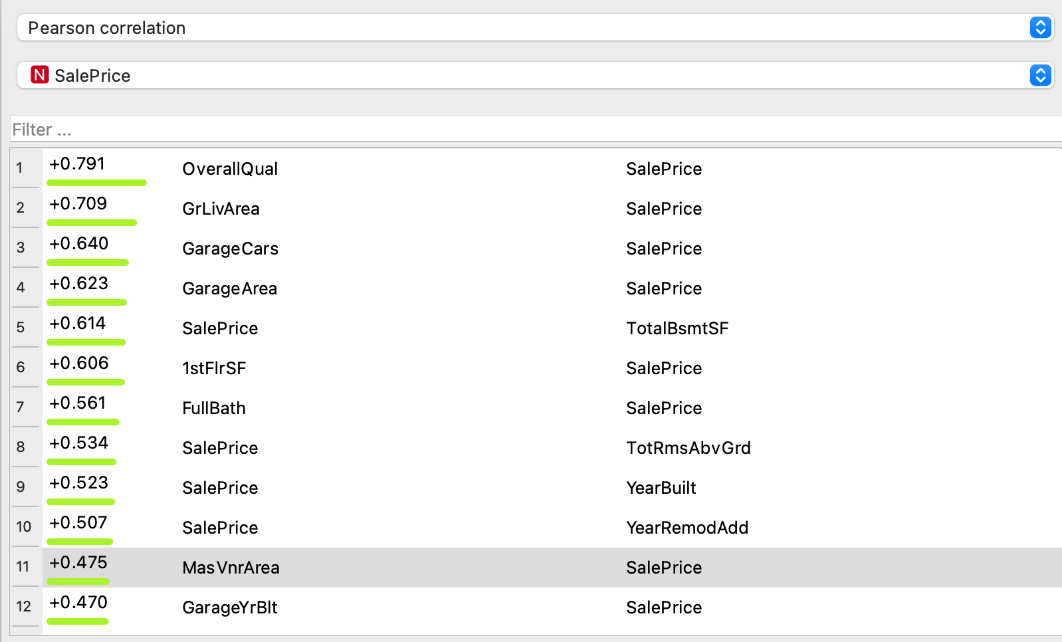
The right plot shows the year built against the sale price. We could expect that the newer the house the higher the price. Based on the graph, it appears we are mostly correct, however there older houses at somewhat high prices.

Given the distribution of the SalePrice, we have grouped the SalesPrice into 11 groups to help our modeling and understanding of the data. Using this new feature and the overall quality is much higher among the higher prices.



Now that we have reviewed the correlation of some of the most obvious features, we will take a look at what other features have a high correlations. The image below shows some of the ones we tested and others.



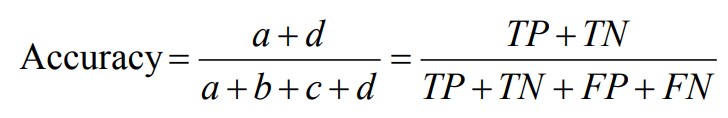
## Predictive modeling

In section, we will test different techniques necessary for a model to make a prediction. One of the challenges we will face is how to properly measure the quality of a given model without quantifying its performance on the training and testing. There are several ways to approach this and typically is it done using some type of performance measurement, whether it is through calculating the error, the goodness of fit, or the accuracy of a model. Regardless of how we measure the quality of a given model, it is critical that the same metric is used to compare all models.

There are several ways to validate a model’s results, we will be using a technique called Cross Validation. This technique is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods. The number of k-fold will be 10.

For the purpose of our testing, we will be using the following performance metric to select the best fitting model:

**Classification Accuracy**: the fraction of predictions our model predicted correct.



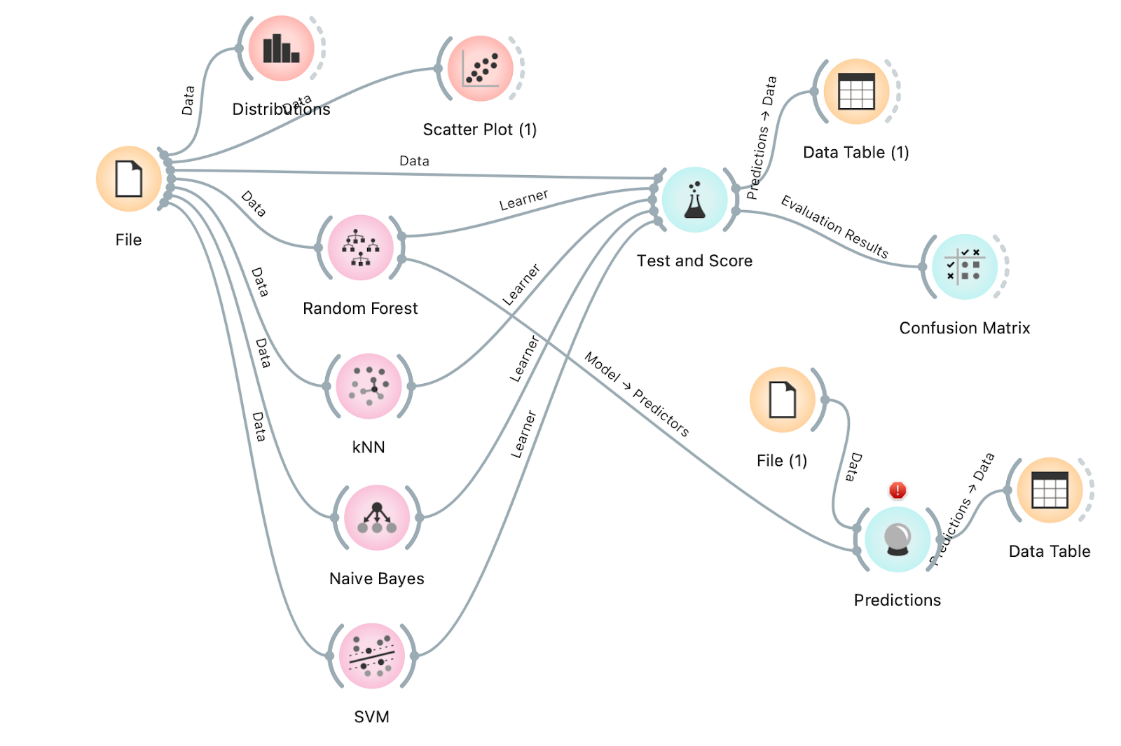
This section will cover the following five methods:

* Naïve Bayes Model
* Random Forest Model
* K-Nearest Neighbors Model
* Support Vector Machine Model

## Predictive Results Modeling

For this project, we will be utilizing the powerful tool Orange Data Mining (Orange). Orange is an open-source data visualization, machine learning and data mining toolkit. It features a visual programming front-end for explorative rapid qualitative data analysis and interactive data visualization.

The diagram below shows the machine learning techniques and the workflow to creating each model.

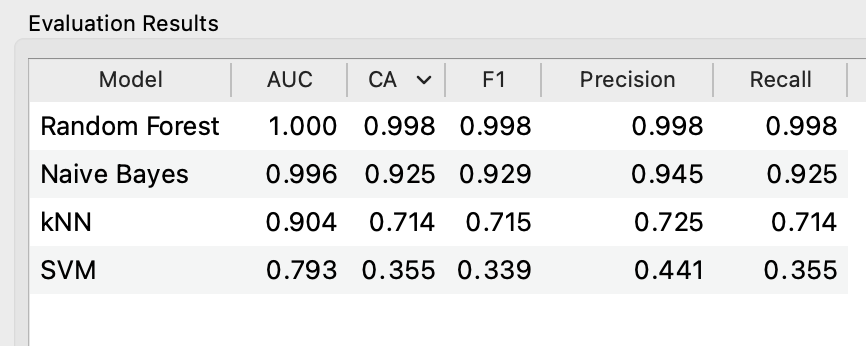


**Naïve Bayes Model**

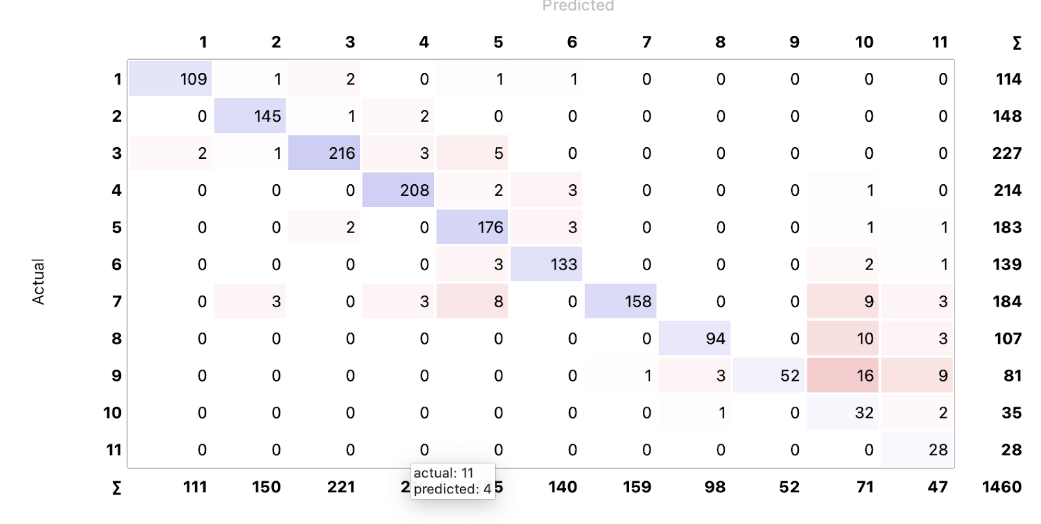
This model is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

