Final Report

Group: Need For Geek
Bowen Yi A59010676
Jiatai Li A59013807
Zhenghua Ye A59006278
Haoning Qiu A59011030

executive summary

In this project, we obtained sale prices data for real estates and stocks, wanted to use machine learning methods on computer to make an effective price prediction and ultimately feed back into investment decisions. In this part, we want to get the best price forecast for real estate. The basic idea of our prediction is to first preprocess the data, try different machine learning models, such as: random forest, knn, logistic regression, etc., get different results and then compare the accuracy of training, validation and test data to find the most successful model and best suitable prediction of sale price.

Data processing

At first, we used the OneHotEncoder package to handle the data. One-hot encoding is the process by which categorical data are converted into numerical data for use in machine learning. Categorical features are turned into binary features that are “one-hot” encoded, meaning that if a feature is represented by that column, it receives a 1. Otherwise, it receives a 0. So one-hot encoding is actually a process by which categorical data (such as nominal data) are converted into numerical features of a dataset. This is often a required preprocessing step since machine learning models require numerical data.

Then, to fill in the blank data, we used the mean of each feature, and dropped these features which lose more than 50% data.

At last, we used StandardScaler package to de-mean and variance normalization. And it is done for each feature dimension, not for the sample.

Models(used):

To get optimal results, we used different machine learning models to make predictions, and then compare the accuracy of training, validation and test data to find the most successful model and best suitable prediction of sale price.

1.Random Forest

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees.

But according to our test, the random forest did not perform well in this task. The validation accuracy is close to 0. At last we did not use this method to complete this competition.

2.KNN

In statistics, the k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method. It is used for classification and regression. k-NN is a type of classification where the function is only approximated locally and all computation is deferred until function evaluation. Since this algorithm relies on distance for classification, if the features represent different physical units or come in vastly different scales then normalizing the training data can improve its accuracy dramatically.

The same bad outcomes with Random Forest. It could not precisely predict the real estate sale price, so we finally dropped this model.

Logistic Regression

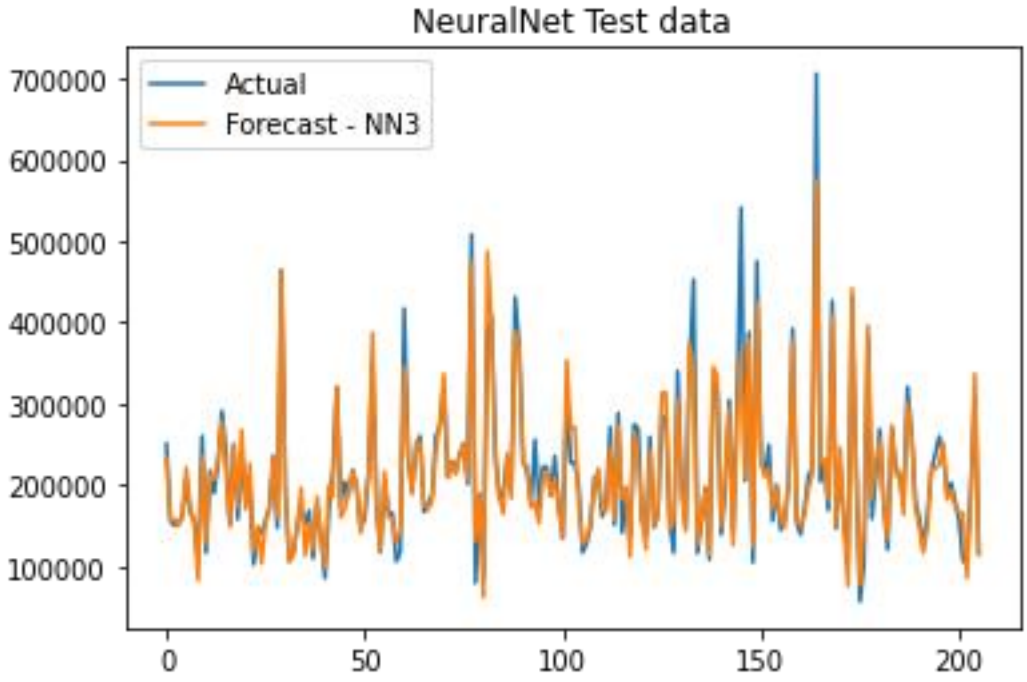
In statistics, the logistic model is a statistical model that models the probability of one event (out of two alternatives) taking place by having the log-odds (the logarithm of the odds) for the event be a linear combination of one or more independent variables ("predictors"). In regression analysis, logistic regression is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by a indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 and 1, hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.

