Analysis of used Cars

Under the guidance of:

Prof.Patrick Shafto, PhD

Prof. Pushpi Paranamana, PhD

Course: Statistics and Machine Learning

Rutgers Business School, Rutgers University, Newark, NJ 07102, USA

|  |  |  |
| --- | --- | --- |
| **SECTION** | **TOPIC** | **PAGE** |
| 1 | Introduction | 2 |
| 2 | Cleaning, Exploration and Analysis | 3 |
| 3 | Methodology | 4 |
| 4 | Results | 4 |
| 5 | Conclusion | 5 |

**Table of Contents**

Section 1: Introduction

**Abstract:** The vehicle industry of the United States represents about 50 percent of the country’s auto retail market with an annual sale of USD 350 billion. This massive share constitutes the largest market segment of the economy. With the impact of digital disruption to the market, online retailers and ecommerce websites have become a convenient channel of trading for consumers aiding in quick buy and sell.The digital transformation has fueled the availability of pricing information and more transparency in vehicle conditions thereby, avoiding middle parties like car dealerships and traders. Admittedly, this availability of information is productive for all the stakeholders, a downside is that consumers can view different price listings of a single car model on various websites on the internet. This could confuse the customers posing a dilemma about purchasing a model from a certain website. Our motivation is to address the primary interest of a consumer regarding purchasing a used car. We have focused on attributes which may positively impact in consumers making an informed decision. Our parameters include the price point of a used car, whether it is appropriate in terms of obtaining a good deal for a car, vehicle history reports, brand of the car, its geographical location, mileage, transmission, size type, et cetera.

**Goal:** Our goal is to understand and analyze the factors that contribute to the pricing of a vehicle and to utilize these relevant features to build an efficient model that could effectively predict the pricing of a used vehicle.

**Audience:** Private buyers or traders looking online to purchase a used car.

**Dataset:** The dataset is made publicly available on Kaggle and is obtained from Craigslist, the world's largest collection of used vehicles for sale. This data is updated every few months and the set used consists of values till January 2020 and consists of 25 features and 0.5 million values.

**Features**: Below listed and features of the dataset that did not contain null values and are relevant to the analysis.

* **Region**: Geographical location of the car comprising 403 unique values of different locations.
* **Price**: Pricing of the car listed on the website
* **Year**: Full Date of registration of the car.
* **Manufacturer**: Brand name of the car. For example – Ford, GMC.
* **Model**: Model name or number of the car.
* **Odometer**: Provides the total distance the car has travelled in number of miles or kilometers
* **Fuel**: Type of fuel used by the car – gas, diesel or others.
* **Transmission**: Provides data regarding the transmission type of the car.  
  Our dataset comprises of 90% automatic cars, 7% manual cars, 3% other transmission types.
* **Size**: Size of the car – Full size, mid-size or compact

Section 2: Cleaning, Exploration and Analysis

**Data Visualization and Insights:** This section documents the questions we have tried to answer by data exploration and visualization. Following are some of the notable questions we have tried to address with insights gathered from answering them

* *Top 10 manufacturers:* Ford is the largest manufacturer of cars in the U.S. followed by Chevrolet
* *Top 10 years in car production:* Highest production of cars took place in 2013 by car manufacturers. Closely, followed by 2014
* *Condition of the vehicle:* More than 50% of the vehicles are in excellent condition. Of the remaining, majority are in good conditions
* *Fuel type versus Number of vehicles:* Vehicles running on gasoline takes up the lion’s share of the market
* *Price variation by transmission type:* Automatic transmission vehicles are always expensive than manual transmission vehicles
* *Price variation versus type of vehicles:* SUVs and Sedans are the most expensive type of vehicles. Moreover, Bus and Off-road are the cheapest type of vehicles
* *Price variation versus manufacturer:* Ford and Chevrolet are the most expensive car brands in the market.
* *Price distribution of vehicles in year:* 2019 has witnessed the prices of cars to be the most. However, this estimation is excluding outliers which were present in our dataset.
* *Car sales per state:* California has the highest number of cars sales followed by Texas
* *Number of each type of vehicle:* Sedan and SUVs are mostly in use

**Data Cleaning**: The feature columns that had NA values were removed. The data for the feature ‘condition’ was changed to the standard rating in KBB website. Before implementing any regression techniques, we used Label Encoder to encoded the mainly nominal categorical variables into numerical ones to ensure better performance. Log transformation was done to make the distribution of odometer data follow normal distribution. Listing of cars from 1990-2020 were taken for creating the model. From the price column zero values were removed and prices were limited at 150k.All zero-valued odometers were removed.

