Machine Learning Application in Predicting Used Cars Price

DSCI 3415 Fundamentals of Machine Learning

Habeeba Hossam
Data Science
The American University in Cairo
Cairo, Egypt
habibahossam@aucegypt.edu

Youssif Abuzied
Computer Engineering
The American University in Cairo
Cairo, Egypt
youssif.abuzied@aucegypt.edu

Introduction

The last phase of the project was preprocessing and cleaning the data. We ended up with 200000 samples and about 1699 features. The next step is to decide on a model to use for the project. This report includes the models we tried, experimental setup for each of them, parameter choice, performance evaluation , and pros and cons for the model and fit for the problem.

Experimental setup

To ease a comparative analysis between all the models introduced in this study, each model was tested on the dataset using 4 fold validation with the precision, recall and accuracy to each test set where appropriate. R-Squared (R^2) was used to assess the accuracy of regression models discussed below. All experiments were conducted using the sklearn library in Python.

To elaborate, R^2 is a goodness of fit measure meaning the value of 1 indicates a perfect fit, while zero indicates the fit is not better than predicting the mean of the y values.

Generally, precision is a performance metric that measures the proportion of true positive predictions among the total predicted positives. As for recall, it identifies all of the predicted positive points out of all actual positive samples in a dataset. Accuracy measures how well a model is able to correctly predict the class labels of a dataset. It is the fraction of correctly classified samples over the total number of samples in the dataset.

Discretizing the Label

Discretizing the label in machine learning is a technique for transforming numerical input or output variables to have discrete ordinal labels. For the problem at hand, label price in USD is quantitative. To preserve the main characteristics of the label it will be divided into 10 bins based on the 10th percentile. Many machine learning algorithms prefer or perform better when numerical with non-standard probability distributions are made discrete. This technique is used to make the objects recognizable and understandable for machine learning models. For instance,

Implemented Machine Learning Algorithm Insights

The researchers implemented the following machine learning models; K-Nearest Neighbors (KNN), Linear Regression, Decision Trees, Random Forests, Naive Bayes, Logistics Regression and Artificial Neural Network. A detailed description of the experimental analysis along performance metrics for each of the aforementioned models is provided.

k-nearest neighbors

Model Description: K-nearest neighbors is an instance based learning model. When a new instance is provided for classification, a set of previously classified instances with comparable traits that were stored from the training set are retrieved and utilized to determine the class of the new instance based on the most frequent class in the retrieved set. Because similar items, in this case used automobiles, typically have similar prices, the researchers believe that KNN is an appropriate model for the problem.

Experimental setup and parameter choice: It must be noted that for this model no binning or any other label discretization technique was used. As a rule of thumb, the value of parameter K should be smaller than the square root of the number of observations. As an initial start, the value of parameter k is 100. Also, the price of the car using KNN was determined by assigning weights to the nearest 100 cars, weighing the most similar cars more weight. We then used 4 fold cross validation to test this model. Concerning the similarity measure, Euclidean distance was employed.

Performance Evaluation: The calculated R^2 score is 0.81. This means that this model is relatively good for fitting this data. However, there are some drawbacks that make this model not very suitable for our data. First of all, the training time required by this model is very short. However, prediction is not time efficient as similarity should be calculated all over the dataset and then sorting the cars takes place which makes the process very slow. From a user point of view, this model will take a lot of time to predict a single car. As a result, the model is not time efficient at all.

Linear Regression

Description: Linear regression relies on correlation between the features and the labels to infer a predicted result. This algorithm learns by identifying the line, plane, or hyperplane which minimizes the error between itself and the actual observation. As evident there are more than one feature, thus, a multivariate linear regression was used. To minimize the error between the predicted values of the label and the actual label, the MSE method was used. Since the main issue at hand is a regression problem in the classical sense, Linear regression is a must tried algorithm. Moreover, it is credited for its time and computational efficiency.

Experimental setup and parameter choice: Since this model is a regression model, no binning or other label discretization method was utilized. This model is very straightforward. As a result, there were no parameters that we were able to set to a specific value. We used 4 fold cross validation for this model

Performance Evaluation: The yielded performance results indicate an improvement in R^2 , which is 0.88, in relation to the aforementioned score of KNN. This model was very fast in training and prediction time and it was really good for fitting the data. However, what makes us concerned when using this model is that there are not many improvement opportunities as this model does not have many parameters that we can edit to get better results.

Decision Tree

Description: Decision trees rely on inductive learning methods which draw general conclusions from a successive set of facts / feature properties. The algorithm is not only the easiest to understand as well as implement but generally successful and commonly used. Decision trees are used mainly for regression and classification. Moreover, decision trees excel in being able to handle both numeric and categorical data at the same time. Furthermore, it is efficient in terms of time as it is logarithmic in time.

Experimental setup and parameter choice: We set the criterion for deciding to split on which feature as the entropy. All the remaining parameters were set to the default of scikit learn. We also used 4 fold cross validation in this model.

Performance Evaluation: For the problem at hand, implementing decision trees resulted in the following value for R^2, 0.880. Decision trees yielded a very good value for fitting the data. Also, it did not take a long time for training and prediction. However, one major issue for decision trees is that it might generate hugely complicated tree structures which might end up in overfitting.

