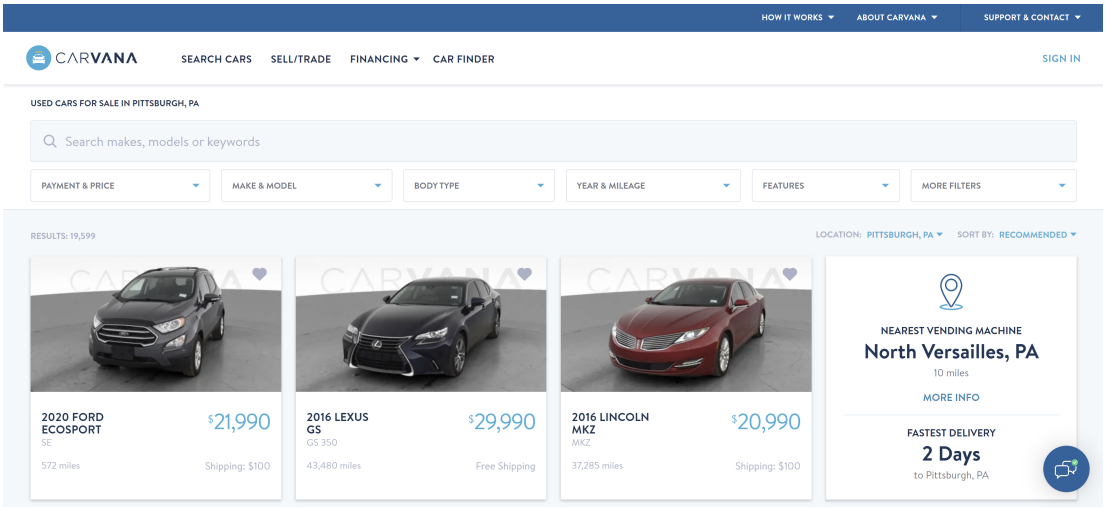
Used cars market: Preliminary Data Analysis

Sabbella Prasanna

4/9/2021

Scope

Context: Carvana is an online platform that allows users to sell/trade their cars. The marketing strategists of Carvana knows that the success and popularity of their online brand highly depends on the number of customers who sell and buy used cars through Carvana. It is often difficult for any customer to decide on what is the best sale price of their car and would end up either selling it at a very lower price than at what it could have been sold or at a rather higher price than usual. In both cases either of the buying or selling customers are dissatisfied. This leaves a negating effect on the Carvana's customer base. In order to attract both selling and buying customers the strategists came up with a unique idea, to invest in launching a new feature on their app/website that allows the customer to know what the best price could be to sell his/her car at.



Need: The strategists wanted to suggest the selling price to their new selling customers in a way that attracts both buyers and sellers so that it can end up as a win-win for both parties. They would like to investigate on what major aspects/features of a car did a buyer really looked for and then match these parameters with the car that is about to be listed on sale and make a rightful selling price suggestion.

Carvana		Price suggestion by the app		
		Too high	**Reasonable**	**Too low**
Customer	**Seller**	Gain	Satisfied	Loss
	Buyer	Loss	Satisfied	Gain

Vision: The strategists will make use of already available data history of sold-out cars and investigate what features are really causing a shift in the selling price and measure the sensitivity of change for these features so that they could build a predictive model for predicting the price. The best predictive model is to be chosen by running a set of already available machine learning models and pick the one with the least prediction error which could be Root mean square error or absolute mean error. This predicted price is then used as a selling price suggestion in the new feature.

Outcome: 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 be monetized. Once convinced, they would invest higher volumes of budget into the project for gathering more data and make the app more reliable.

Exploration

Problem statement: Its Required to model a price calculator to suggest a price to the customer. The best machine learning model to predict the selling price from previously available data is to be chosen.

Corresponding variable to predict would be "price_usd". The variable is a continuous numeric class which makes the problem statement regression by nature.

