COMP1816 - Machine Learning Coursework Report

Emirhan Ülgen - 001150599
Word Count: No more than 2000 words (1997)

1. Introduction

In this modern world, artificial intelligence is getting popular day by day. Machine learning surely has the share of this development since it is a part of artificial intelligence and it is essential for machines, as the name implies. It enables machines to perform several operations without the help of humans. With the significant improvement of artificial intelligence in technology, machines have a great importance in society. One of the abilities in machine learning is making predictions. This coursework aims to test the regression and classification abilities that take part in predicting some values and putting items in separate categories. To meet this goal, different models used in such abilities were developed and tested. The objective is that the system estimates the median house values and also predicts if the passengers in Titanic were survived or dead. Different evaluation metrics were used to measure the performances.

2. Regression

The dataset is the housing dataset with 1000 data. ”median house value” is the dependent variable. Polynomial regression, random forest, and gradient boosted trees are selected as the regression models because the dataset is huge and these models are considered the best for this particular dataset. Gradient boosted trees (GBT) and random forest use ensemble learning method so they use decision trees. Since there are 1000 data, these models could be good to predict the target values with the help of trees. Polynomial regression uses a different approach from the others. The relationship between feature and target value shows an ”*n*-th” degree polynomial. Mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R2 score are used as metrics for regression. The most important metric considered is the R2 score since it represents the accuracy of the prediction.

2.1. Pre-processing

The housing dataset had 1000 rows and 11 columns. One variable was categorical so it also needed to be numerical. This was done using label encoding where each category is assigned a unique number. Another column was created for all encoded values. After that, the status of the columns was checked to find any null value. There were 11 null values in total, all of them was in feature variables. ”median house value” was selected as y. ”No.”, ”median house value”, and ”ocean proximity”, encoded, were dropped and the rest was combined to set x. After that, the data was split to train and test data for x and y. The test size was specified as 0.2, 20 percent, and the random state was specified as 0. Imputation method was used to fill the gaps in train and test data for x, so all the NaN values were replaced with the mean. Feature scaling was used to reduce the variation among the values in train and test data.

2.2. Methodology

Gradient boosted trees (GBT) was selected as a main model. It is a model that uses weak learners, decision trees, to make predictions. These weak learners are added to the ensemble in order to correct, or minimize, the errors of the existing model. The prediction for this model is made of the sum of the predictions from all the weak learners. This is the general equation:

$$F(x) = b_0 + f_1(x) + f_2(x) + ... + f_m(x). \tag{1}$$

where $F(x)$ is the final prediction, y, b_0 is the initial prediction, usually mean of the target value, $f_i(x)$ is the prediction of trees. It works until the number of estimations are reached. The learning rate, γ, could be used to reduce overfitting.

	MSE	RMSE	MAE	R2
Gradient Boosted Trees	3.245201e+09	56966.66	40710.22	0.72
Random Forest	3.572204e+09	59767.92	44134.88	0.69
Polynomial Regression	3.915675e+09	62575.35	45835.04	0.66

Figure 1. The comparison of the performances based on the metrics

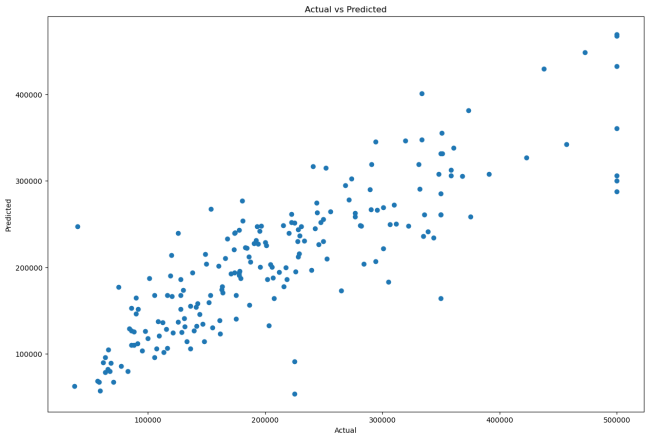

Figure 2. The plotted results of gradient boosted trees

It manipulates the contribution of each tree to the overall model. Each individual tree, is multiplied with γ and added to the final prediction, y. GBT aims to work out thoroughly by using decision trees for prediction to bring out the best result. Hence, it was thought that this model would bring a successful result.

