PREDICT CAR SALE VALUE USING REGRESSION ANALYSIS

SUBMITTED BY:

HARSHITA PATIL

AGENDA

- Why Prediction Model?
- About Data
- Why Regression?
- Understanding of algorithms of models
- Modeling
- Metrics (RMSE and R-Squared)– Both training and validation dataset
- Correlation Accuracy of Actual Vs Predicted
- How Model can benefit you?
- Future Scope

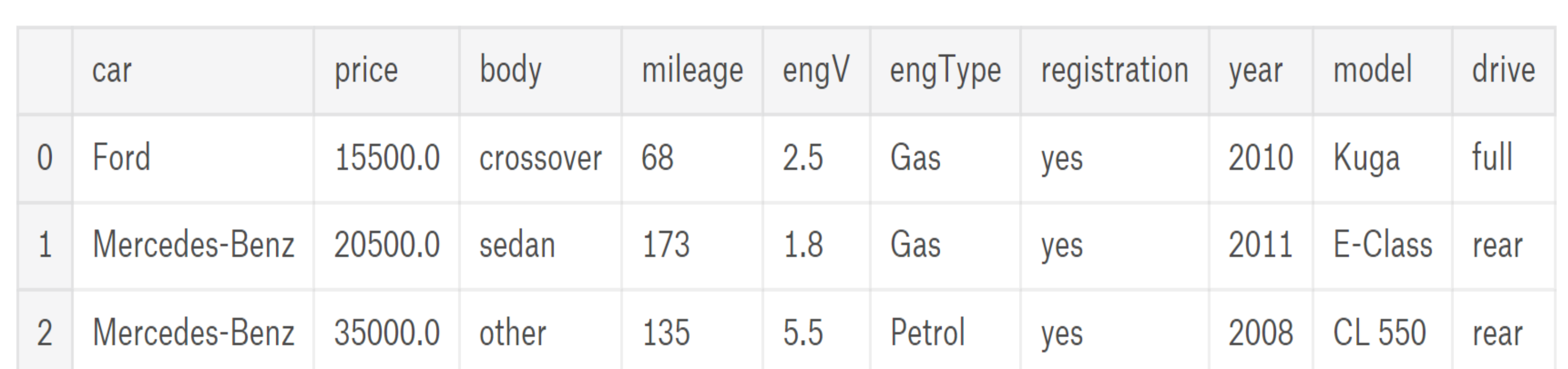
WHY PREDICT?

- The United States is home to the second largest passenger vehicle market of any country in the world, second now to China. Overall, there were an estimated 263.6 million registered vehicles in the United States in 2015, most of which were **passenger vehicles. -** *Passenger vehicles in the United States page of Wikipedia says*
- It is a proposal to car sales business.
- The model is going to predict the sale value of the **used** car on the basis of the data of its features gathered from the advertisements.
- Predicting the value of product in business helps business to gain the sales which are likely to be missed and allows to plan the business growth by making intelligent business decisions.
- It is also beneficial for the customers. When shopping for a used vehicle, typically an overriding concern is: *Am I paying too much?*
- Determines if the asking price for a particular car is reasonable given the information provided in the listing.

ABOUT DATA

- Collected from the car sales advertisement for study.
- Contains data for more than 9.5K cars of years ranging from 1953 - 2016 .
- Price of car ranges from $ 259 to $ 547,800.
- These are all used cars so it opens the possibility to analyze features' significances to determine car value.

- Contains 9576 rows and 10 variables with essential meanings.
- Variables are both categorical and numerical types.
- Data has gaps. Variables has N/A value which needs imputations.
- Price of car is $ 0 which is impossible.