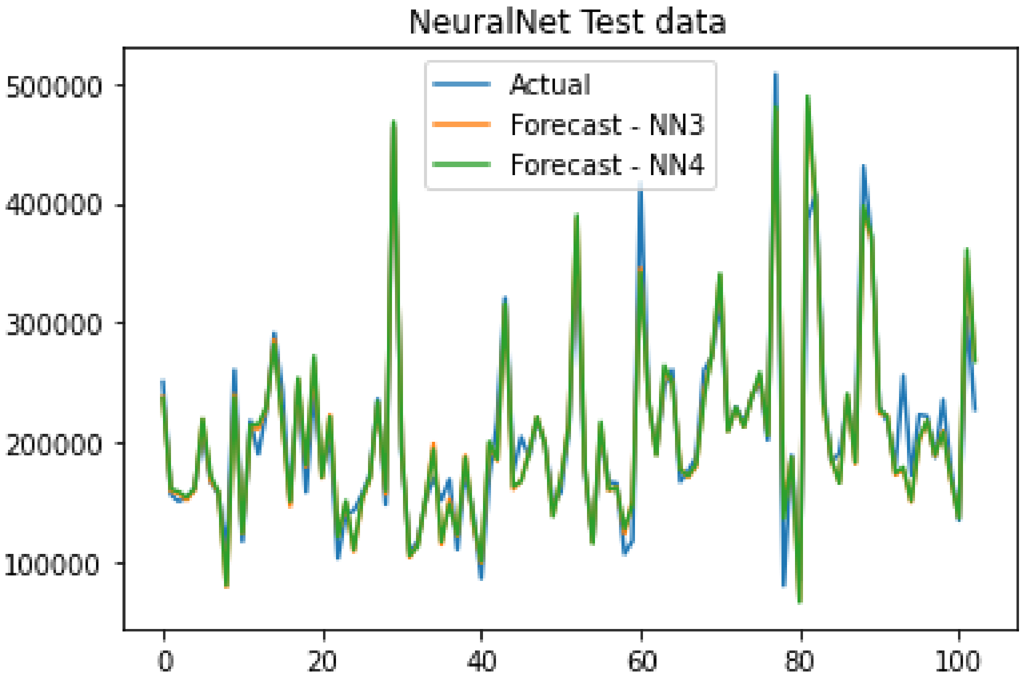
The Logistic Regression provided a more precise model comparing with Random Forest and KNN. However, it did not allow us to get particularly high scores in the kaggle competition, so we thought of trying to use deep learning methods.

MLP

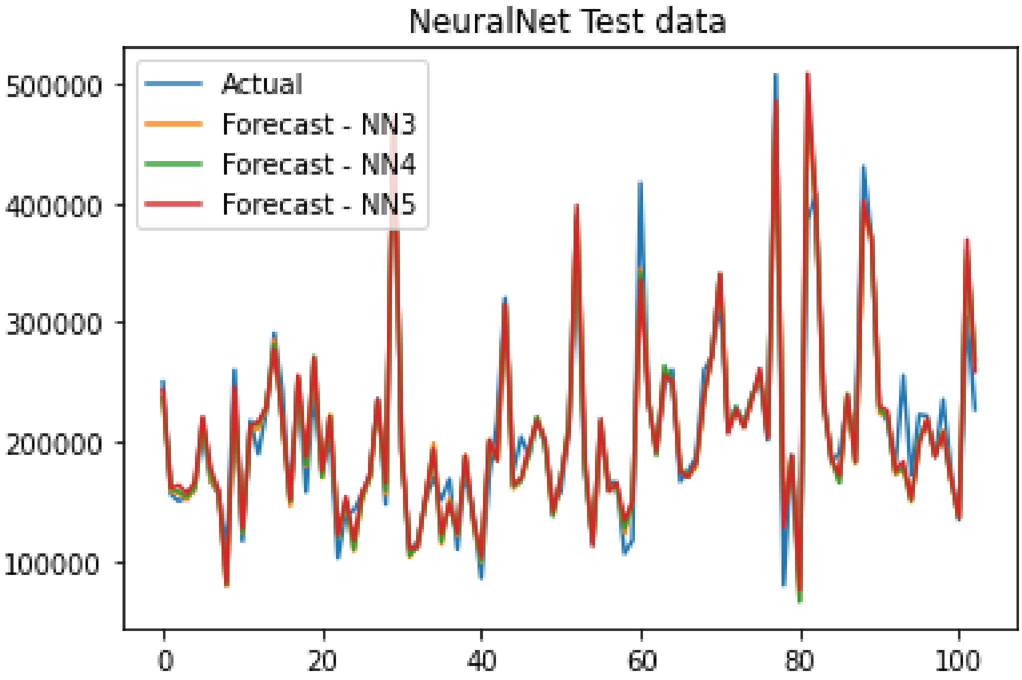
A multilayer perceptron (MLP) is a fully connected class of feedforward artificial neural network (ANN). An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable. What is most important is that the MLP consists of three or more layers (an input and an output layer with one or more hidden layers) of nonlinearly-activating nodes. Since MLPs are fully connected, each node in one layer connects with a certain weight to every node in the following layer.