Random Forest Regression

Random forest is designed to overcome the high variance issue of decision trees. It is implemented through applying parallel decision trees. As a result, the output depends on multiple decision trees which results in low variance. In classification problems, the majority voting classifier is used to determine the final output. On the other hand, in aggregation problems, the final output is the mean of the outputs of all decision trees. Some of the pros of random forest is that it is efficient, especially in this case of large datasets. In addition, it solves the problem of datasets with high dimensionality while ensuring little overfitting related issues.

Experimental setup and parameter choice: One salient parameter that the researchers specified is the maximum depth of the tree which is chosen to be 10. Like other models, we used 4 fold cross validation in this model.

Performance Evaluation: The value of R^2 is 0.86. It is the best value among other models. However, it can be improved by increasing the maximum depth of the tree. The drawback is that it is time consuming and computationally expensive. In reality, this algorithm took the most time to run among all of the other models, taking an hour.

Artificial Neural Networks

Description: Artificial Neural Networks algorithm is a brain-inspired algorithm that is used to foresee problems and model complex patterns. This model consists of multilayer perceptrons. It mainly works by two main algorithms which are the feed forward algorithm for generating the output and the back propagation algorithm for training and updating the weights. This is mainly done by gradient descent and stochastic gradient descent. Artificial neural networks possess many advantages. First, it is very flexible and can work for classification and regression problems. Also, it is very fast for predictions as once it is trained, it can generate the output very fast. However, neural networks do have some drawbacks including very long training time. Also, neural networks are hugely affected by the nature of the training data.

Experimental setup and parameter choice: The main parameter indicated by the researchers is the maximum number of iterations which is 100 iterations. We had two hidden layers, each of them had 10 perceptrons. The activation function was the identity function. We used 4 fold cross validation in this model like other models.

Performance Evaluation: The yielded result is that R^2 is the highest, 0.90. The algorithm took a considerable time for training. However, the predicting time was really fast as it is simply a feed forward algorithm. Also, this algorithm is very promising as it has a lot of parameters like the number of hidden layers, their sizes, number of iterations , and the activation function. By setting and trying different values for the different parameters, the performance of this model could be pushed higher.

Logistic Regression

Description: Logistic Regression is a model that employs statistical inference and probability theory to estimate the uncertainty within a given dataset. In this model, learning is dictated through the use of gradient descent since minimizing the error results in transcendental equations with no closed form solution.

Experimental setup and parameter choice: Since this model is used for classification and our problem is a regression problem, Discretizing the label was necessary. As previously stated, the label price was divided to fit into 10 bins. The maxim number of iterations is 500. This yields the following performance metrics. Finally, we also used 4 fold cross validation in this model.

Performance Evaluation: The precision is 0.610, the recall and accuracy are equal to is 0.612. The accuracy score for this model was 0.61. These results are not satisfactory to us. This model does not fit our data. This could be explained by the fact that logistic regression is a classification model but our model is a regression. Binning the data does not seem to help in this model.

Naive Bayes Classifier

Description: It is a classification technique based on Bayes' Theorem with an independence assumption 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. The main advantages of this model is being very fast in training and prediction. Also, it is very useful in the case of having multi class classification and having categorical features. However, some of the disadvantages include the assumption that the data are independent. As a result, this model might not be very reliable if some input features are dependent on each other.

Experimental setup and parameter choice: For this model Discretizing the label was necessary as this model is a classification model and our problem is a regression problem. As previously stated, the label price was divided to fit into 10 bins. Also, we used a specific type of naive bayes which is the gaussian naive bayes as the normal naive bayes cannot work with numerical input features. Finally, we also used 4 fold cross validation in this model.

Performance Evaluation: The model yielded the following performance metrics. The precision is 0.4, the recall is 0.316 and accuracy is 0.315. Despite being the fastest model across all the models we tried, this model is clearly the worst.

Comparative Analysis of Top Performing Models

In this section of the report the main aim is to compare the top three models in terms of performance. It is salient to note that the aforementioned performance metrics, precision, recall and accuracy, of the Logistics Regression and Naive Bayes Classifier did not show any significantly good scores or figures. In reality the highest is 60 % which is low. For that reason, researchers will be mainly focused on other mentioned models. The models that are best suited for the problem at hand, predicting the prices of used cars, are the following; Artificial Neural Network, Linear Regression and Decision Trees. The highest performing model is Artificial Neural Network with the best R squared figures. The reason that the Artificial Neural Network might be performing better than the runner up, Linear Regression model, is that it is capable of capturing non-linear relationships between input features and the label. Furthermore, one of the main reasons Artificial

Neural Network might be performing better than Decision Trees could be its ability to identify complex patterns in data through its layered structure. Each layer learns to represent more abstract features based on the input data. In terms of time, neural networks had the greatest training time. However, it was capable of producing outputs really fast. From all of these models, the decision tree was the fastest in both training and prediction times.

Final Choice

The researchers choose Artificial Neural Network to be the main model for future model implementation in this project over Linear Regression and Decision Trees for several reasons. First, Artificial Neural Network has the highest R squared 0.90. Second, the ability to handle large complexity of non-linear data. Lastly, it is more robust when compared to Linear Regression and Decision Trees in terms of noise. Moreover, Artificial neural network was very attractive to us as it has a lot of parameters that we can edit and try different values for which might lead to improved efficiency. For more details, the researchers are aware of the time and computational complexity of implementing an Artificial Neural Network. To be clearer, it took almost an hour to complete the training process, while for Linear Regression and Decision Trees, it only took nearly ten minutes. To conclude, the researchers hold the firm belief that the Artificial Neural Networks model is the most suited for the data.