Summary of "price_usd" *without transformation, skewness at* **2.235308**

```
## [1] 2.235308
```

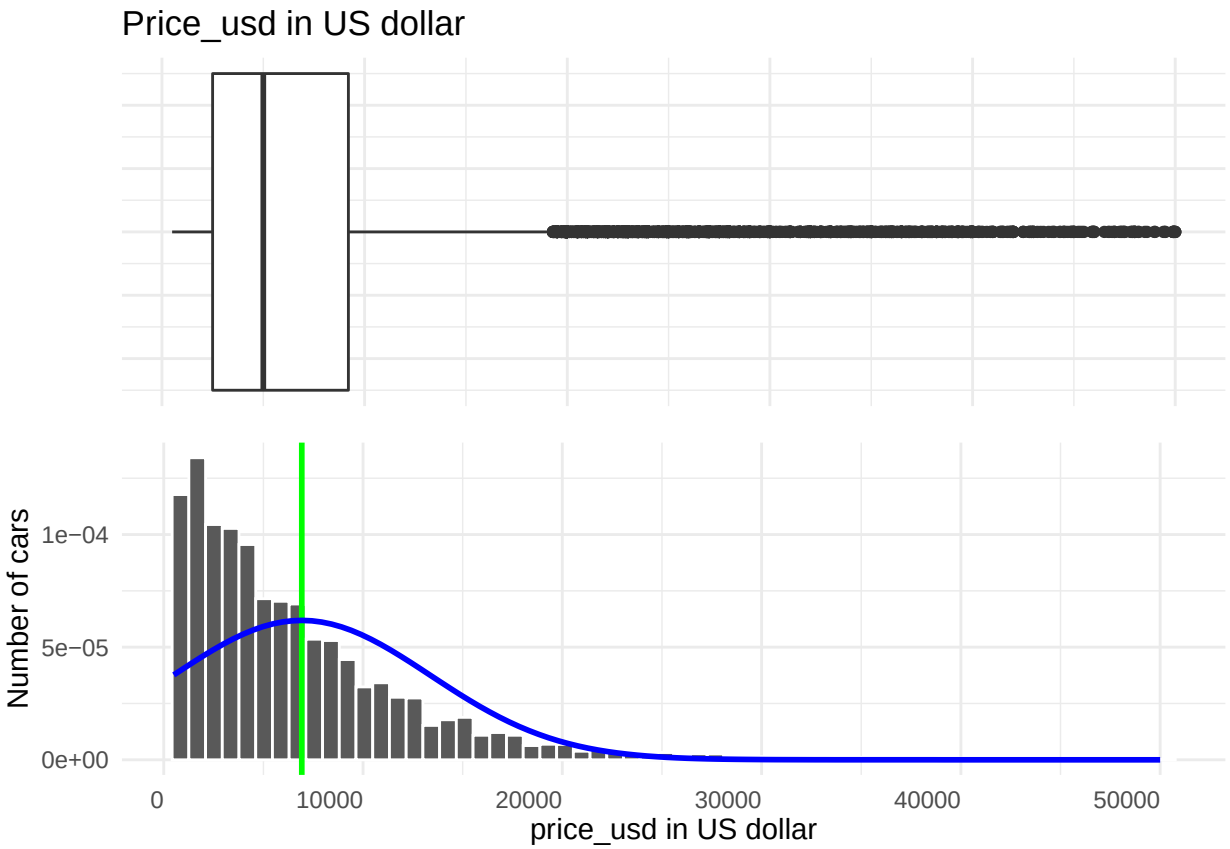
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   501.5  2499.0  5000.0  6925.4  9200.0 50000.0
```

Summary of "price_usd" *with log transformation, skewness at* **-0.1938202**

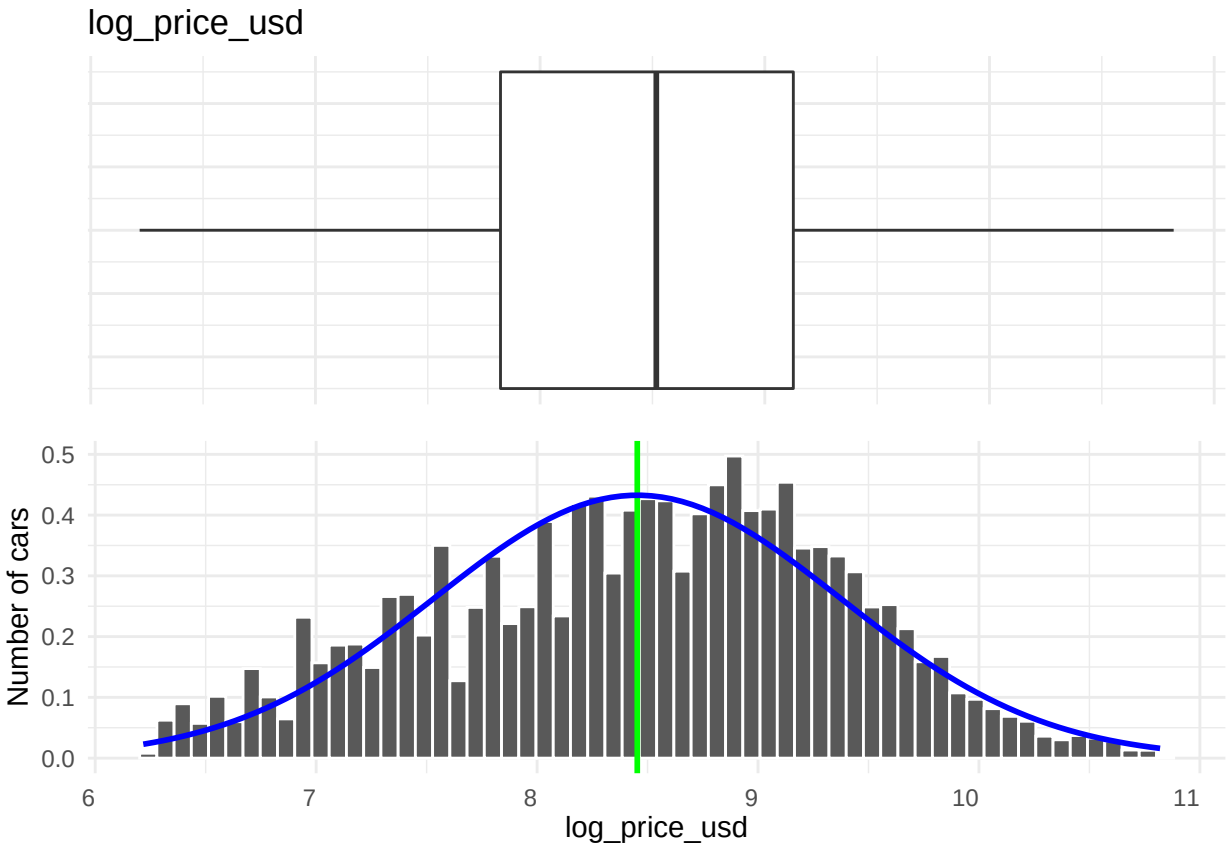
```
## [1] -0.1938202
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.218   7.824   8.517   8.454   9.127  10.820
```

Boxplot and distribution of "price_usd" *without transformation, mean at $6925 (green line)*



Boxplot and distribution of "log_price_usd" ***price_usd with log transformation, mean at 8 (green line)***



