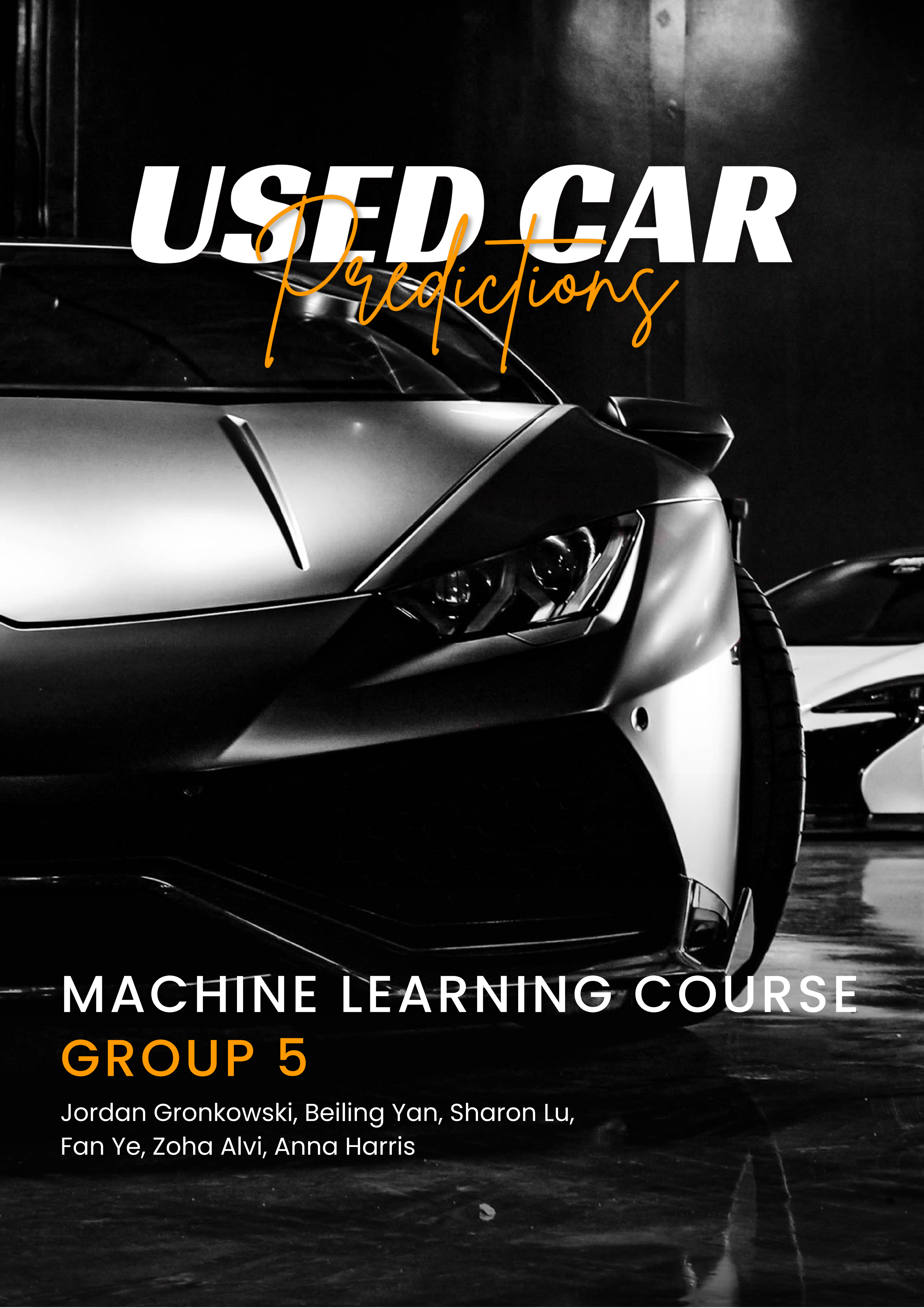

USED CAR Predictions

MACHINE LEARNING COURSE

GROUP 5

Jordan Gronkowski, Beiling Yan, Sharon Lu,
Fan Ye, Zoha Alvi, Anna Harris

Table of Contents

INTRODUCTION, HYPOTHESIS AND GOAL

Problem

A common perception about automotive vehicles is that they depreciate in value from the moment they leave the show room. The cost of a used vehicle is dependent on multiple factors, such as age, reputation, model etc. During the course of the COVID-19 pandemic, we even saw vehicles appreciate due to their supply and demand. In our analysis, we would like to evaluate which factors have an impact on the price, as well as to what extent. This would help establish if an asking price is reasonable for a vehicle, after evaluating its characteristics.

Hypothesis

Our hypothesis is that we can predict the selling price of a used vehicle, given the knowledge of conditions such as year, gas type, transmission type, location, etc. We will be utilizing and comparing these four models - linear regression, random forest regression, lasso regression and a support vector machine - and their predictive capacity for the price. In addition, we predict that in particular, the location factor of a car would considerably affect the resale price of the vehicle. In order to further test this, the modeling will be performed on three differently cleaned datasets and compared with each other to see how location affects the predictive capacity of auto vehicle pricing.

Goal

Our goal is to determine one tested model with the highest predictive capacity /accuracy of three tested models, based on various evaluation methods relevant to each model. The R2 value will be of the highest consideration when making a judgment on the effectiveness of a model.

EDA (DATA PREPARATION, CLEANING AND PIPELINE)

Data Preparation and Data Cleaning

