Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Program: Bachelor of Science in Computer Science and Engineering

Course No: CSE 4108

Course Title: Artificial Intelligence Lab

Title: Used Car Price Prediction

Submitted to:

Mr. Faisal Muhammad Shah
Associate Professor, Department of CSE, AUST.

Mr. Md. Siam Ansary
Lecturer, Department of CSE, AUST.

Submitted by:

Group No: A101

Name: Mustofa Ahmed	**Student ID:** 18.01.04.005
Name: Samia Sabrina Nabi	**Student ID:** 18.01.04.006
Name: Sadia Mobasshira Hridi	**Student ID:** 18.01.04.014

1. Description of the problem:

Toyota and Honda are one of the most popular car brands in the USA. As a lot of people buy and sell these cars, there is a need for an application that will predict used car prices for the brand's Honda and Toyota. In our project, we first performed data visualization and found some insight that helped us to compare the two brands. Our dataset consisted of 6 label features and 5 numeric features. We had to preprocess these features before passing them to our models. We then used the models to predict the price of used Toyota and Honda cars.

2. A brief description of the dataset:

The data was collected from the CarMax.com website. On the website, people pictures of the car they want to sell. The dataset consists of more than 400 entries and a total of 11 features. Our dataset consists of 2 brands and from each brand, we choose 10 different models. One of the features is mileage (k miles), which implies how many distances the car had traveled. It also gives engine information of the car, the model the car was built, exterior color, etc. Finally, it consists of the price (in $dollars) of the car which is our target column.

3. Description of the used ML models:

Linear Regression: Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e it finds the linear relationship between the dependent and independent variable. A Linear Regression model’s main aim is to find the best fit linear line and the optimal values of intercept and coefficients such that the error is minimized.

Decision Tree Regression: A decision tree is a decision-making tool or a model of decisions that use a flowchart-like tree structure**.** Decision tree regression examines an object's characteristics and trains a model in the shape of a tree to forecast future data and create meaningful continuous output. The output/result is not discrete, in the sense that it is not represented solely by a discrete, known set of numbers or values.

Random Forest Regressor: An ensemble of decision trees is referred to as a random forest. A Random Forest is made up of many trees that have been built in a "random" manner**.** Each tree is made up of a distinct sample of rows, and each node is divided up into a separate set of features. Each tree generates its own unique prediction. The average of these forecasts is then used to give a single outcome. On huge datasets, it performs well as we are working with subsets of data. We can easily work with hundreds of features.

Bayesian Ridge Regression: Bayesian ridge estimates a probabilistic model of the regression issue. A spherical Gaussian provides the prior for the coefficient:

$$p(w|\lambda) = N(w|0,\lambda^{-1} Ip\ )$$

Prior over and are gamma distributions, which are the conjugate prior for Gaussian precision. Bayesian Ridge Regression is the name given to the resulting model, which is comparable to the classic Ridge. It performs well in cases of large multivariate data.

K Nearest Neighbor Regression: KNN regression is a non-parametric method that approximates the relationship between independent variables and continuous outcomes by averaging data in the same neighborhood in an understandable manner. The KNN algorithm predicts the values of any additional data points based on 'feature similarity.' The analyst must set the size of the neighborhood, or it can be decided using cross-validation to find the size that minimizes the mean-squared error.

Voting Regressor: A voting regressor is an ensemble meta-estimator that fits many base regressors, one on top of the other, on the entire dataset. The individual guesses are then averaged to give a final prediction.

4. Comparison of the performance scores of the models:

Name	**Mean Absolute Error**	**Mean Square Error**	**Root Mean Square Error**	**R2 Score**	**Cross-Validation Score**
Linear Regression	$2.9963^{\wedge 13}$	$7.8108^{\wedge 28}$	$2.7947^{\wedge 14}$	$-9.2618^{\wedge 20}$	$-9.3094^{\wedge -22}$
Random Forest	**876.9465**	**1637363.3326**	**1279.5950**	**0.9806**	0.7353
Decision tree	3551.7241	24379310.3448	4937.5409	0.7109	0.5703
Bayesian Ridge	1709.0295	4740035.8410	2177.1623	0.9438	**0.7979**
K Nearest Neighbor Regressor	2440.6130	11482758.6207	3388.6219	0.8639	0.6701
Voting Regressor	2451.7314	10865876.7667	3296.3429	0.8712	0.7568

5. Discussion

Amongst the model we have tested, the best performing model for our dataset is Random Forest. It outperformed all the other models in various evaluation metrics. Not only did it get good results, but the result also had a huge difference between the other models. This model has the highest accuracy as the prediction is made by averaging each of the predictions of each tree of the forest. Although the performance of the model was good, it performed poorly in cross-validation. Other models performed better. The model with the poorest performance is the linear regression model, which is as expected because most of the time, it is considered as the base model. The 2nd model that performed best was the Bayesian Ridge Model. It has the highest cross-validation score. The other models like K Nearest Neighbor Regression, Decision Tree showed somewhat similar results. Decision Tree could not perform well as Poor Resolution on Data With Complex Relationships Among the Variables. Example: the complex relationship between price and drive type. Bayesian ridge performs well in cases of large multivariate data. For our dataset which includes the majority of points from others, that’s why K Nearest Neighbor Regression could not perform that much good. Linear regression is appropriate for datasets where there is a linear relationship between the features and the output variable.

6. Contribution of Group Members

180104005: 33.33%

- Data Visualization
- Data Pre-processing
- Worked with two Machine Learning models- Bayesian Ridge and Voting Regressor.
- Wrote the introduction and brief description of the project.

180104006: 33.33%

- Collected 200 data for Honda.
- Worked with two Machine Learning models- Linear Regression and K Nearest Neighbor Regression.
- Collectively contributed to the discussion and the description of the models.

180104014: 33.33%

- Collected 200 data for Toyota.
- Worked with two Machine Learning models- Random Forest Regression and Decision Tree Regression.
- Wrote the documentation of the dataset.
- Collectively contributed to the discussion and the description of the models.