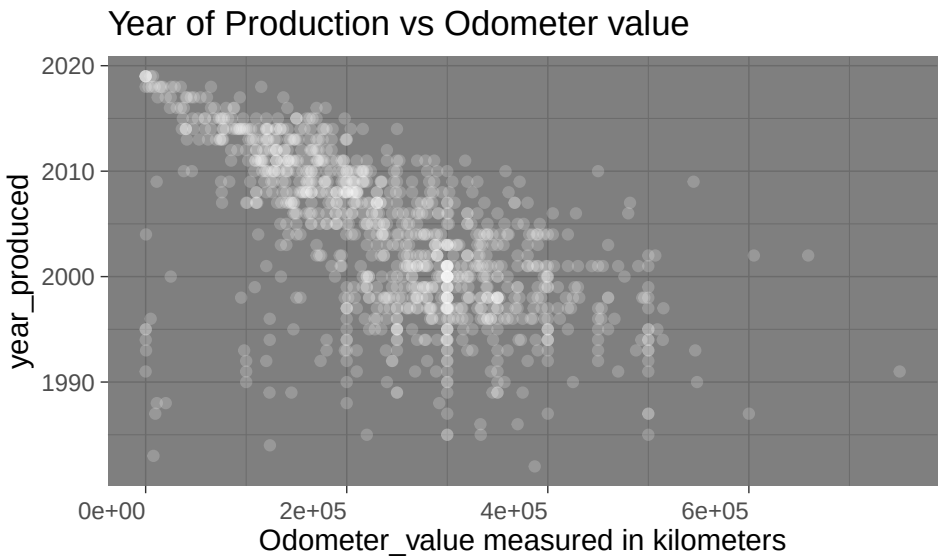
List of predictors in the data set

- **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.

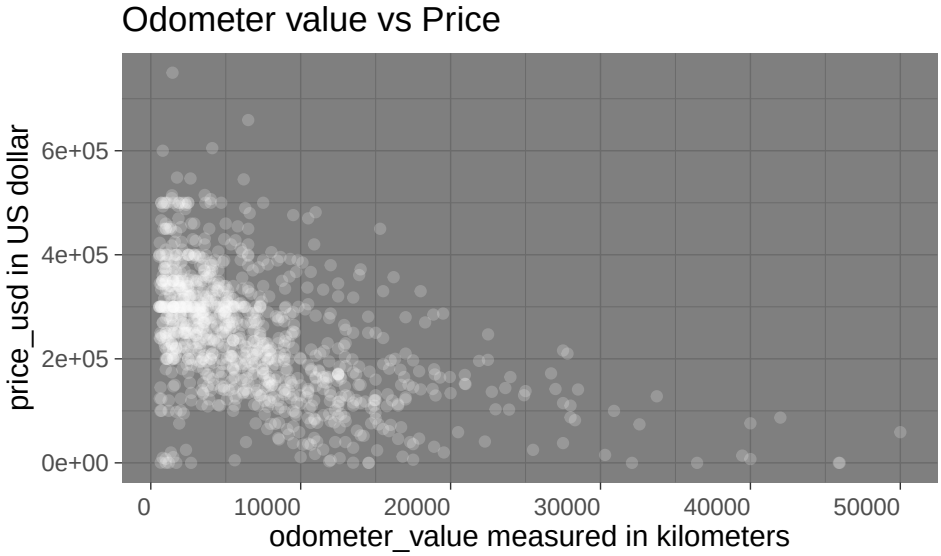
Potential predictors

year_produced, odometer_value and price_usd shaping the trends

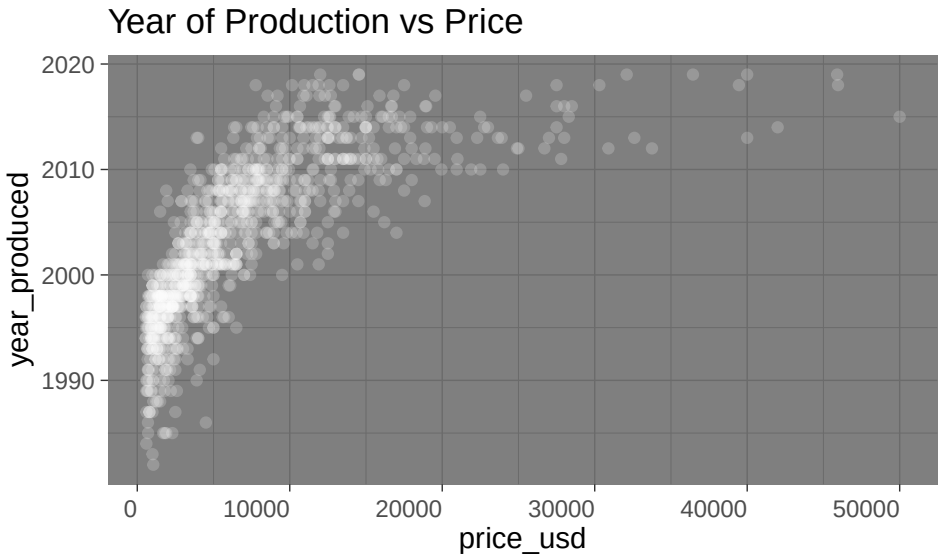
older cars are driven more so they have higher odometer values.



Higher odometer values make a car less reliable, and the prices are low.

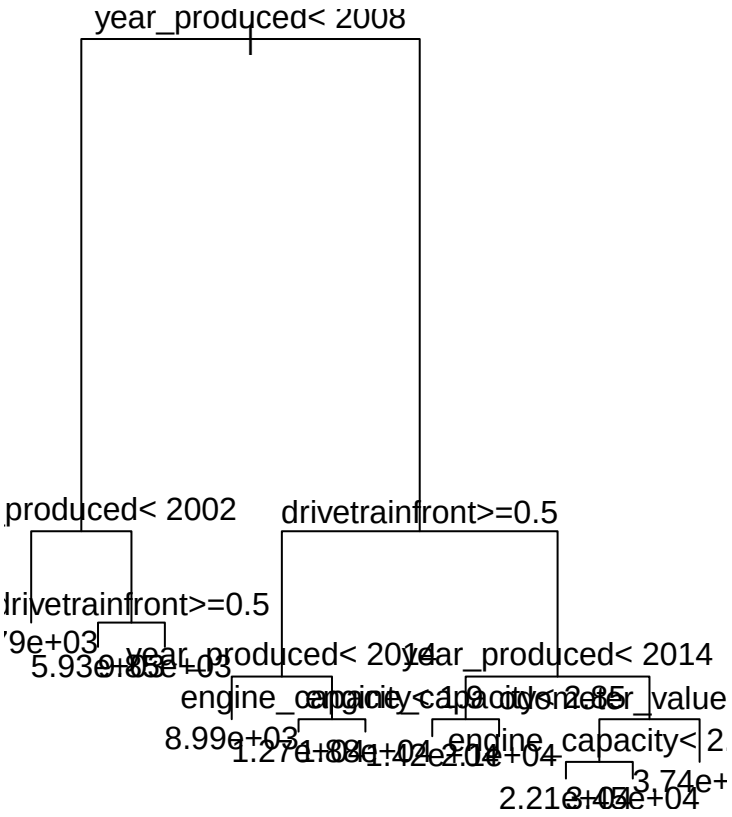


As a result, older cars are cheaper



Interesting fact: Year of Production governs most of the trend in the data while the other predictors are responsible for the variation.

The machine learning regression tree model showing year of production at the top of the tree.


year_produced< 2008
produced< 2002
drivetrainfront>=0.5
drivetrainfront>=0.5
9e+03
5.93e+03
year_produced< 2014
year_produced< 2014
engine_capacity< 1.9
city_meter_value
8.99e+03
engine_capacity< 2
3.74e+
2.21e+04

Anomalies in the data set

- 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.

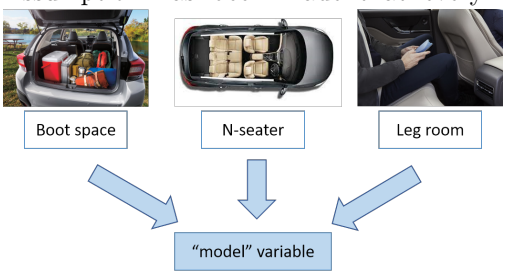
- If appearance is unknown the data might me deceptive

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.



Prediction models

Linear model with no transformation. RMSE at $3359 and MAE at $2197