The data was originally web-scraped from Craigslist postings and posted on Kaggle at Data Source. The data has 25 columns with a maximum of 539759 unique values.

On closer inspection, it became clear that the data had a lot of NaN values in it. For missing manufacturer's it was common to find the manufacturer listed with the model under the Model column, with the manufacturer full of null values. A few different scripts were run to pull the manufacturer name from a list of confirmed manufacturers first from the Model column, and then to check the description field to see if a recognizable manufacturer name was listed there.

The following columns were dropped during cleaning because they were either unnecessary or contained too many null values to be useful to analysis: id, county, vin, url, paint_color, image_url, region_url, size. Transmission, fuel, and cylinder columns had some null values imputed from the description column. Where no value could be found, those rows were dropped from the dataframe. Outliers that are far away from median values of "years", "odometer" and "prices" were dropped.

We were able to impute missing values for many rows from the description field. Scripts were run to locate missing odometer values, and to locate odometer values listed as "123k" and convert that to the correct integer. Corrections were made to misspellings of manufacturers, for example, Porsche was frequently misspelled as Porche.

At the end of the cleaning we had 250,573 records and 14 features (region, year, manufacturer, condition, cylinders, fuel, odometer, title_status, transmission, drive, type, state, longitude and latitude) in the cleaned file. The price of the vehicles ranges from $1,000 to $48,545 with a median of $11,999, the year of the vehicles ranges from 1985 to 2021, and the odometer ranges from 0 to 300,000 miles.

Figure 1 and 2 show the relationship between prices and year and odometer of vehicles, respectively. The price of vehicles decreases with increasing miles and years.