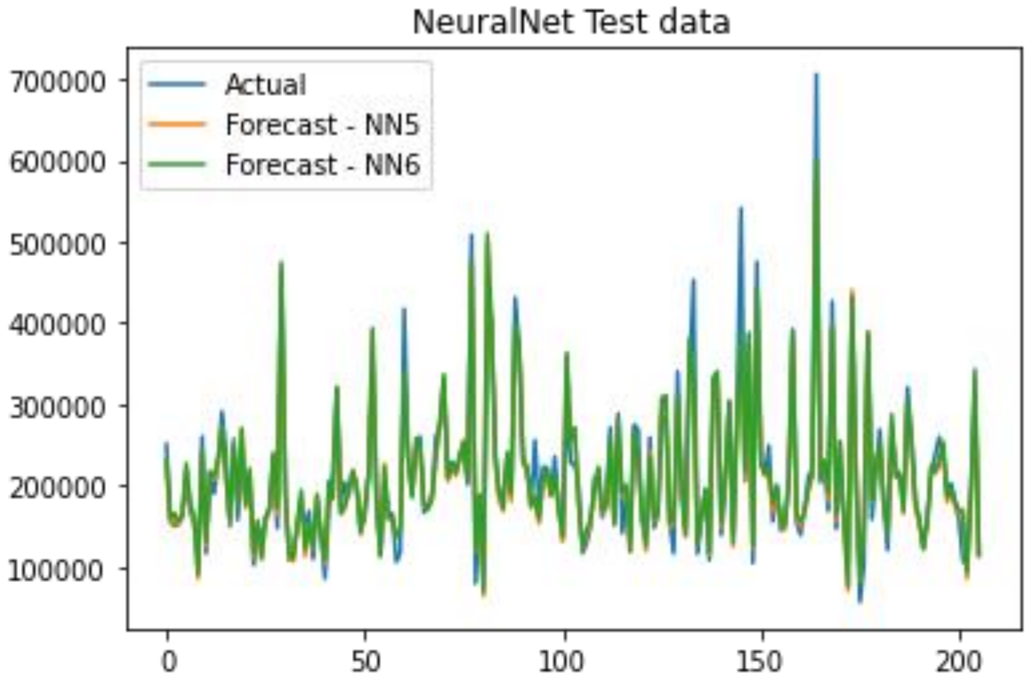
At first, we used Sequential() function to scale the input variables. Then we first used three hidden layers and four hidden layers to get sale price prediction. The result is shown below:





To get a better prediction, we added the hidden layers and compared their results:





Conclusion:

Before conducting the analysis, there was a consensus within our group that regression models may be more effective than classification models for price forecasting, since the final output is a fixed price. Even so, we tried different regression and classification models.

As above, in the process of solving this problem, Random Forest, KNN and Logistic Regression did not provide very good results. MLP gives the best results, but the selection of internal parameters of its model, such as the number of hidden layers, also affects the prediction results of the model from different aspects. However, machine learning and even deep learning are a learning process without standard answers, so the model may still have a lot of room for improvement to achieve a better prediction accuracy.