Used cars market: Final report

Sabbella Prasanna

4/17/2021

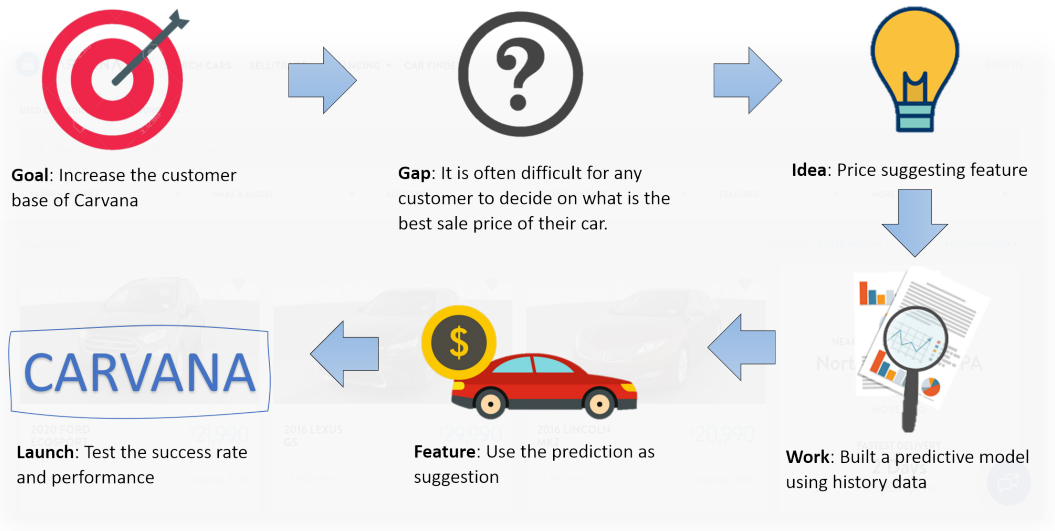
1. Introduction

Motivation

Since the past 10 years Russia always stood among top 10 car makers in the world. About 1.8 million cars were sold in 2019. Fuel cars are now being replaced with electric cars and the demand for used cars is never going to fall anytime in the future and there will always be a need of an online platform to sell and buy used cars.

Background

Carvana is an online platform that allows users/customers to sell and buy used cars and like numerous other brands we might want to expand the customer base of the website. It is often difficult for any customer to decide on what would be the best sale price of their car. Thus we came up with an idea to work on a new feature that will be able to suggest a sale price to a selling customer. This can be accomplished by building a predictive machine learning model using the previous sales history data and use this prediction as the suggestion price. Once executed the board members will need to assess the achievement rate and performance of the site.



Goals

The predictive model will be summarized as a report to the board members of Carvana who could further test/validate the price suggestion feature (probably by launching it as a trial feature for one quarter) before its launch. This sheds light on how impactful the feature is on the customer base of Carvana. During the trial period, the customer base growth/shrinkage is strictly monitored and in order to justify if the growth/shrinkage is really a consequence of the new app feature, feedback is collected from targeted customers who must have used the feature. If growth is the consequence, the suggestion feature can take its place on the website to be fully functional. Once convinced, they would invest higher volumes of budget into the project for gathering more data and make the app more reliable.

2. About the dataset

The dataset contains 36,533 observations and 29 different variables of different used car sales in 6 major cities of Russia in 2018. The data was collected by a freelancing data scientist in the year 2018 on personal interest.
Interesting Fact: There are 29 different car manufacturers in Russia among which 19 are currently active. Although the data collected is of the sales in Russia, Only LADA from the dataset belongs to the list of Russian car makers among 50 different car makers. This reflects the influence and proportion of market share owned by international car brands in Russia.

List of predictors in the data set

- **model_name** *(Categorical)*: Name of the vehicle model
- **manufacturer_name** *(Categorical)*: Brand of the car
- **transmission** *(Categorical)*: There are two types of transmission... automatic & manual
- **color** *(Categorical)*: Color of the outer body of the car
- **body_type** *(Categorical)*: Body type of the car
- **engine_type** *(Categorical)*: There are two types of engines... Gasoline and Diesel
- **engine_fuel** *(Categorical)*: Different types of fuels
- **drivetrain** *(Categorical)*: 3 types of drives are present..Front, Rear and All wheel drive
- **location_region** *(Categorical)*: There are 6 different locations where the cars are listed on sale.
- **has_warranty** *(Categorical)*: Shows whether a car on sale has a warranty
- **year_produced** *(Discrete Numeric)*: Year when the car got manufactured.
- **duration_listed** *(days) (Continuous Numeric)*: For how long was the deal on sale?
- **is_exchangeable** *(logic)*: States whether the car is exchangable with other cars.
- **engine_has_gas** *(logic)*: States whether the car's engine runs on gas
- **state** *(Categorical)*: Ownership status of the car. Owned/emergency/new
- **engine_capacity** *(Liters) (Continuous Numeric)*: Volume swept by all pistons in one engine
- **odometer_value** *(Kilometers) (Continuous Numeric)*: Total distance covered by the car.
- **number_of_photos** *(Discrete Numeric)*: Number of photos of the listed car uploaded on the website
- **feature_0 to 9** *(logical)*: Features such as Alloy wheels, fog lamps and other accessories in the car that adds on more price to it.