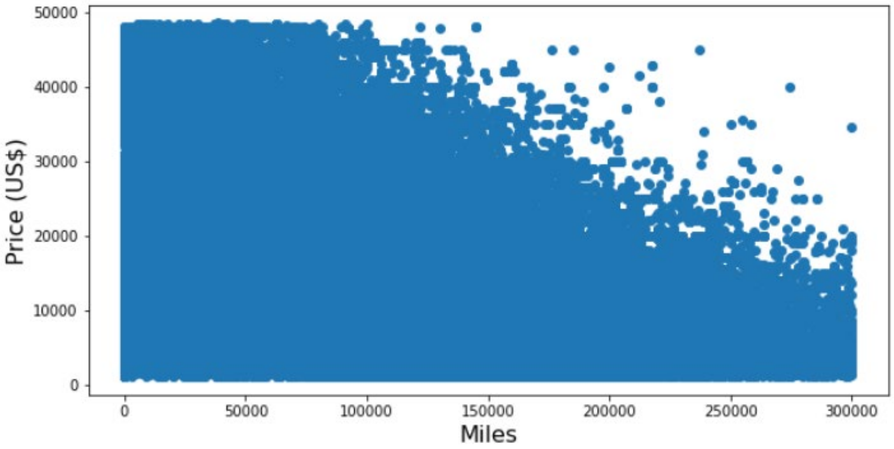

Figure 1. Dependence of price on miles of vehicles

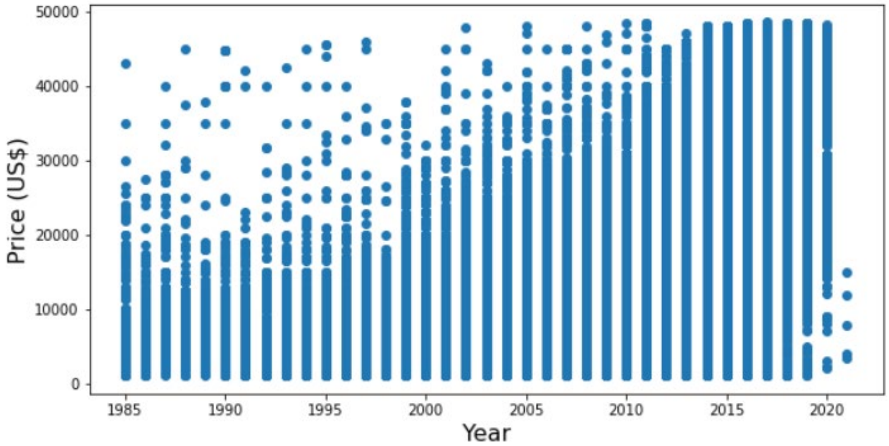

Figure 2. Dependence of price on year of vehicles

Figure 3 describes the price ranges of vehicles based on different conditions. There are six classes of conditions, including new, like new, excellent, good, far and salvage. It was found that the vehicles with the condition of new, have the highest average price, while the vehicles with the condition of fair have the lowest average price.

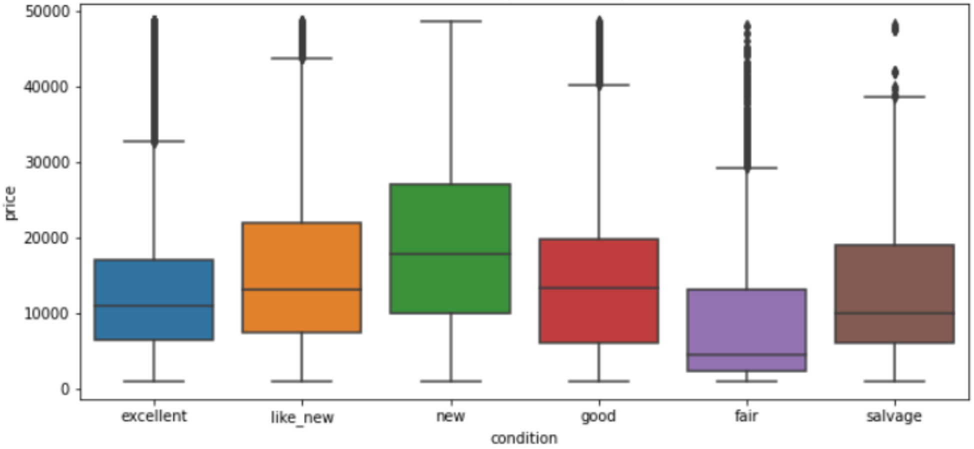

Figure 3. Dependence of price on conditions of vehicles.

