ID5059

Knowledge Discovery and Data Mining

University of
St Andrews

Assignment

Summary of Machine learning Models for predicting car prices.

Student ID: 220024321

INTRODUCTION

The present study is aimed at leveraging machine learning models to forecast the market value of used vehicles. This prediction is based on a diverse set of predictors such as the automobile model, year of manufacture, mileage, and other pertinent data.

METHODS

An exploratory data analysis was conducted to identify any patterns, correlations, and distributions within the dataset. This approach facilitated an understanding of the data, which was instrumental in selecting an appropriate machine learning model to address the problem. Following the exploratory analysis, the data underwent preprocessing such as standardization, feature scaling, feature engineering and numerical encoding of the dataset variables to render it compatible with the machine learning models.

Regression models were chosen for this study, owing to their applicability in predicting quantity, such as the price of a car. Specifically linear regression, penalized regression models such as LASSO and Elastic Net were employed due to their penalty terms (L1 and L2 Norm), which optimizes the bias-variance trade-off. In addition, a random forest regressor was used, as it is less sensitive to outliers and has a better bias-variance trade-off by combining multiple decision trees to make predictions.

RESULTS

The unscaled dataset was initially used to fit the machine learning models, with each model performing at par except for the random forest regressor in terms of the R-squared score. However, using the scaled dataset slightly improved the models, except for the Elastic Net model which performed poorly. One reason for its poor performance could be attributed to its L2 norm penalty term, which shrinks insignificant covariates. However, scaling the data can lead to a loss of information, resulting in a suboptimal model. Tuning the Elastic Net made the model worse since an increased alpha value meant increased penalty term therefore forcing the L1 norm to shrink covariate to zero leading to a model with fewer features. The random forest regressor outperformed the other models with a total R-Squared score of 85%, achieved through hyperparameter tuning using cross-validation technique. One of the notable features of the ensemble regressor is its robustness to outliers and its ability to achieve a better bias-variance trade-off by combining multiple decision trees to make predictions.

The selected evaluation metrics for the best model (the Random Forest/Ensemble Regressor) were the **mean absolute error (MAE**) and **Root Mean Squared Logarithm Error (RMSLE**). The **MAE** was chosen because it measures the forecasting distance of the model to the true value and the model achieved a 3,580 MAE score. The **RMSLE** measures the ratio between the predicted and actual values of the target variable using the natural logarithm of those values. Because of the skewness of the variables, the RMSLE is ideal for measuring performance. The ensemble model achieved a RMSLE of 0.20, indicating a great performance. Finally, the Random Forest regressor performed great judging by its performance on the various evaluation metrics however further research could explore using statistical techniques such as Principal Component Regression, which transforms predictor variables into a smaller set of uncorrelated variables, imputation techniques for Null variables reducing bias or models which handles the impact of multicollinearity.