3. Obstacles

Difficulties

Handling 36,533 observations into random forests is quiet overwhelming.
Run time for Random forest is about 8 hours.

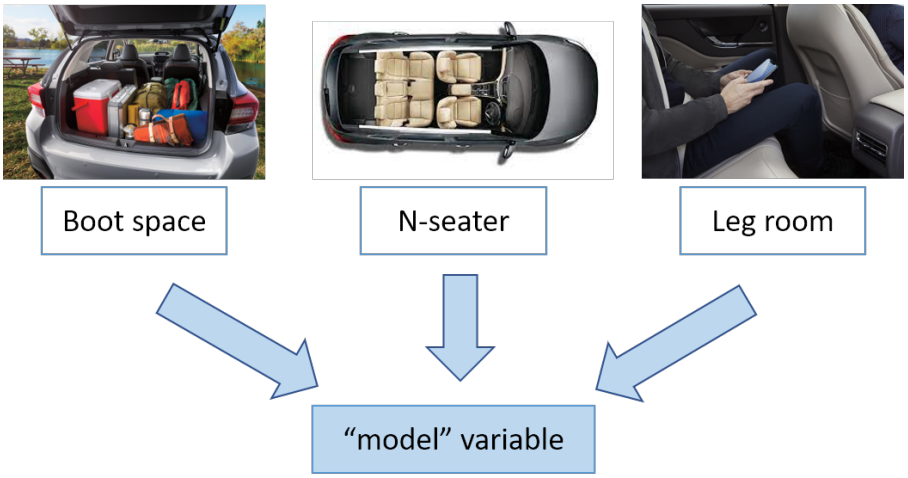
Dealing with missing information

Not all variables are defined
Variable name "up_counter" is not defined. Variable names "feature_0", "feature_1" ... "feature_9" (Alloy wheels, tubeless tyres, fog lamps) hold no correspondence with particular attribute.
The best way to deal with these variables is by not dealing with them as the analysis does not alter. Yet the app's suggestion feature may not work as the customer may not be able to input his car features to know the price.

If appearance is unknown the data might me deceptive
Two cars with same data may not portray the same aesthetics. Customer may choose these cars at different prices.

Customer may prefer this at $10,000 Customer may prefer this at $5,000

Missing information that matters Aspects of car that almost every customer looks at such as boot space, n-seater, leg room etc. are missing in the data set.
Assumption has been made that every model has a unique value/attribute of these missing details.