Figure 4 describes the price ranges of vehicles based on various numbers of cylinders of engines. Electric vehicles which have no engines are classified as “no cylinders”. For most vehicles that have 4, 6, and 8 cylinders, the average price increases with the number of cylinders. However, the vehicles equipped with 5-cylinder engines have the lowest average price probably due to the unpopularity. The cars with 3-cylinder engines have almost the same average price as vehicles equipped with 8-cylinder engines. This is probably caused by hybrid vehicles which usually have engines of 3 cylinders.

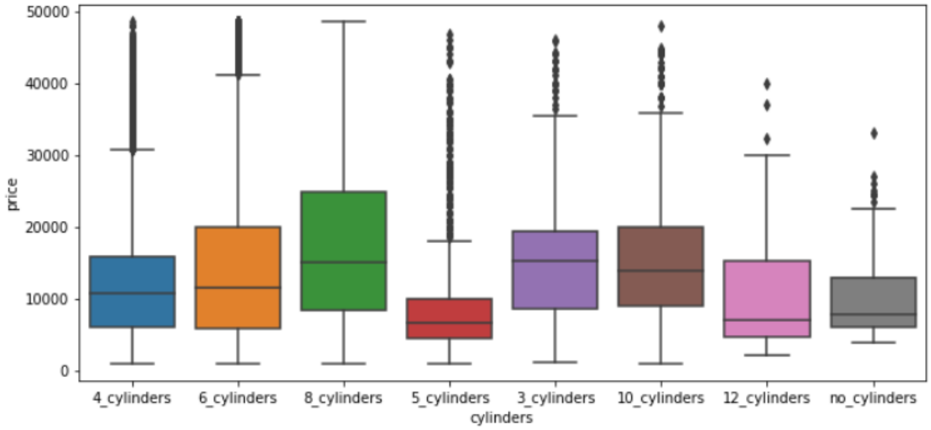

Figure 4. Dependence of price on number of cylinders of vehicles.

Figure 5 describes the price ranges of vehicles based on fuel types. It can be seen that electric vehicles have higher prices than vehicles using gas. It is also interesting to find that vehicles using diesel have the highest average price.

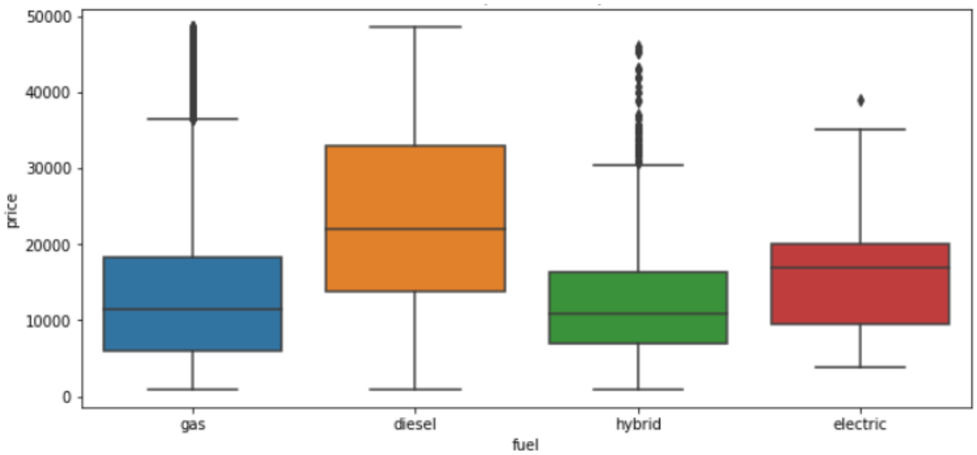

Figure 5. Dependence of price on types of fuels of vehicles.

Figure 6 shows the price ranges of vehicles based on their title statuses. We can see that vehicles with clean titles have higher average price than vehicles with salvage, rebuilt and parts only titles. However it is surprising to see the vehicles with titles of lien have the highest average price, probably because these vehicles are relatively new and just sold in recent years.

