Applied Statistics & Machine Learning

Regression Analysis Report

Group B

Group Project ML - CA2

Problem Statement: To identify the car prices depending on the Make and Model of the car using regularised linear regression and regularised support vector regression techniques.

Team:

Mrinal Kokate (20025561)

Akash Jayakumar (20020040)

Neha Jagtap(20015927)

Submission Date:
27th April 2024

Under the Guidance of:
Dr. Assem Abdelhak

Vehicle Sales and Market Trends Dataset: This data set provides us with the valuable information regarding the Make and Model, Body Type, year, transmission, exterior and interior colours, seller information.

Potential uses of this data:

1. **Market Analysis**: We can identify what are the market trends and suggest to the dealers which types of models they can start selling depending on the demand.

2. **Predictive Modelling** : We can predict the sale price of a particular model depending on various important factors by identifying the features which regulates the prices. There would be some feature sets which we can drop depending on the feature reduction techniques.
3. **Business Insights**: Dealers, Financial Institutions relevant industry professionals can identify what vehicles are customer interested in, market demand and forecast and pricing fluctuations.

Data Preparation:

1. **Data Collection:** We identified the dataset for vehicle sales (from Kaggle Refence - https://www.kaggle.com/datasets/syedanwarafridi/vehicle-sales-data)
   There are 13 features (columns) and 3,906 rows within this data set. In order to begin with linear regression, we first followed the **CRISP-DM (Cross Industry standard process for Data Mining)** to understand the Business, then the data. We then started to perform data preparation wherein we used several steps. In order to read the data we selected PANDAS and imported relevant libraries to read the data further.

2. **Data Cleaning / Feature Selection**: We tried to check if there were any missing data which needed to be cleaned and handle those missing values. Also, in order to avoid any BIAS error(due to linear nature there could be features which are not representing the true relationship between the features and output) hence we need to remove those duplicate features.

3. **Data Transformation:**
   a. Standardisation of data using STANDARDSCALER is a common requirement which significantly improves the quality and trustworthiness of the data.
   b. We also use dimensionality reduction techniques like heat_map, manual reduction to use relevant features.
   c. Time-series data conversion - we used **to_datetime** to ensure we capture the date functionality correctly.

4. **Encoding Categorical Data: We used several method to encode the categorical data to numerical data using to_datetime, Mapping, target encoding.**

Impact of L1, L2, and elastic net regularisation:

The impact can be 2 fold. First on Coefficients and on Performance.

<u>Effects on Coefficients</u>:

1. **Without Regularization** :
   Best Parameters - (**eta)** η**: 0.001, max iter = 5000**
   Best Result (R2) - 0.0642 which is near to 0, meaning that the output feature dependent variable's (Price $) variability is not explained much by the independent variables. **This indicates a Poor Fit.**
   Intercept = β0 - 27939.308

2. **With L1 (Lasso)** regularisation , **it helps to reduce the coefficients to zero to reduce the features and making the model more interpretable and simple.**
   Best Parameters - (eta) η **: 0.001,** α **= 100, max iter = 5000**
   Best Result (R2) - 0.065415
   Intercept = β0 - 27881.443

3. **With L2 (Ridge) (square of coefficients)** regularisation, the model becomes more general and the coefficients are nearly zero but not completely removed, thus making the model more multicollinear.
   Best Parameters - (eta) η **: 0.01,** α **= 0.1 , max iter = 5000**
   Best Result (R2) - 0.0643
   Intercept = β0 - 28068.576

4. **ElasticNet:** Combines both L1 and L2 regularisation to strike a balance between the two.
   Best Parameters - α: 100, (eta) η : 0.01, 'l1_ratio' (λ): 1, 'max_iter': 5000
   Best Result (R2) - 0.0654154
   Intercept = β0 - 27881.4439

   We are getting the best result from Elastic Net as L1 (Lasso) Regression itself.

<u>Performance:</u>

In our model, we see that L1-ratio (λ) is 1 which is more prevalent making it a Lasso Regression. Lasso helps in reducing the feature making the model more simpler than making it more accurate. Hence it helps to perform well when there are less features impacting the output.

Whereas L2 (Ridge) performs well by improving the multicollinearity and making the model more generalised.

Hence there is a tradeoff between simplifying and model being more accurate. And hence the note is accurate - models built with fewer variables are considered more interpretable.

Impact of L2 regularisation on support vector regression performance and interpretability:

Without C -
Best parameters: {'epsilon': 1500, 'kernel': 'linear'}
Best result: -0.07765343204000122

Without C we get a negative R2 which implies that this is a poor fit model and does not explain the independent and dependent variables.

Interpretability: Since its a linear kernel we can consider that the model can be relatively interpretable.

With L2 Regularization (With C):
Best parameters: {'C': 5000, 'epsilon': 5000, 'kernel': 'rbf'}
Best result: 0.041934324761546106

Interpretability: With SVR we use the kernel “rbf”. The penalty is applied to large coefficients. This reduces overfitting for the model.

Analysis:
From the above outcomes, we see that the SVR kernel with “RBF” is performing better (Higher Result). L2 regularisation helps to mitigate overfitting, leading to a more generalizable model.

Implementation of Random Forest Regression:

Best parameters: {'n_estimators': 100}
best_score: 0.11682598310467898
modified_r2: 0.11370502635003732

Performance : In our dataset and performance of our model, random forest regression performed really well and produced high predictive accuracy compared to regularised linear regression model and regularised support vector regression.

To explain further, our RF regressor produced the best result of % compared to the best results we got from L1 and L2 Linear regression. I.e. _____ % L1

This % increase shows the best result and hence we went with RF Regression.

Interpretability:
In our model, the linear regression model offered straightforward interpretability and a disadvantage of coefficients’ associated with each feature directly indicating their impact on the target variable. However, in the case of support vector regression, models with non linear data like Radial Basis Functions (RBF) tend to create more challenges, hence to avoid and counteract all of that, we went with RF Regression. And since our data set is not a high

dimensional data set interpreting the individual effect of features has not been very challenging.

Performing a prediction for one of your models using new data:

x=[[0.52858, 2.109840, -0.387078, 0.945662, 0.985087, -1.270394, -1.550933, 0.900005, 1.272238, 1.793974, 1.392819]]

Output : $36395.13