**Result Naïve Bayes Model:**

The screenshot below shows the results of the “Naïve bayes” model. We can see that the classification accuracy for the model is .925, is a very promising start.



The confusion matrix for the Naïve Bayes Model shows where the model incorrectly made a prediction. It appears the model’s performance for sales prices in group 10 and 11 is not as good as the other groups. This could be caused by high range of prices in these groups.

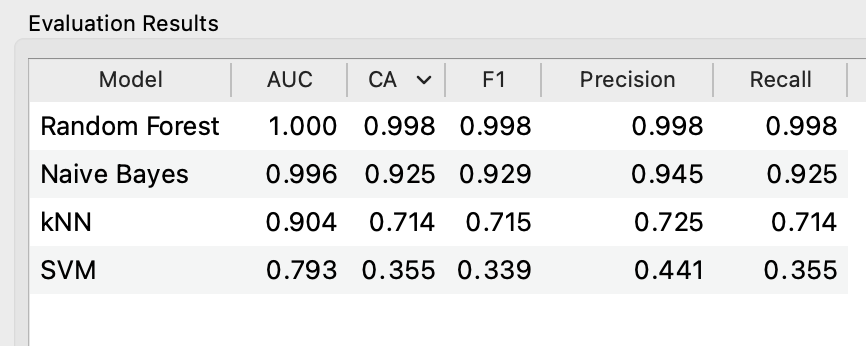


**Random Forest Model**

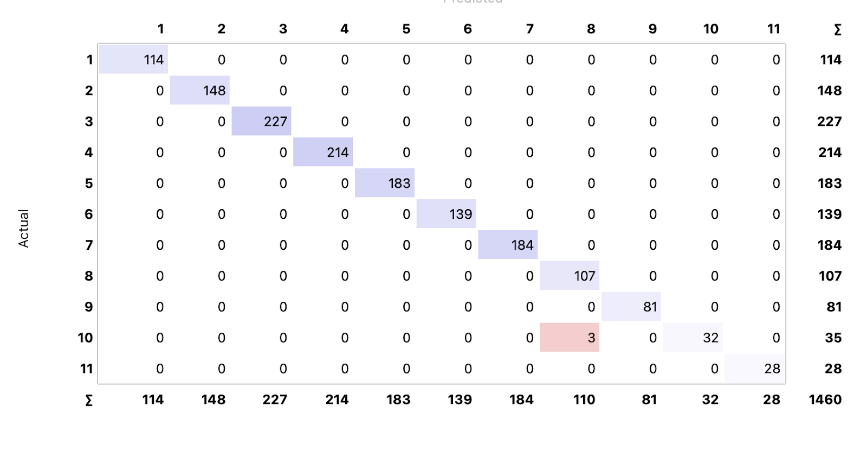
It is a learning method in which number of decision trees are constructed at the time of training and outputs of the modal predicted by the individual trees. RF act as a tree predictors where every tree depends on the random vector values.

**Result - Random Forest Model:**

The screenshot below shows the results of the “Random forest” model. We can see that the classification accuracy for the model is .998, which is almost perfect and definitely better than the Naïve Bayes Model.



The confusion matrix for the Random Forest Model shows where the model incorrectly made a prediction. It appears the model’s performance for sales prices in group 8 is not as good as the other groups as it misclassified 3 records.