Section 3: Methodology

**Machine Learning Techniques:** Based on the goals of our project, we have implemented the following machine learning techniques for creating a predictive model

1. ***Multiple Linear Regression****:* An extension of Simple Linear regression, we have used this model to predict the value of a dependent variable, price, based on the value of two or more other independent variables in the dataset. Benefit of using this model is that it provides a good predictive capability. However, multiple linear regression was not effective for our dataset because the dataset is not in linear form.
2. ***Regularization (Ridge and Lasso)****:* This regression model constrains the coefficient estimate approximating to zero. That is, the technique does not allow learning a more flexible model to avoid overfitting. Therefore, results of regularization proved that there was no overfitting.
3. ***Random Forest Regression****:* The Random Forest is an additive model that makes predictions by combining decisions from a sequence of base models, with the base model being a decision tree. Unlike linear models used above, random forests can capture any non-linear interactions between the features and the target
4. ***XGBoost Regression****:* XGBoost is well known to provide better solutions than other machine learning algorithms. It is an optimized distributed gradient boosting library and faster performance.

Section 4: Results

|  |  |
| --- | --- |
| Machine Learning Techniques | Accuracy |
| Multiple Linear Regression | 0.5921322727066449 |
| Random Forest Regression | 0.857641848801272 |
| Random Forest after feature selection | 0.8586867344974106 |
| XGBoost Regression | 0.7924107122756809 |
| Ridge Regression | 0.5919121723274952 |
| Lasso Regression | 0.5921323055765577 |

**Accuracy scores:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Metric** | **Multiple Linear**  **Regression** | **Random Forest**  **Regression** | **Ridge Regression** | **Lasso Regression** |
| MAE | 4213.472237178738 | 1996.2395025225414 | 4209.2794349548785 | 4213.472447284642 |
| MSE | 38064256.33218949 | 13363624.644346008 | 38084797.20044197 | 38064253.26460482 |
| RMSE | 6169.623678328322 | 3655.629172159836 | 6171.288131374355 | 6169.623429724445 |

Section 5: Conclusion

From the above accuracy scores and comparison of errors, we can safely presume that random forest regression is the best performing model for the dataset compared to other models. The reason for the underperformance of multiple linear regression is that the data is not linear. Since random forest captures non-linear interactions between the features and the target, it works well with our dataset. Random Forest combines multiple decision trees at training time with subsets of data eliminating any interactions between the trees and adds randomness that could reduce overfitting.

**Features and Advantages of Random Forest**:

* It is one of the most accurate learning algorithms available. For many data sets, it produces a highly accurate classifier and runs efficiently on large databases.
* It can handle thousands of input variables without variable deletion..
* It generates an internal unbiased estimate of the generalization error as the forest building progresses.
* It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data is missing.

Irrelevant or partially relevant features can negatively impact model performance. Feature selection performed on the data based on p-values(level of significance = 0.01) in the hope of improving predictive accuracy. The model deemed ‘region’ as an irrelevant feature after filtering through all the features. By implementing random forest regressor with the selected features resulted in a slight increase in the overall accuracy (0.11pc). Performed 10-fold cross-validation to make sure there is no overfitting of the data. From the cross-validation results, we concluded that the model generalizes well with predictive accuracy.

In alignment with our goal to predict the sales price of used cars with a good predictive accuracy, we can say that random Forest model is successful in predicting the prices of cars with an accuracy of 85.86%.

**References:**

* <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>
* <https://towardsdatascience.com/regularization>[-in-machine-learning-76441ddcf99a](https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a)
* <https://towardsdatascience.com/random-forest-and-its-implementation-71824ced454f>
* <https://medium.com/@pushkarmandot/how>[-exactly-xgboost-works-a320d9b8aeef](https://medium.com/@pushkarmandot/how-exactly-xgboost-works-a320d9b8aeef)
* <https://en.wikipedia.org/wiki/Cross-validation_(statistics)>
* <https://www.datacamp.com/community/tutorials/xgboost-in-python>
* <https://xgboost.readthedocs.io/en/latest/parameter.html>