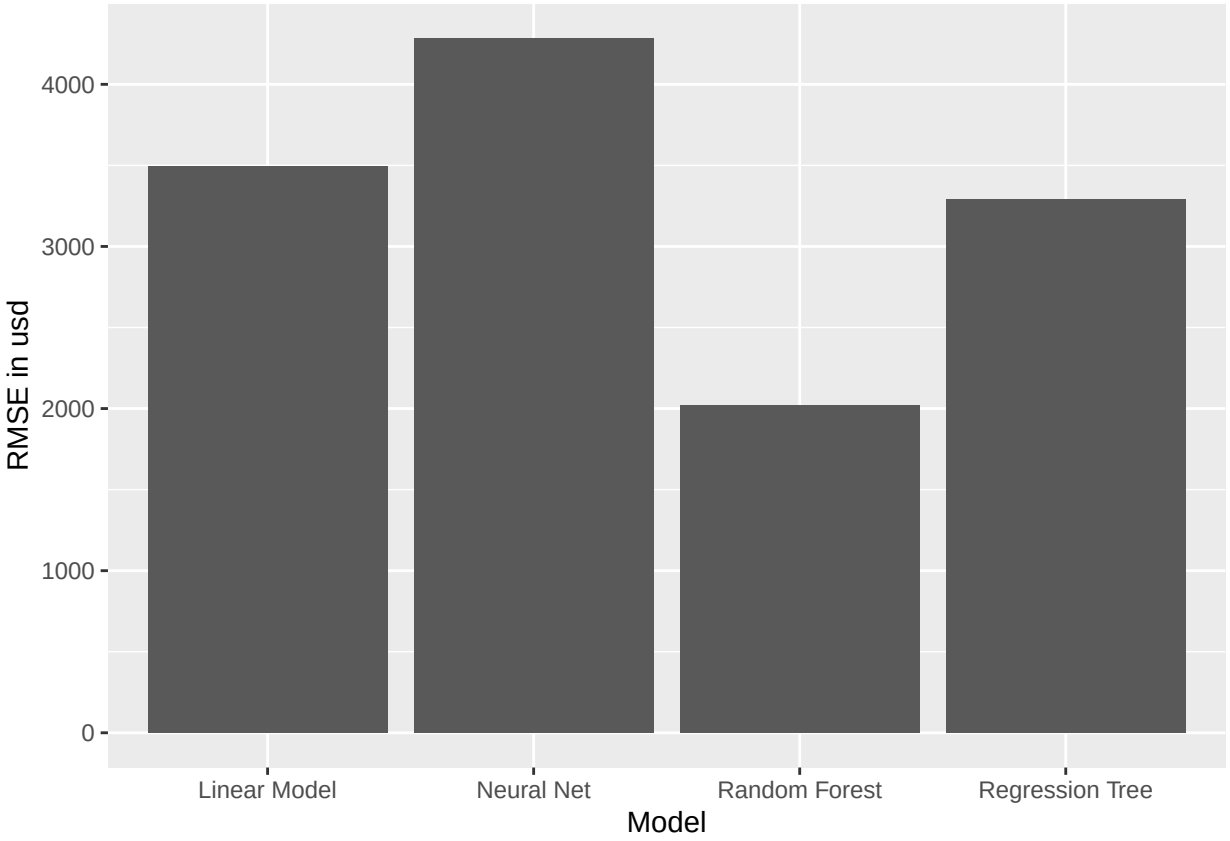
4. Analysis

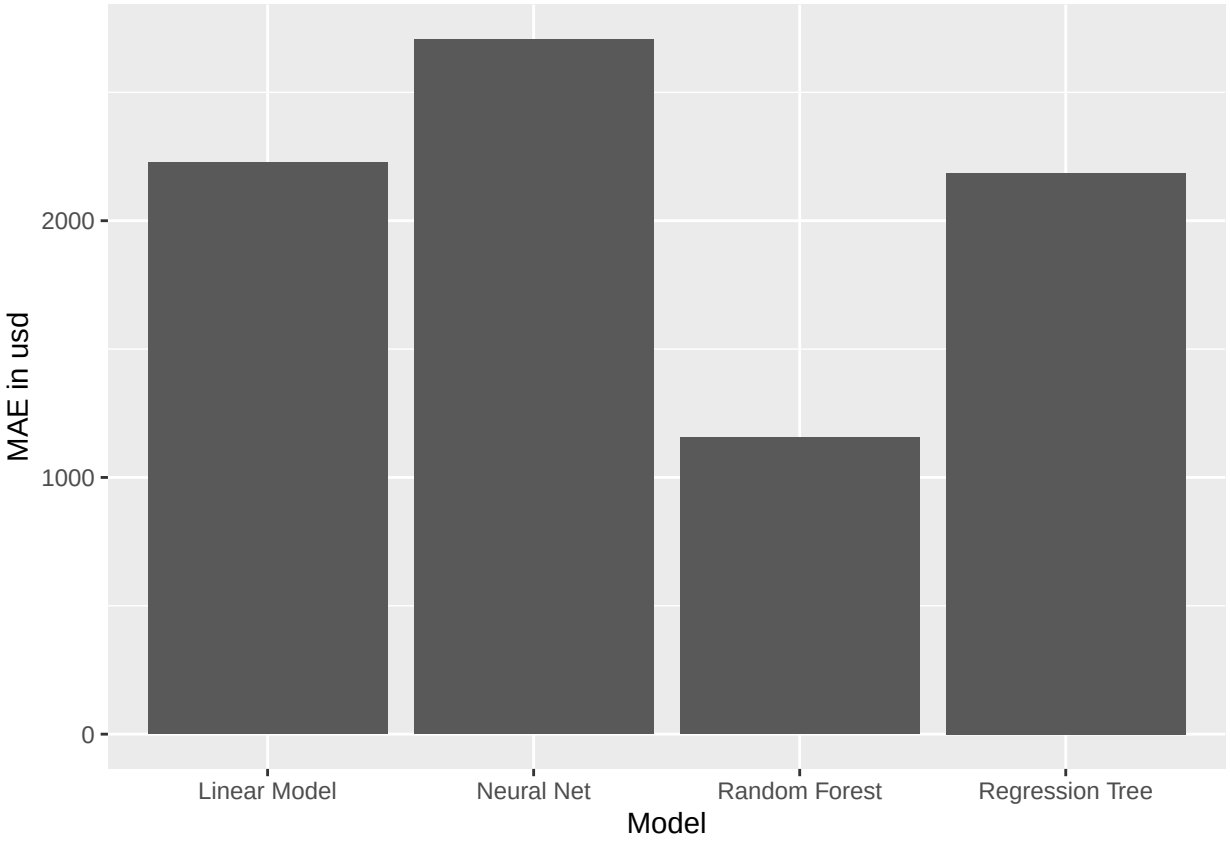
Preprocessing

Categorical variables in the data set were converted into numeric. No log transformation has been applied on the outcome variable and no predictors processing.