2.3. Experiments

2.3.1. Experimental Settings

Random forest and polynomial regression are the baseline models. 2 different models were prepared for each one. First ones for each had no parameter while second ones had parameters for random forest and GBT. For GBT, number of estimators and learning rate were used. On the other hand, for the random forest model, the second model had number of estimators, max depth, minimum samples to split the internal node, and minimum samples for the leaf node. This also succeeded to improve the performance and reduced over fitting slightly. For polynomial regression, only the degree was used as the parameter. The first model had degree 5 and the second one had 2. The cost function decreased and the accuracy increased so over fitting reduced in all models, especially in polynomial.

2.3.2. Results

MSE and its root, MAE, and R2 score were calculated as evaluation metrics. MSE was used as a cost function, MAE was taken to see the absolute average of errors between the actual and predicted values in the models, and R2 was calculated for the accuracy of the prediction. The result of each model, after parameter tuning, was compared. GBT had the lowest MSE with 3 245 201 000 followed by random forest with 3 558 601 579 and polynomial regression with 3 915 674 532. The RMSE, of GBT was 56967 while that of random forest and polynomial were 59654 and 62575 respectively. The MAE of GBT, random forest, and polynomial were 40710, 43977, and 45835 respectively. GBT had the R2 score 0.72, random forest had 0.69, and polynomial had 0.65. The comparison of all the model performances is summarized in Figure 1. Additionally, the plotted graphs of all the models are demonstrated as Figures 2, 3, and 4.

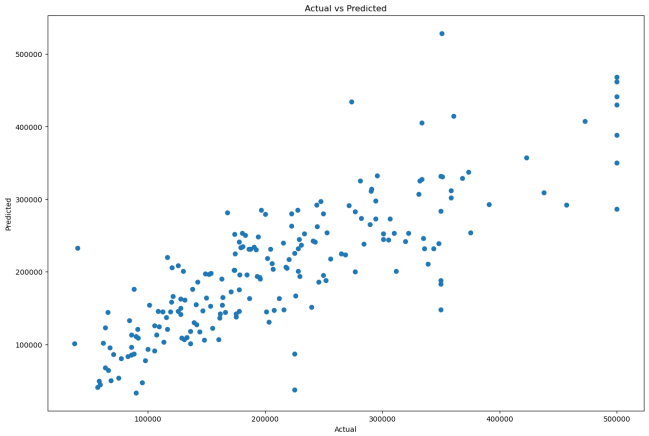

Figure 3. The plotted results of random forest

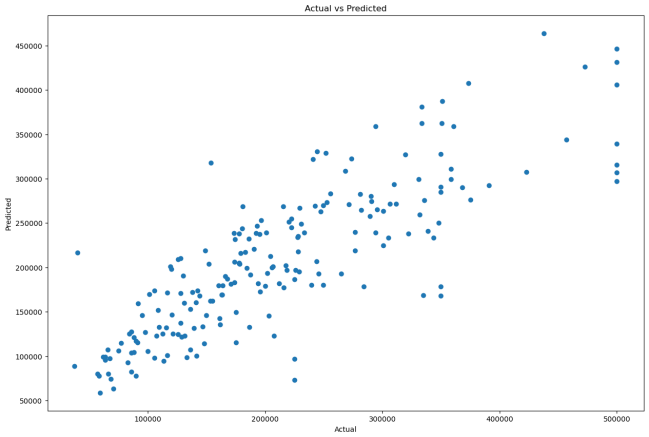

Figure 4. The plotted results of polynomial regression

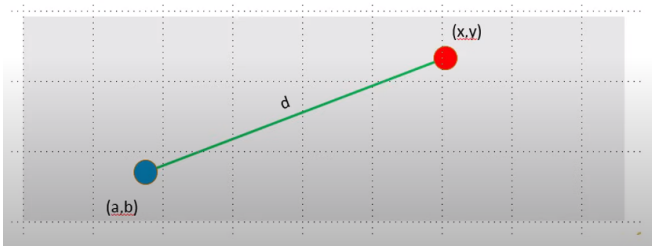

Figure 5. Visual indication of euclidean distance

2.3.3. Discussion

When MSE was calculated as a cost function, huge values came out. RMSE and MAE were used to normalize it. The RMSEs and MAEs of the models could indicate the quality of the predictions are good where the mean of the target variable is 207767. The R2 score is the easiest metric to interpret the model performance since it ranges from 0 to 1. R2 scores are greater than 0.65, closer to 1, so they all performed well. GBT has the best values. The plotted graphs confirm that gradient boosted trees has more actual and predicted plots closer compared to the baseline models. GBT and random forest use ensemble learning. They use decision trees while making predictions. They perform a classification technique even doing regression. The key point GBT, the main model, outperformed the others is that the trees used in GBT are dependent on each other, unlike random forest, so they correct one another leading to reducing the error of the model.