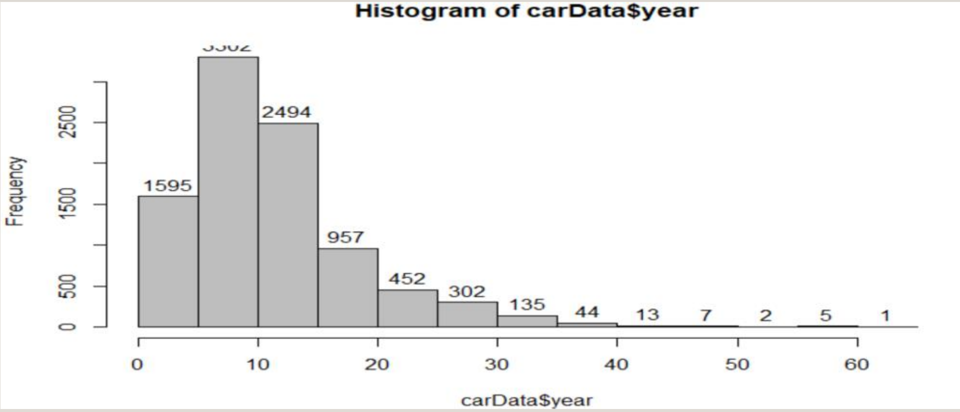


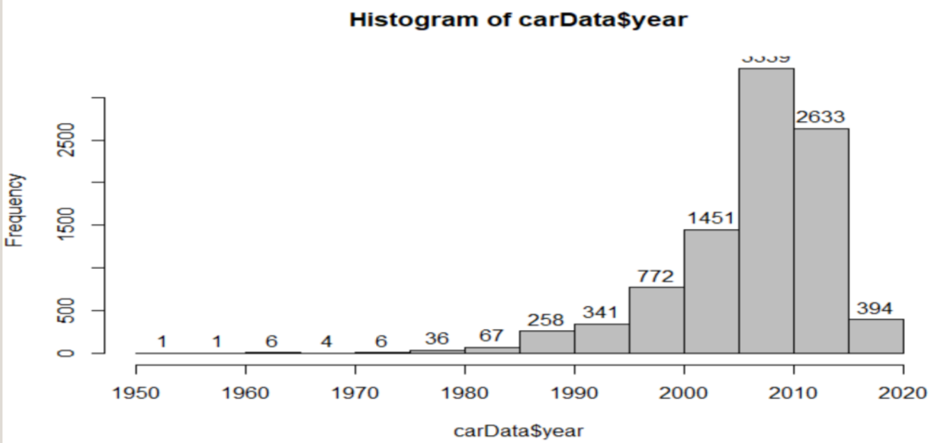
	car	price	body	mileage	engV	engType	registration	year	model	drive
0	Ford	15500.0	crossover	68	2.5	Gas	yes	2010	Kuga	full
1	Mercedes-Benz	20500.0	sedan	173	1.8	Gas	yes	2011	E-Class	rear
2	Mercedes-Benz	35000.0	other	135	5.5	Petrol	yes	2008	CL 550	rear

AFTER WRANGLING

The final datasets contains –

- 9309 unique records
- 87 Makes
- 888 Models
- Oldest make is of year 1953.
- The model considers variety of cars.
- Regression Analysis to derive model



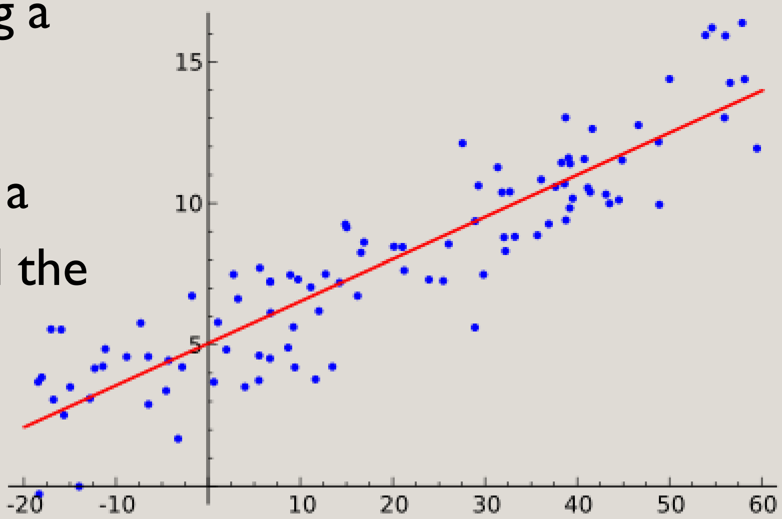


WHY REGRESSION?