```
##       RMSE      MAE
## 1 3359.927 2197.775
```

Linear model with log transformed outcome. RMSE at $8724 and MAE at $6586

```
##       RMSE      MAE
## 1 8724.154 6586.262
```

Linear model with transformed outcome and predictors. RMSE at $8695 and MAE at $6581

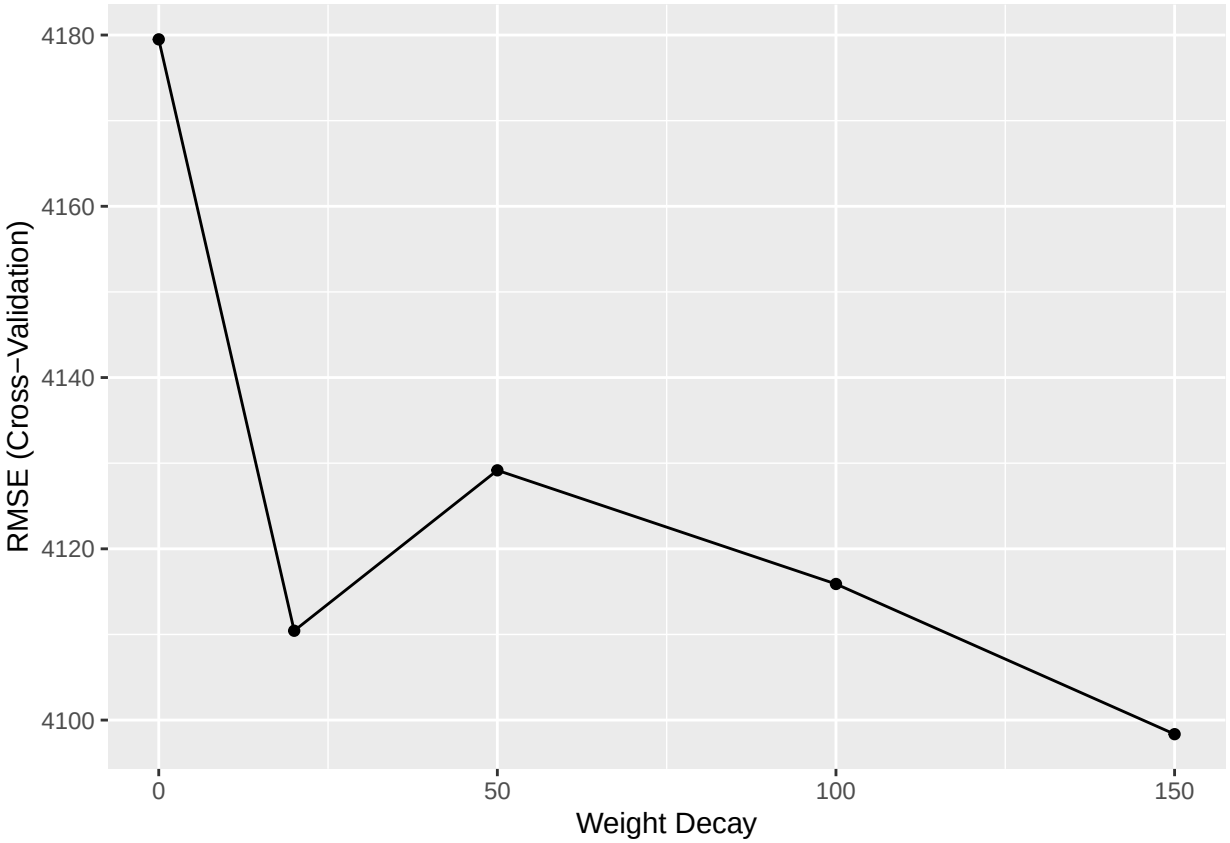
```
##       RMSE      MAE
## 1 8695.936 6581.093
```

Linear model with cross validation 10 folds (No transformation). RMSE at $3549 and MAE at $2264

