Submission by: Pavana Lakshmi Venugopal
UBIT ID: 50464513

Here ACB Auctions, wants to predict the market values(prices) of Toyota Corollas so that could help buyers or dealers. This could help them know which value they could bid with. This is a regression problem. I have selected KNN model, Linear Regression Model and Decision tree model. I have decided to drop 3 columns which would not be of mush use when it comes to prediction. The columns are Id, Mistlamps and Cylinders. ID would not add much value as its just numbers which would not help for the price of the cars. Mistlamps had more NA values, close to the actual shape of the dataset, so I decided to drop it. Column Cylinders had the same value throughout so I do not think it would add any value to the prediction task. All other columns I do think it would help in prediction and chose them as predictors. The Price column would be my response variable.

Since the dataset has 1436 rows and there were only few NA values, Price=1, Color=9, CC=5, Mfr_Guarantee= 1, Airco=1 I decided to drop them. The columns Fuel_Type and Color had categorical values, so I flagged them. After that I used z-score normalization to scale my data. For Knn I added leaf_size=30, could not change anything for Linear and for Decision Tree added splitter="best". For Knn in training I see the RMSE value as 1391.1371 and in testing it is 1804.4122. Definitely Overfit. For Linear I see the RMSE value as 1351.1285 and in testing it is 1322.8666. This is not Overfit. But performed better than Knn. Lastly, Decision tree I see the RMSE value as 1067.0795 and in testing it is 1173.5588. Not a huge difference. And I would say it is not Overfit as well. I would recommend ACB to choose Decision Tree as it performed way better than the other models. Less error as well compared to other models. I decided based on the RMSE values, and how well it performed compared to training.