3. Classification

The dataset is the titanic dataset with 890 data in which ”Survival” column is the dependent variable. Logistic regression, decision trees, and K nearest neighbor (KNN) were selected as the classification models because these are the popular models and used widely. Logistic regression, regardless of its name, is a classification method and it uses the logistic function. Decision tree uses trees to classify the categorical values. It can be either used alone or by other models which makes it a part of ensemble learning. KNN picks the closest neighbors to predict values. Accuracy, precision and recall scores are used as evaluation metrics.

3.1. Pre-processing

The dataset had 890 rows and 11 columns. There were many categorical values in this dataset. They were converted to numericals using label encoding. ”Survival” was assigned as y while the rest, apart from ”PassengerId”, index, was all assigned as x. X had 178 NaN values in total but y had none. When splitting the data, test size was set to 0.27. 42 was used as random state since it might affect the result of the performances positively. Imputation method was used to deal with the missing values. Feature scaling was used for the same reason as regression.

3.2. Methodology

K nearest neighbour (KNN) was selected as a main model for classification. It is a classification method that makes predictions based on its neighbours. ”k” parameter represents the number of nearest neighbours to include. It affects the accuracy of the model so it is important. The square root of ”n”, the total number of data point, is taken for ”k” and selected if it is an odd value. Otherwise, the nearest odd value to the result should be taken. Euclidean distance is calculated the find the nearest neighbors. The distance between two points with coordinates (x, y) and (a, b), shown in Figure 5, is calculated by:

$$d = \sqrt{(x-a)^2 + (y-b)^2}. \tag{2}$$

This distance, ”d”, is used with ”k” to predict the class of a variable. All the closest values of ”d” to ”k” are picked to make estimation. A model that relying on its neighbors would produce better results, that’s why it was selected as main model.

	Accuracy	Precision	Recall
Decision Tree	0.83	0.82	0.75
KNN Neighbors	0.84	0.86	0.73
Logistic Regression	0.83	0.80	0.79

Figure 6. The comparison of model performances

3.3. Experiments

3.3.1. Experimental Settings

Decision tree and logistic regression are the baseline models of classification. Two models were generated for each one, just like in regression. Decision tree had no parameter in the first model but in the second one, complexity parameter and max depth of the tree were assigned as 0.01 and 4 respectively. The result of all the metrics increased. For logistic regression, regularization model, C, was specified as 0.01 and fit intercept was set as true. In the second model, fit intercept remained as true but C was increased to 1. Increase in C caused the accuracy and recall to rise while it lowered the value of precision. For KNN, the square root of y test values was calculated to specify K which was 15. One of the parameters was the number of nearest neighbors while the others were p, 2, and the metric, euclidean. For the second model, the number of neighbours was set as 19 and the others remained the same. This increased the accuracy and precision; however, the recall score declined.

3.3.2. Results

Accuracy, precision and recall scores are used as evaluation metrics. Precision calculates the performance of positive predictions while recall measures how the model performs at getting all positive instances. But the most important metric is the accuracy because it, as the name implies, calculates the overall accuracy of predictions. The result of all the models after hyper parameter tuning was compared. KNN had the highest accuracy with 0.84 whereas the baseline models had the same score of 0.83. KNN had still the same performance when it comes to precision with 0.86 followed by decision tree with 0.82 and logistic regression with 0.8. However, KNN had the worst recall score with 0.73 while logistic regression comes first with 0.79 and decision tree is the second with 0.75. Figure 9 shows the metric scores of all the models.

3.3.3. Discussion

Since these metrics are between 0 and 1, it would be much easier to interpret compared to regression metrics. All the scores are above 0.7, all the scores are close to 1 which shows that they all performed well on predictions. KNN was the main model and it came first in accuracy and precision, but worst in recall. It is really good at making positive predictions; however, not bad at getting positive instances. KNN relies on its neighbors to make predictions, as mentioned earlier. The reason why KNN got behind the baseline models in recall might be that its neighbors could have more negative instances. Despite that, it did better than the others in precision as well which acknowledges its accuracy of positive predictions. Besides, the most important feature was the accuracy and KNN, the main model, despite slightly, outperformed the others. Using only the closest values instead of the entire data must have facilitated the model's performance.

4. Conclusion

The models have been implemented to predict the values required for regression and classification method at the best level. The report has the description of the model performances and data processing, including splitting data and imputation method. By tuning the hyper parameters, the models have become more robust, and over fitting is avoided as much as possible. A number of models were experimented and those included in the report were tried repeatedly to get the best possible result. In conclusion, the report involves a wide overview of the models and their performances.