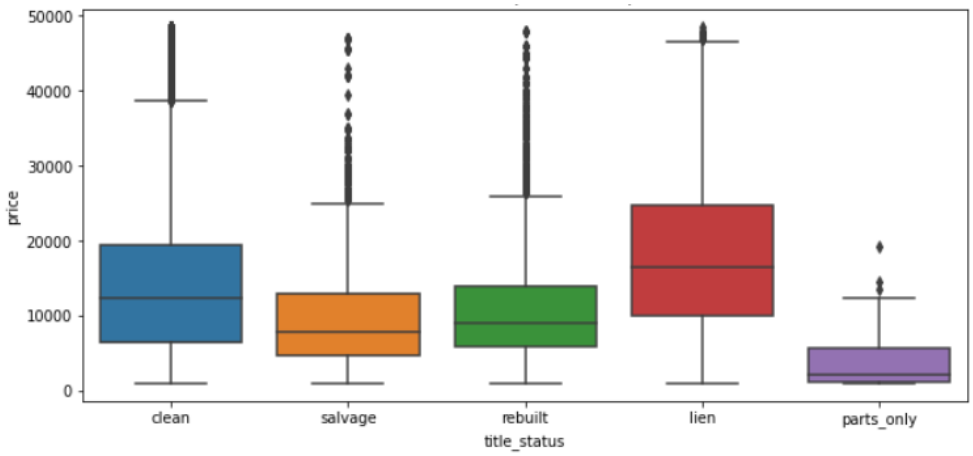

Figure 6. Dependence of price on title statuses of vehicles.

Figure 7 shows the price ranges of vehicles based on their transmission types. The vehicles with manual transmission have the lowest average price. The vehicles equipped with CVT transmission have the highest average price, probably because electric vehicles that have E-CVT are classified as CVT vehicles.

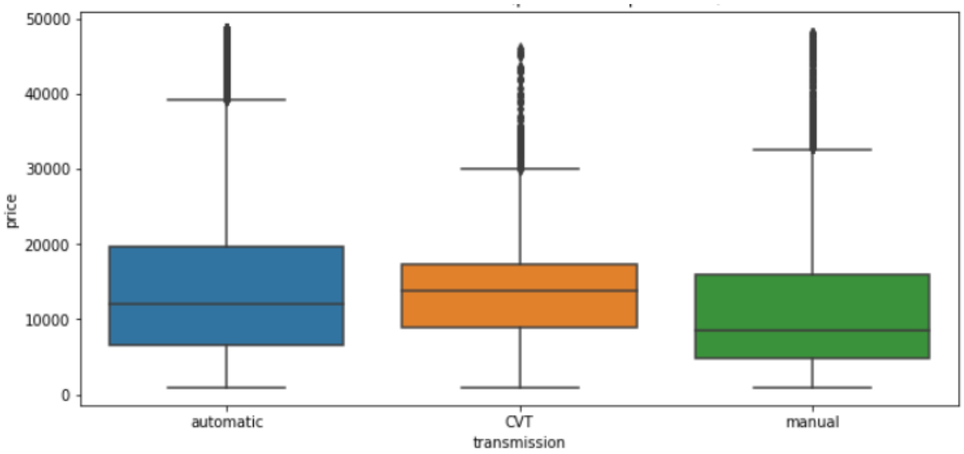

Figure 7. Dependence of price on transmission types of vehicles.

Figure 8 shows the price ranges of vehicles based on drive types. The vehicles equipped with AWD or 4WD are more expensive than other vehicles.

