Thoucentric ML Internship
Assignment-02
Report

Pintu Kumar
IIT Bhilai
pintuk@iitbhilai.ac.in
7440934396

PREDICT USED CAR PRICES

Problem Statement :

Dataset :
https://www.kaggle.com/karimali/used-cars-data-pakistan/downloads/used-cars-data-pakistan.zip/2

Use the above data set to predict the used car prices

Expected Outcomes

1. *An excel files containing outputs for the test data set*
2. *All the code needs to be written using the python programming language*

The code should be shared as a GIT repository

Solution :

Improvement Timeline

When first time dataset it trained just after preprocessing , it gave R2 score of 10 % . After further preprocessing and doing explanatory analysis , it is found that there were too many missing values in almost all the columns except “Price”. It was about 2000 -4000 missing values in those columns.

Therefore I decided not to just remove those NaNs blindly. I replaced them with dummy variable “unknown” . When trained it again then accuracy score increased to about 44%.

After further analysis it was found that “Price” column in train dataset has outlier. So I removed the outlier. Again trained the model and found that the accuracy score increased to about 89% in case of RandomForestRegressor and about 80% in case of Linear Regression.

After doing important feature analysis and using only 30 important features, the accuracy score of Linear Regression gets reduced to around 72% -74%.

Finally after filling the missing values in column "KMs Driven" with the average value, the accuracy score is about 88% in case of RandomForestRegressor and 80% in Linear Regression.

10-Fold Cross Validation score is about 86% for RandomForestRegressor and 80% in case of Linear Regression. Therefore we can say that our model is not overfitting. It generalizes well and will do well on unseen data in future.
Now I did feature engineering.
I add a new feature "Damaged". If the "KMs Driven" is more than 2000000 then car is expensive otherwise it is cheap. When trained using this features along with others, the training R squrared score for RandomForestRegressor improved to 89% and 10-Fold cross validation score improved to 85% and test score is 85%.

For Linear Regression , train score is 82% , 10-Fold cross validation score is 80% and test score is also 80%.

Versions in git Repository :

Versions	Git Commit Hash	Result	Remarks
1.0	78646a4	R squared score = 22%	RandomForestRegressor gives score = 0.22
2.0	d3eecf5	R squared score = 42%	RandomForestRegressor gives score = 0.42 when missing values are replaced with dummy variable in categorical features
3.0	0bf3d09	R squared score = 86%	R squared score increased to 86% on test data when OUTLIER from the Price is removed and trained on model RandomForestRegressor
4.0	6edb317	R squared score = 82%	Linear Regression
5.0	ce5c5d7	In RandomForestRegressor R squared score = 85% In Linear Regression R squared score = 73%	Imputed the missing data in KMs Driven And R^2 accuracy score is 85% on test data for RandomForestRegressor model and 73% for Linear Regression when used only 30 important features after doing one hot encoding
6.0	35fa8ed		Final Model
7.0	2e14574	R squared score = 85%	Final Model for submission with feature engineering

Model Comparison:

	Linear Regression	RandomForestRegressor
Test set R^2 Score	80%	85%
Training R^2 Score	82%	89%
Train RMSE	212906.11	168002.12
10-Fold Cross Validation R^2 Score	80%	85%

Best Model : RandomForestRegressor
Training R-squared score : 89%
10- fold Cross Validation R-squared score : 85%