# CS-559-WS Group 12 Report

## Training Data Process

Before the models were trained, the data was processed. First, certain features were dropped based on several conditions. Next, nans were replaced instead of being dropped from the dataset. Categorical features were then encoded as integers. Afterwards, all features were standardized. Finally, PCA was used to drop the number of remaining features to thirty-five. Below is a more in-depth description of each step of the process.

Features were dropped if they met one of several conditions. If a feature had a high percentage of values that were nan, then it was dropped. If a feature was highly correlated with another feature, then one of those two features would be dropped. Categorical variables where most samples fall within one category are not good predictors, so they will be dropped. Any features that lacked relevance were dropped, such as sample id numbers.

When dealing with nan values, suitable nan replacements were found. Samples with nans were not dropped so that the models could be trained on the maximum amount of data. For numerical features, nans will be replaced with the average value. For categorical features, nans will be replaced with the most common category. With these replacements ready, any nans were replaced by these values.

Categorical features were encoded as integers. The Sklearn LabelEncoder was used to handle the transformations.

All features and the prices were standardized with the Sklearn StandardScaler. The prices were standardized to retain numerical stability and prevent exploding/vanishing gradients. When computing the MSE, the prices are destandardized before calculating the MSE.

The remaining feature count was reduced to 35 using PCA.

## Training Process

Once the data was preprocessed, kmeans was used to cluster the data. The kmeans cluster ids were used as ground truth labels for the classifier. Using the elbow method, it was determined that three was the best k value. A support vector machine was trained to predict which cluster a particular sample belonged to.

The classifier’s predictions were used to determine which model each sample belonged to. This process created k mini datasets, that would be used to train k models. If the classifier predicted that a sample belonged to model 0, then that sample and its ground truth would contribute to training model 0.

Once the data was split into groups, a small segment of each training set was reserved for validation. Each group model was hyper tuned and evaluated against the validation data. Once the best hyperparameters were found, the final models were trained.

## Test Process

Test predictions were generated by running the preprocessed test data through the trained models. The test data was preprocessed using the same process as the training data. The same features were dropped. Any nans were replaced with the values selected during the training process. Labels were encoded using the LabelEncoders that was fit during training. The Data was standardized using the StandardScaler that was fit during training. The remaining feature count was reduced to 35, using the PCA model that was fit during training.

Once the test data was preprocessed, the classifier was used to assign each sample to a cluster. Using the classifier predictions, the samples were split up into their assigned groups. Each group of data was then passed into the group’s respective model for inference. Since the prices were standardized, the StandardScaler’s inverse\_transform function was used to destandardize the predicted and ground truth prices. With the destandardized prices, the mse for the test data was then calculated.