```
##       RMSE      MAE
## 1 3359.927 2197.775
```

Neural net with cross validation 10 folds (No transformation). RMSE at $4676 and MAE at $2897

```
##       RMSE      MAE
## 1 3827.151 2476.821
```

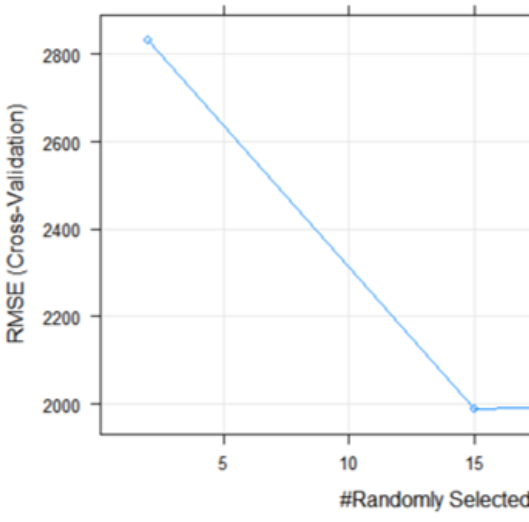


Random Forests with no pre processing. RMSE at $1932.

```
[1] 1932.693
Random Forest

27399 samples
   29 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 21919, 21919, 21919, 21920, 21919
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    2829.856  0.8343126  1733.258
  15    1989.646  0.9058637  1151.228
  29    2012.573  0.9031797  1150.452

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 15.
```

	mtry <dbl>
2	15

1 row



Future plan into analysis

As Random forest shows a promising result, investigation into decision tree is necessary that may help in predicting better outcome.
Improvements to be made with Features Engineering (contribution/sensitivity of predictors to the error), Ensembling techniques (bagging, boosting, bootstrapping), Hyperparameters tuning (complexity parameters)

Difficulties

Handling 36,533 observations is overwhelming.
Run time for Random forest is about 2 hours.