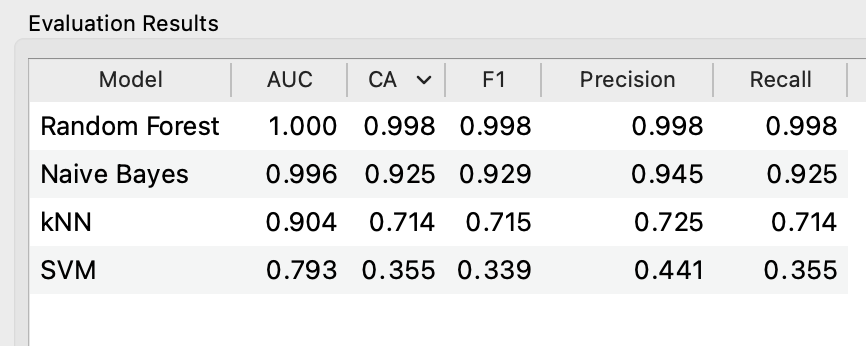


**K-nearest neighbor Model**

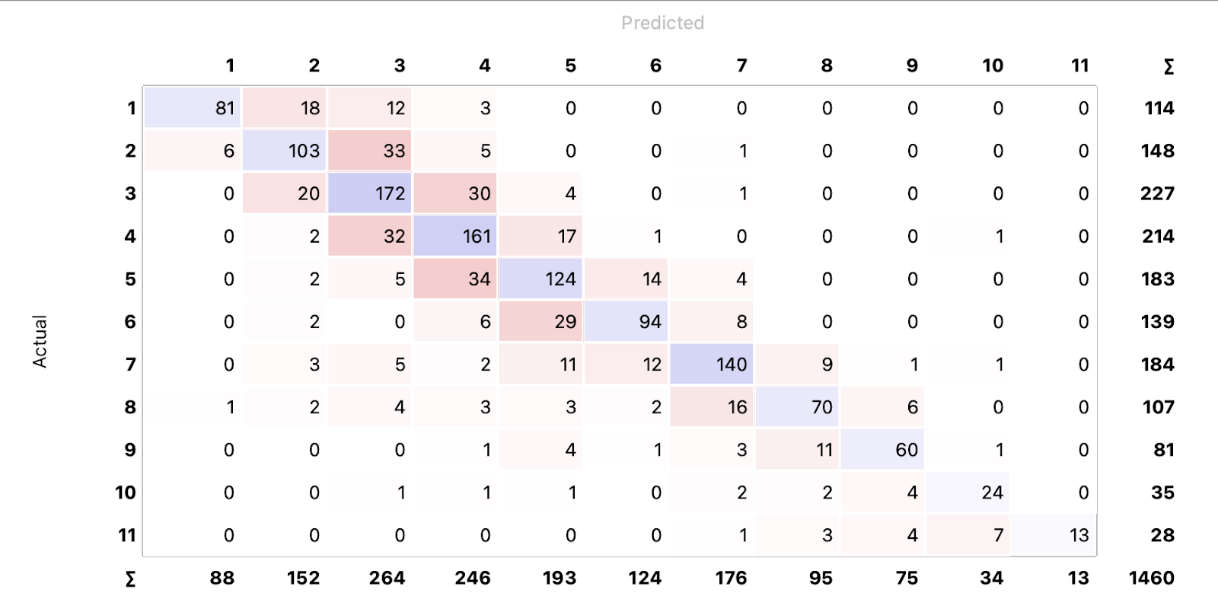
K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

**Result - K-nearest neighbor Model:**

The screenshot below shows the results of the “kNN” model. We can see that the classification accuracy for the model is .714, which is not a good as the other models.



The confusion matrix for the KNN Model shows where the model incorrectly made a prediction. It appears the model’s performance for sales prices across all group is inaccurate.

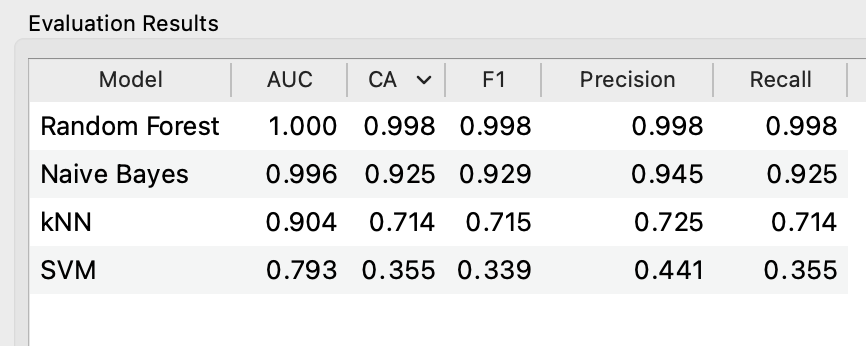


**Support Vector Machine Model**

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points.

**Result - Support Vector Machine Model:**

The screenshot below shows the results of the “SVM” model. We can see that the classification accuracy for the model is .355. This model will not be considered.



Now we take the results from these models to determine which mode should be used to predict the sale price of a home. It was clear that the Random Forest model outperformed all the other models. This model can be used as the engine to help different stakeholders in the Real Estate industry.