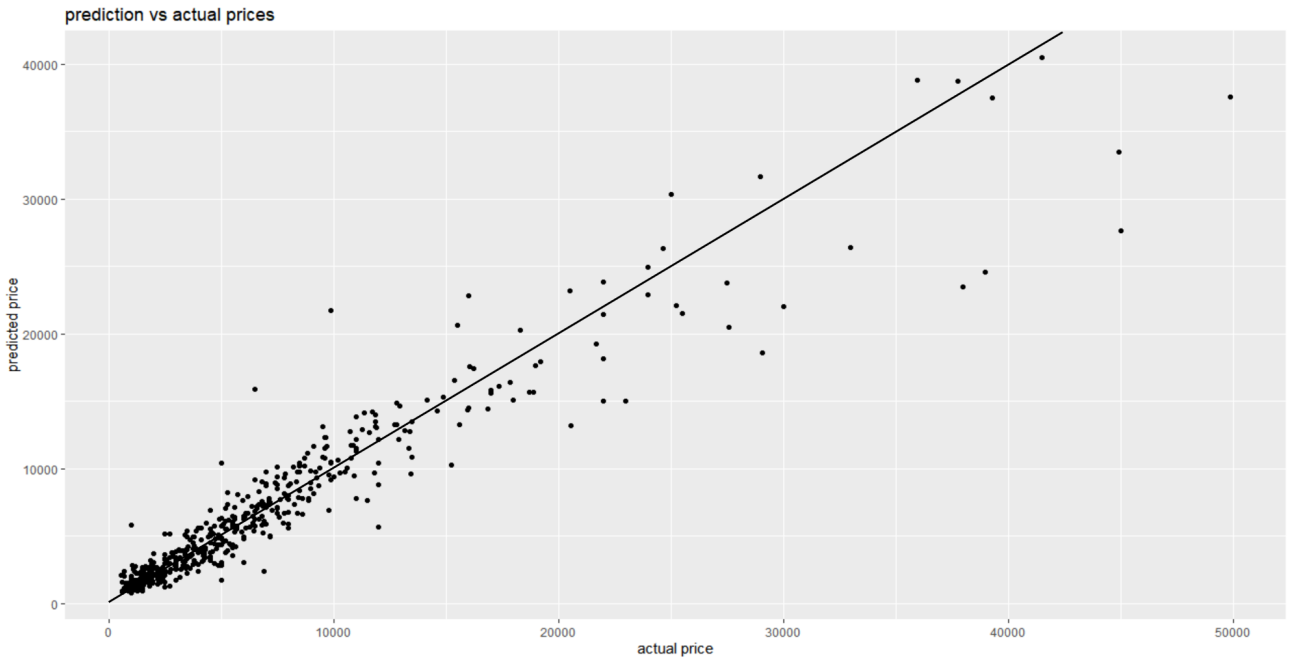
Performance of models

The RMSE and MAE of the prediction with different models is shown below. Both RMSE and MAE are the lowest with the Random Forest model.





The below graph shows the actual value versus the predicted value for random forest.

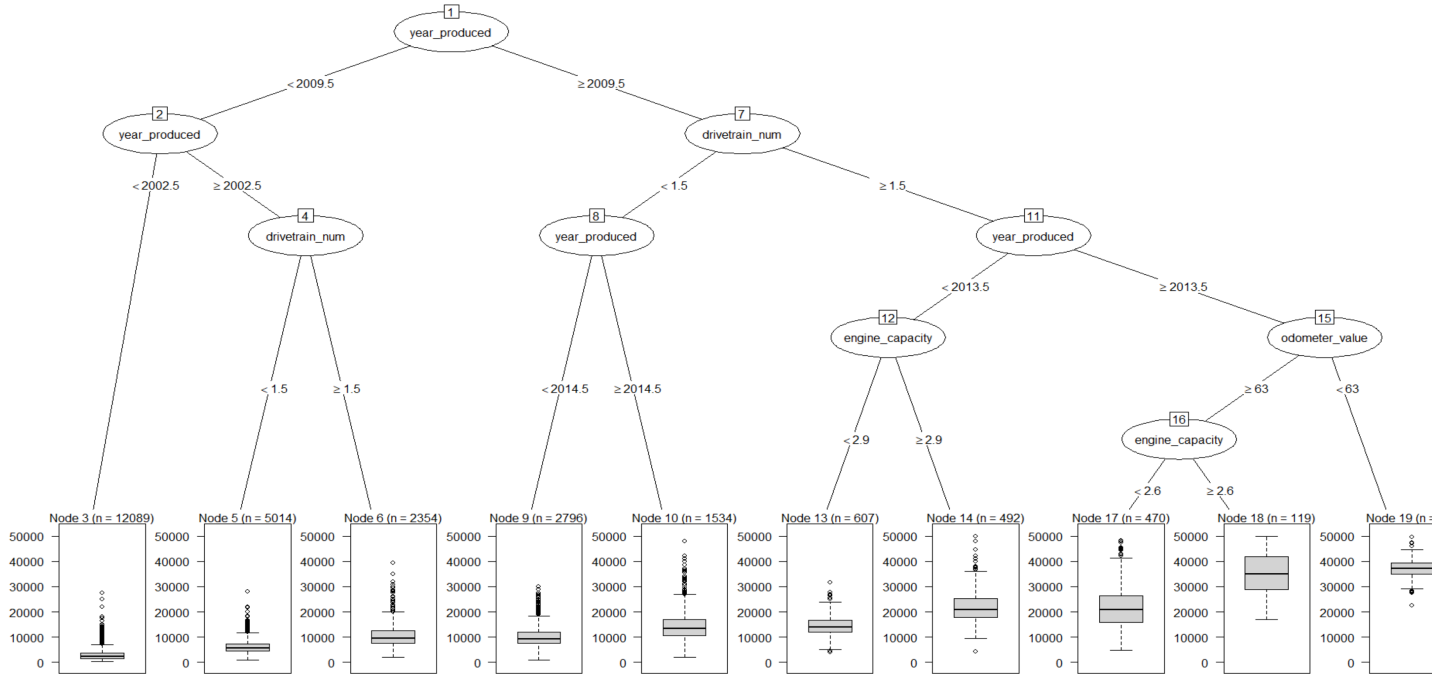


Why are trees better?