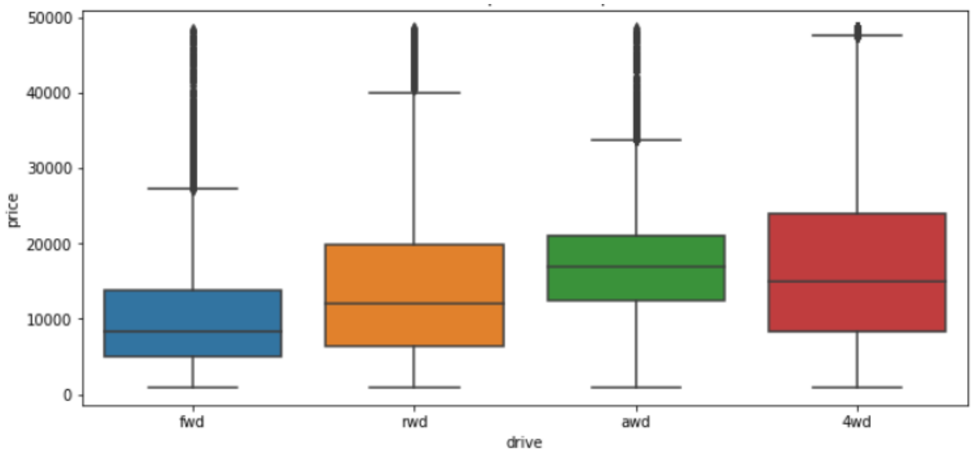

Figure 8. Dependence of price on drive types of vehicles.

Figure 9 shows the price ranges of vehicles based on types. The pickup and trucks have relatively higher prices, while sedan, hatchback and minivans have relatively lower prices.

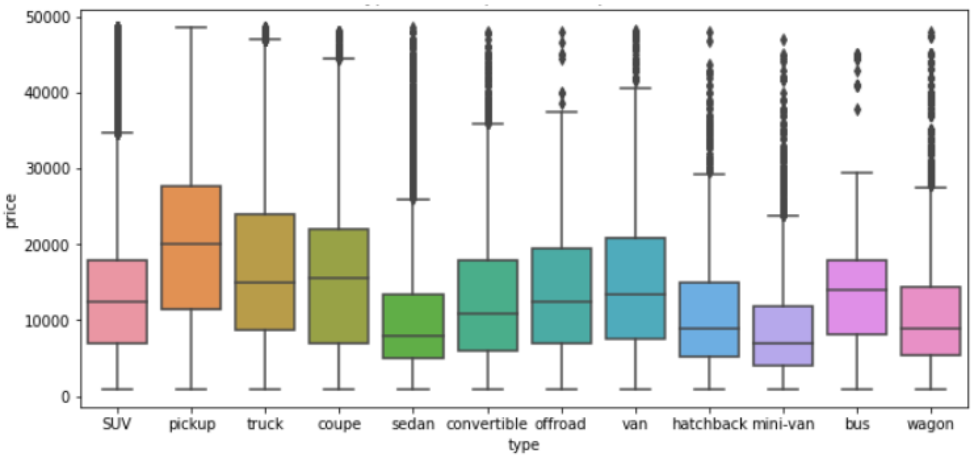

Figure 9. Dependence of price on types of vehicles.

Figure 10 shows the price ranges of vehicles based on the state where they were sold. We can see the vehicles sold in states that have fewer people such as Alaska, Hawaii, Wyoming and Montana usually are more expensive than vehicles sold in other states.

