MARKET PRICE PREDICTION OF USED AUTOMOBILES USING VOTING ENSEMBLE METHOD IN ML

SANJAYA DE SILVA

A dissertation submitted in partial fulfilment of the requirement for

MSc Big Data Analytics

Department of Computing

Informatics Institute of Technology, Sri Lanka

in collaboration with

Robert Gordon University, UK

2021

Abstract

The automobile industry is rapidly changing, and car prices are only going up every day ("Used Car Market Size & Trends Report, 2020-2027," n.d.). With the current global economic situation, the majority of the people tend to buy used cars more than ever and also consider parameters like market value, life span, and reliability when buying them (Puneet, n.d.). Since insurance is also low when insuring used cars, people consider buying them more. Prescient & Strategic Intelligence states that global car market size for used cars is 115.2 million units in 2019, and will reach up to 275.3 million units by the year of 2030. This is a compound annual growth rate **(**CAGR) of 8.7% (Puneet, n.d.). But when deciding the market value of a used car can be more delicate. Since it depends on various factors like mileage, model, manufactured year, country, and how well it is being used, two cars might be having different market values even if they belong to the same car model. This can be problematic when deciding price, since one car model can be comparatively higher while another one being lower. This affects the buyers and they might end up paying more than what it is worth. To overcome this problem this research has focused on developing a Machine Learning (ML) based model to predict the market price of a used automobile.

System has three voting ensemble models and it will pick the best model depending on the dataset and predict the price with minimum error. Random forest, Gradient boosting, and MLP are the models and 0.215, 0.219, 0.222 are current root mean squared errors respectively for each model. Random forest regressor was selected by the voting ensemble method, since it has 0.951 test accuracy, and 0.215 root mean squared error for the BMW data that was used for testing. Having a voting ensemble method will eliminate the algorithm dependent misclassifications.

Keywords - Artificial Neural Network (ANN), Compound annual growth rate (CAGR), Random Forest regressor, Gradient Boosting regressor, scikit-learn, Machine Learning (ML)