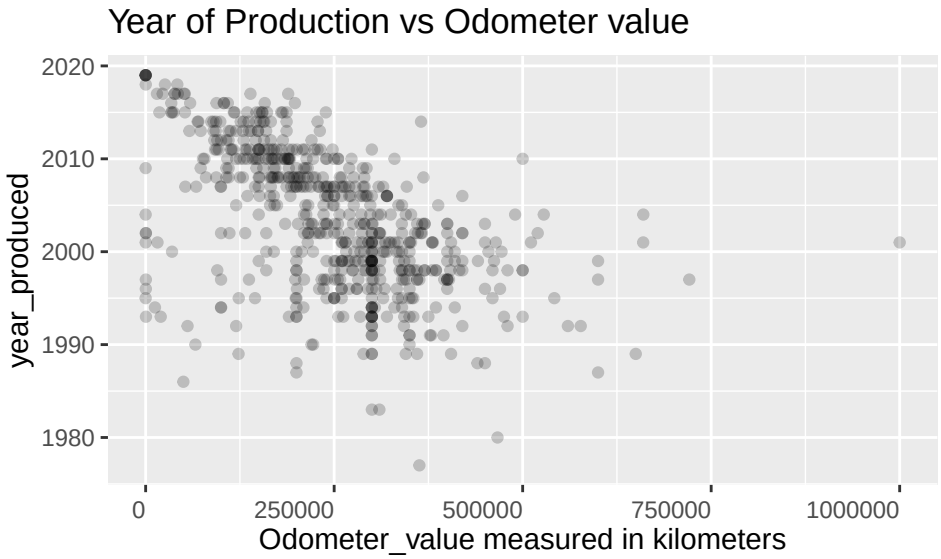
Each "manufacturer" has their own brand value. "*Manufacturer_name*" and each model has its own price range. "*Model_name*". Trees are basically designed to recognize these sets of models within the observations. So regression trees and random forest are better at predicting the sale prices "*price_usd*".

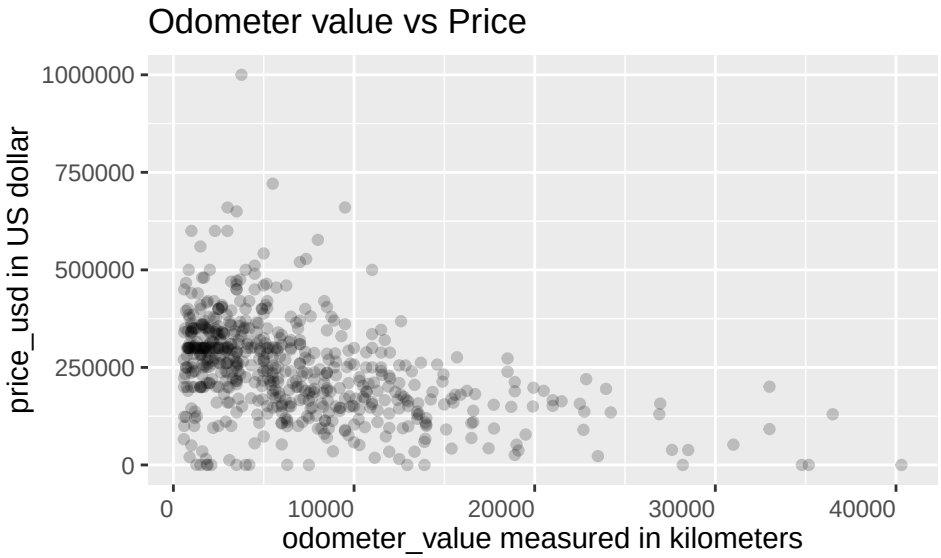
Discussing few aspects of the tree branches

The figure below shows a brief decision tree of the regression tree model.