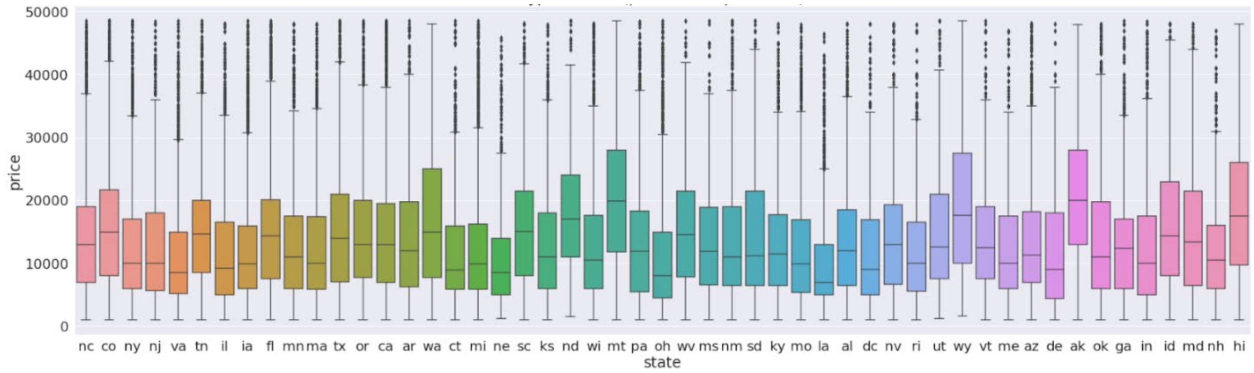

Figure 10. Dependence of price on state

Pipeline

The cleaned dataset has categorical variables of 'region', 'manufacturer', 'condition', 'cylinders', 'fuel', 'title_status', 'transmission', 'drive', 'type' and 'state', continuous numerical variables of 'odometer', 'lat' and 'long', and a discrete numerical variable of ‘Year’. One hot encoder was used to encode categorical variables and MinMaxScaler was used to scale numerical variables.

As longitude and latitude are slightly negatively correlated, as shown in Figure 11, we used Principal Component Analysis (PCA) to replace them with one new feature (‘lat_long_pca’). Similarly, odometer and year are also negatively correlated, so we used PCA to replace them with another new feature (‘odom_year_pca’).

In addition, we find the column ‘region’ has 403 classes and the location information can be represented by ‘state’, ‘longitude’ and ‘latitude’ as well. Therefore we decided to drop this column.

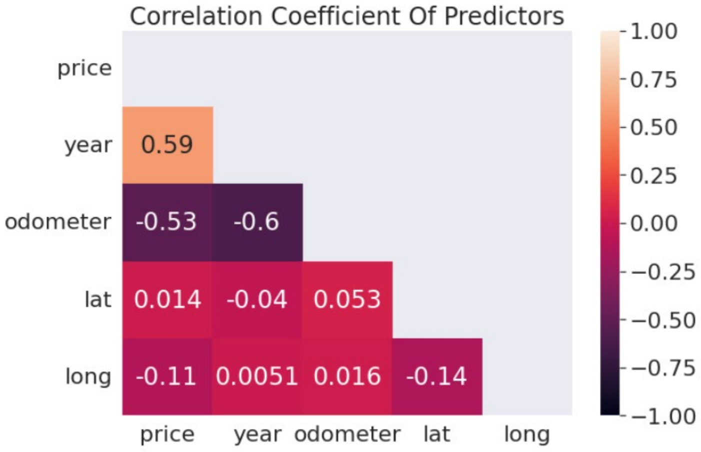

Figure 11. Correlation coefficient between numerical variables.

As a summary, we will train and test three datasets. The first one includes all the features of the cleaned dataset, the second one has no ‘region’, the third one excludes ‘region’, ‘year’, ‘odometer’, ‘latitude’ and ‘longitude’ but includes PCA created new features of ‘lat_long_pca’ and ‘odom_year_pca’.

MODELING BUILDING AND EVALUATION (LINEAR REGRESSION, LASSO REGRESSION, LINEAR SVR, EXTRATREEREGRESSION)

Baseline: Linear Regression

After a team member created the pipeline, it was used to run a linear regression and build the baseline model. First, import the sklearn model for splitting the train & test as test size =0.3 and then fit the X_train and y_train into a linear regression model.

We used r2_score to score the model, including region features receiving r2 extremely negative; excluding region features, 0.755; including PCA features, 0.752.

Apparently, exclusion of region features can help improve r2_score and the inclusion of PCA has little influence of r2_score in the linear regression model. More models were explored to explore better predictions.

Lasso Regression Model

Lasso regression is a machine learning algorithm that can be used to perform linear regression while also reducing the number of features used in the model.

We used r2_score to score the model, including region features receiving 0.755; excluding region features, 0.754; including PCA features, 0.751.