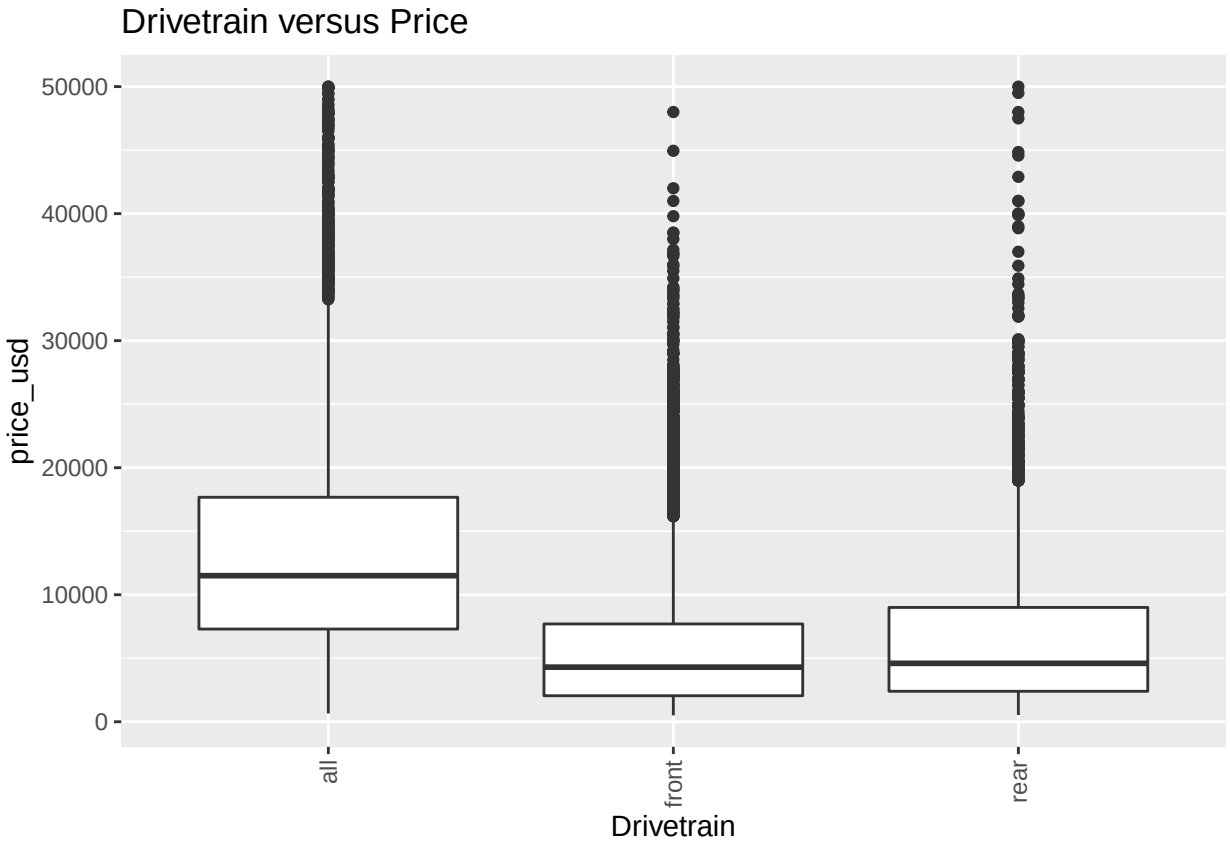


In the data set older cars are driven more so they have higher odometer values, which make a car less reliable, and the prices are lower. See graphs below. Year of Production governs most of the trend in the data while the other predictors are responsible for the variation. So, the decision tree mostly splits the outcome data based on the *year_produced* variable as observed in nodes 1,2,8 & 11.





In the dataset approximately 75% all wheel drive cars are higher in price than 75% of the number of front and rear drive cars. See graph below. Under converting the data set "All" wheel drive corresponds to 1, "Front" wheel drive corresponds to 2 and "Rear" to 3. So the *drivetrain_num* variable which is at nodes 4 and 7 is always split as >1.5 ("1"-All wheel drive) and <1.5 ("2 & 3"-Front and Rear wheel drive).



Dissatisfaction with the Random Forest

The below table explains what variable impacts RMSE of the prediction the most. *year_produced* variable being at the top of the tree is the most impacting variable with value at 100. The trees were hailed for their power of recognizing the fact that every model *model_name_num* has a price range, despite which the *model_name_num* variable's impact level is very low at only 5.43.

	Overall <dbl>
year_produced	100.000000
engine_capacity	21.227462
odometer_value	18.964211
drivetrain_num	18.306737
feature_7_num	15.904704
model_name_num	5.430518
transmission_num	5.146711
manufacturer_name_num	3.607591
feature_3_num	3.301451
body_type_num	3.284924
number_of_photos	2.614446
feature_8_num	2.569770
duration_listed	2.395395
up_counter	1.991034
feature_6_num	1.647638
state_num	1.452709
has_warranty_num	1.392522
feature_4_num	1.287678
engine_fuel_num	1.108852
color_num	1.094311

20 rows

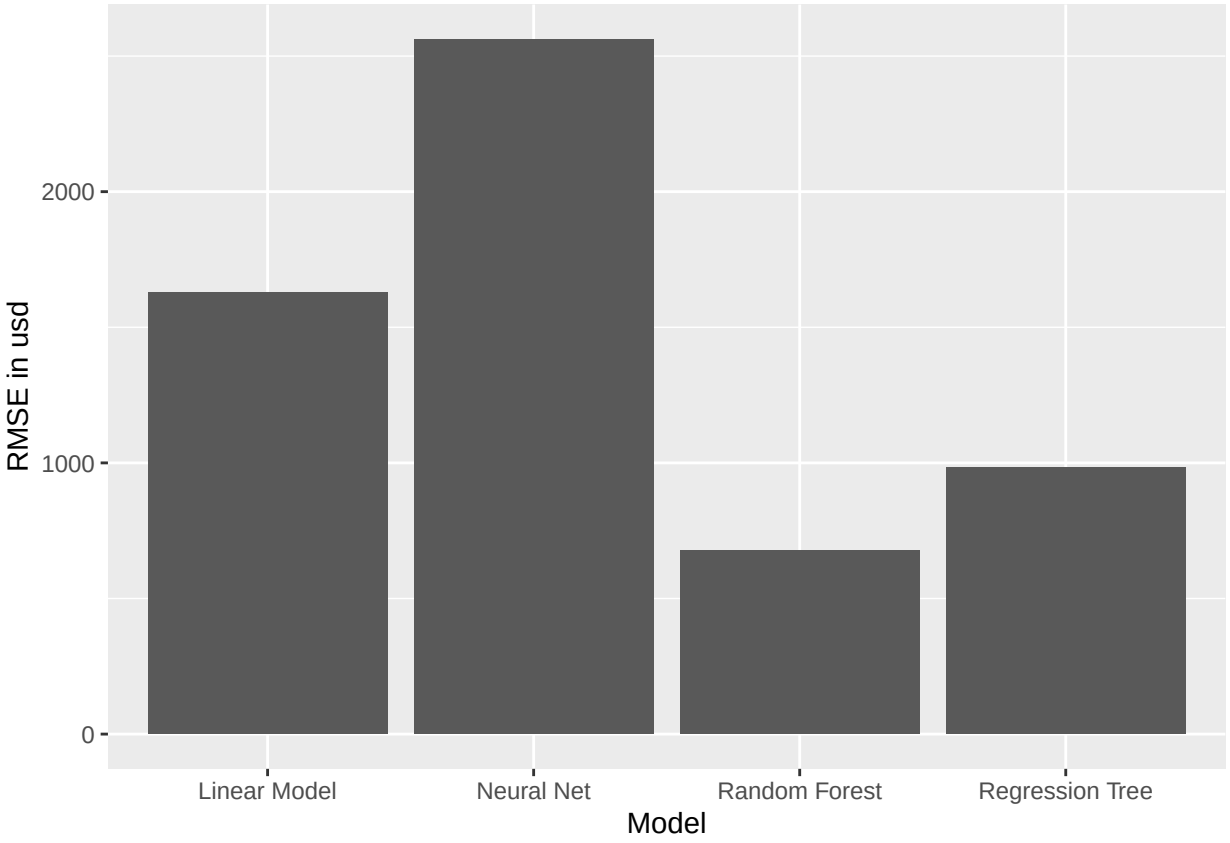
Prediction based on car model

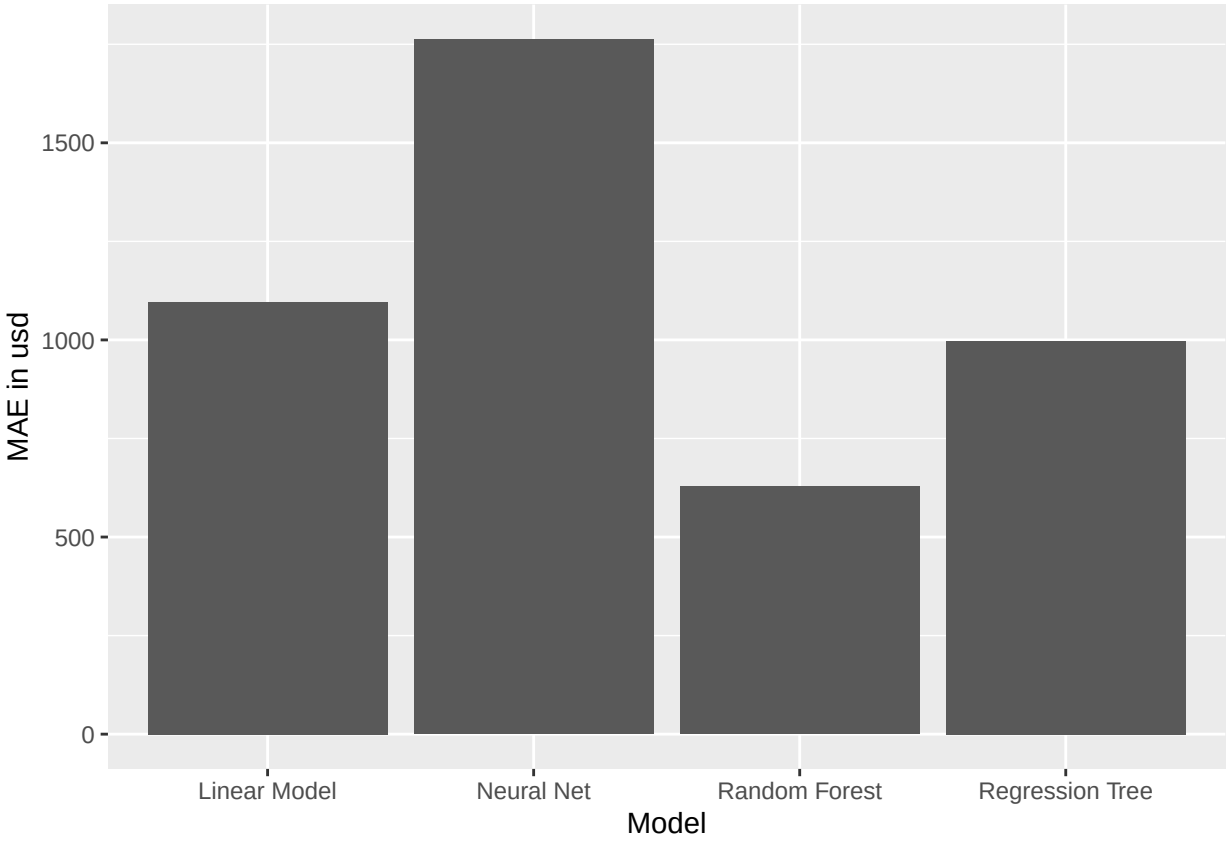
As the previous idea fails to produce a prediction based on the car model it is necessary to look at the RMSE of individual car models.