There are no significant differences between r2_scores no matter whether the region features or PCA features are included or not. And the r2_score of the lasso regression model is almost the same as the linear regression model.

Linear SVR Model

Linear SVR is similar to the SVM method. It also builds on kernel functions but is appropriate for unsupervised learning.

We used r2_score to score the model, including region features receiving 0.607; excluding region features, 0.605; including PCA features, 0.476.

It indicates that this model is worse than the Linear Regression model and Lasso Regression model.

ExtraTreeRegressor

ExtraTreesRegressor implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The default values for the parameters controlling the size of the trees (e.g. max_depth, min_samples_leaf, etc.) lead to fully grown and unpruned trees which can potentially be very large on some data sets. To reduce memory consumption, the complexity and size of the trees should be controlled by setting those parameter values.

In extremely randomized trees, randomness goes one step further in the way splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias.

We then used ExtraTreeRegressor to fit the dataset.

The score was improved.

We used r2_score to score the model, including region features receiving 0.923; excluding region features, 0.900; including PCA features, 0.920.

CONCLUSION

After the data cleaning and data preparation process, we were able to explore various relationships that impacted price. This allowed for insight as to what necessary changes will need to be made in order to make sure the models will evaluate correctly.

During the creation of the pipeline, One Hot Encoder was used for the categorical variables and the MinMaxScaler for numerical variables. Principal Component Analysis (PCA) was used to replace the longitude and latitude variables with a new feature (‘lat_long_pca’). Similarly, PCA was used to replace odometer and year with another new feature (‘odom_year_pca’).

The four models were then trained on three datasets:

1. Cleaned dataset including all features

2. Cleaned dataset excluding the region variable
3. Cleaned dataset excluding ‘region’, ‘year’, ‘odometer’, ‘latitude’ and ‘longitude’ but includes PCA created new features of ‘lat_long_pca’ and ‘odom_year_pca’.

The first model implemented was Linear Regression. This model was simulated in order to act as a baseline of comparison for all the other models. After training the data and evaluating the model, we evaluated metrics across the three datasets. The R2 score significantly improved for dataset 2 and 3, resulting in approximately around 0.75.

The second model implemented was the Lasso Regression, and this was done in order to utilize regularization and improve accuracy of the regression model. We found that the lasso regression model highly improved the R2 score of dataset 1 compared to the linear regression. However it did not significantly improve the R2 of dataset 2 & 3.

The third model implemented was the linear SVR model, mainly because it is useful for unsupervised learning. The R2 value for dataset 1 & 2 was better than for dataset 3. However, the R2 value was averaging close to 0.5, which was not enough to prove a strong regression.

The fourth model implemented was the ExtraTrees Regressor, which is a modified version of the random forest algorithm. The model yielded very high R2 scores for all three datasets, proving to be the best model out of the four for predicting the price of used vehicles.

In conclusion, the best model for predicting the price for all three types of prepared datasets was the ExtraTrees Regressor. The R2 for all three datasets was higher than 0.89, which demonstrated high predictive capacity. It is also important to note that the highest R2 value attained was with the dataset that contained all features. Although it did not improve significantly, it is still important to consider that location did affect pricing. This regressor differs from the random forest algorithm, as it utilizes randomization by splitting nodes randomly instead of ideally. It samples without replacement and nodes are based on random splits. This improves accuracy as well as overfitting, as variance is reduced at the expense of slightly higher bias. The regressor seems to perform better with higher levels of noise as well. In addition, it is not as computationally expensive as the SVR model. The SVR model allows us to decide how much predictive error is acceptable - however it doesn’t work for all data points. This was probably the reason that the model didn’t perform as well, as there were many features that needed to be evaluated for price prediction.

The ExtraTrees Regressor utilizes the entire dataset to train decision trees, and this reduces bias that is otherwise increased due to the randomization. In addition, it had low computation cost in comparison to the other models trained. Because the following model had substantial feature engineering, ExtraTrees Regressor verified itself as the leading model to be used.