Each model has its own brand value and price range and brand value is based on many factors such as History, Country, Marketing strategies, people's obsession, attractiveness, captivating logo etc. of which no information is currently available in the dataset. Why try to predict the price of a car from the whole dataset when the customer would know exactly his car's model name?

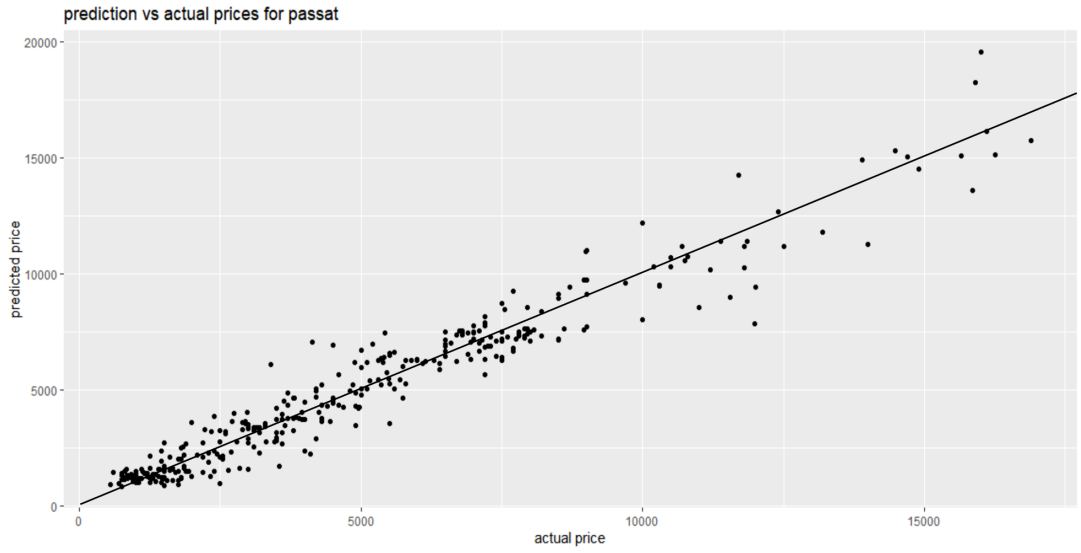
Performance of different models on the legendary "PASSAT"

As usually the Random Forests are performing better with RMSE at 866.7 usd and MAE at 622.5 usd



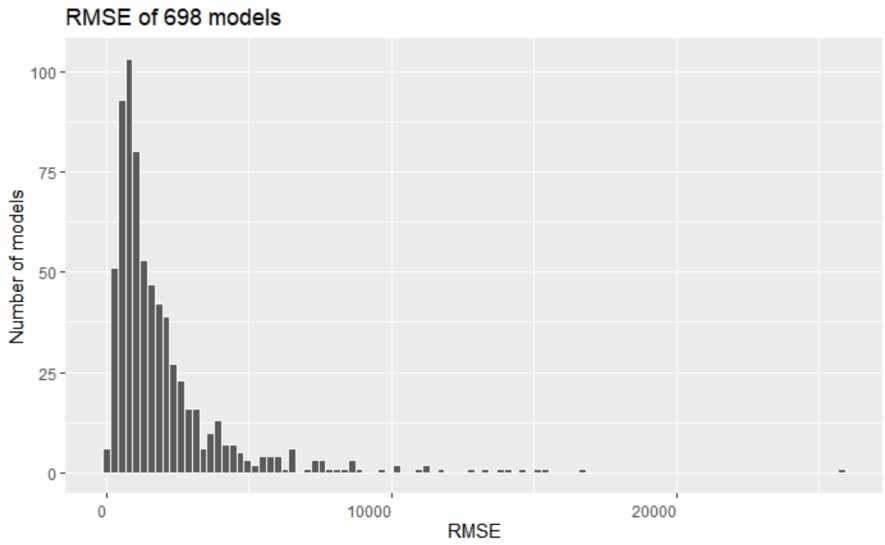


The graph below shows the actual prices versus the predicted prices for Passat



Performance of random forests on 698 individual car models.

Most models fall under the RMSE of 1000 usd.



5. Discussions

Interesting relations in the dataset

All the predictor variables in the dataset are more or less different features that are responsible for fluctuations in price and these features are improvements/upgrades that were developed over time in the automobile industry and this is probably a key point to why the fluctuations in price are time bound, implying that the "*year_produced*" variable governs the outcome variable.

Is the goal achieved

The goal is partially achieved. Prediction on 698 individual models is good but at the same time 324 models were eliminated as very few sales records were available. This eliminates less than 1% of the observations in the total data set which means that only 1% of the total customers are interested in these models. This also leads to several drawbacks for the price suggestion feature. The app's/website's suggestion feature works only if the customers model is within the dataset. Manual suggestion has to be done for the left over 324 models.

Future work

The fact that the initial Random Forest did not recognize well the price sets based on the car models has to be further investigated so that there is a scope of building a prediction model on 324 car models so that manual price suggestions can be eliminated. Wide range of ideas are still required to predict the price of a car which is not in the dataset at all.