- Regression analysis is a statistical analysis, where given a set of independent variables, we can predict the outcome of a dependent variable.
- It is used to predict the result of a quantitative (numerical) variable. In our case the predicted value of car.
- Multiple Linear Regression/ Generalized Linear Model – GLM
- K- Nearest Neighbors – KNN
- Support Vector Machine – SVM

MULTIPLE LINEAR REGRESSION

- Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.
- It is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.
- $Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + b_k X_k$
- Predicted price= a+ b1*car + b2*model + b3*mileage +….+ error

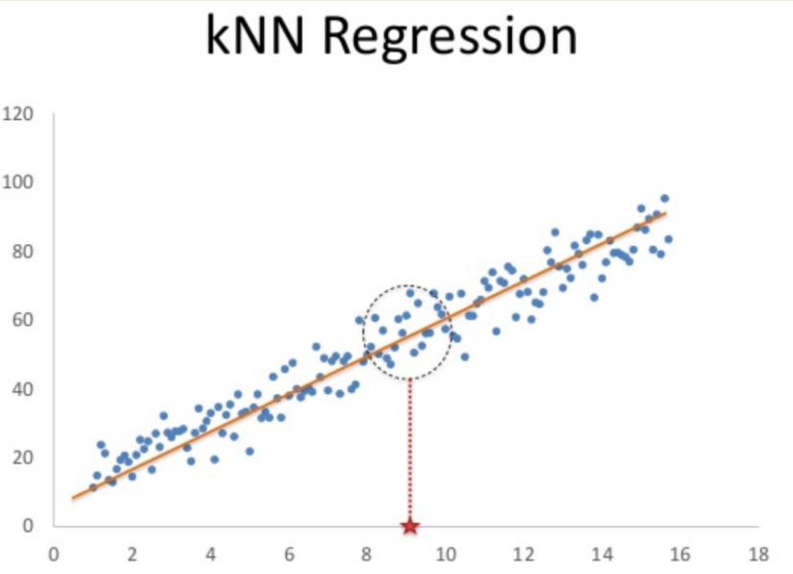


K – NEAREST NEIGHBORS – REGRESSION KNN

- It is non-parametric method so provide more flexible approach.
- K nearest neighbors is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measure (e.g., distance functions)
- Assume a value for the number of nearest neighbors K and a prediction

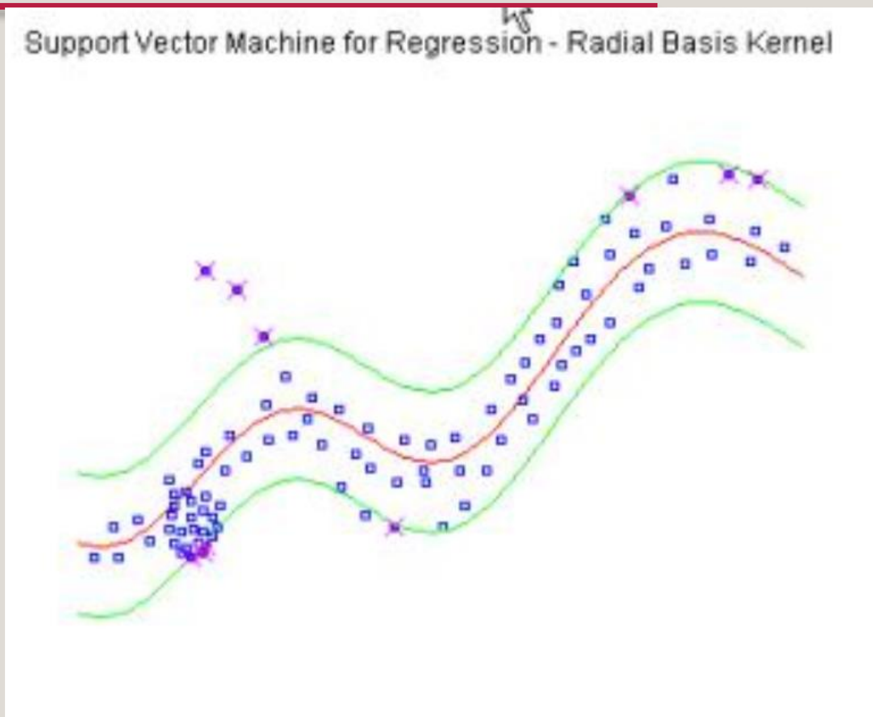
point xo.

- KNN identifies the training observations *No* closest to the prediction point x
- KNN estimates f (xo) using the average of all the responses in *No*, i.e.
- f(xo) = $\frac{1}{K}\ \sum_{xi \in N0} yi$



SUPPORT VECTOR MACHINE - REGRESSION SVM

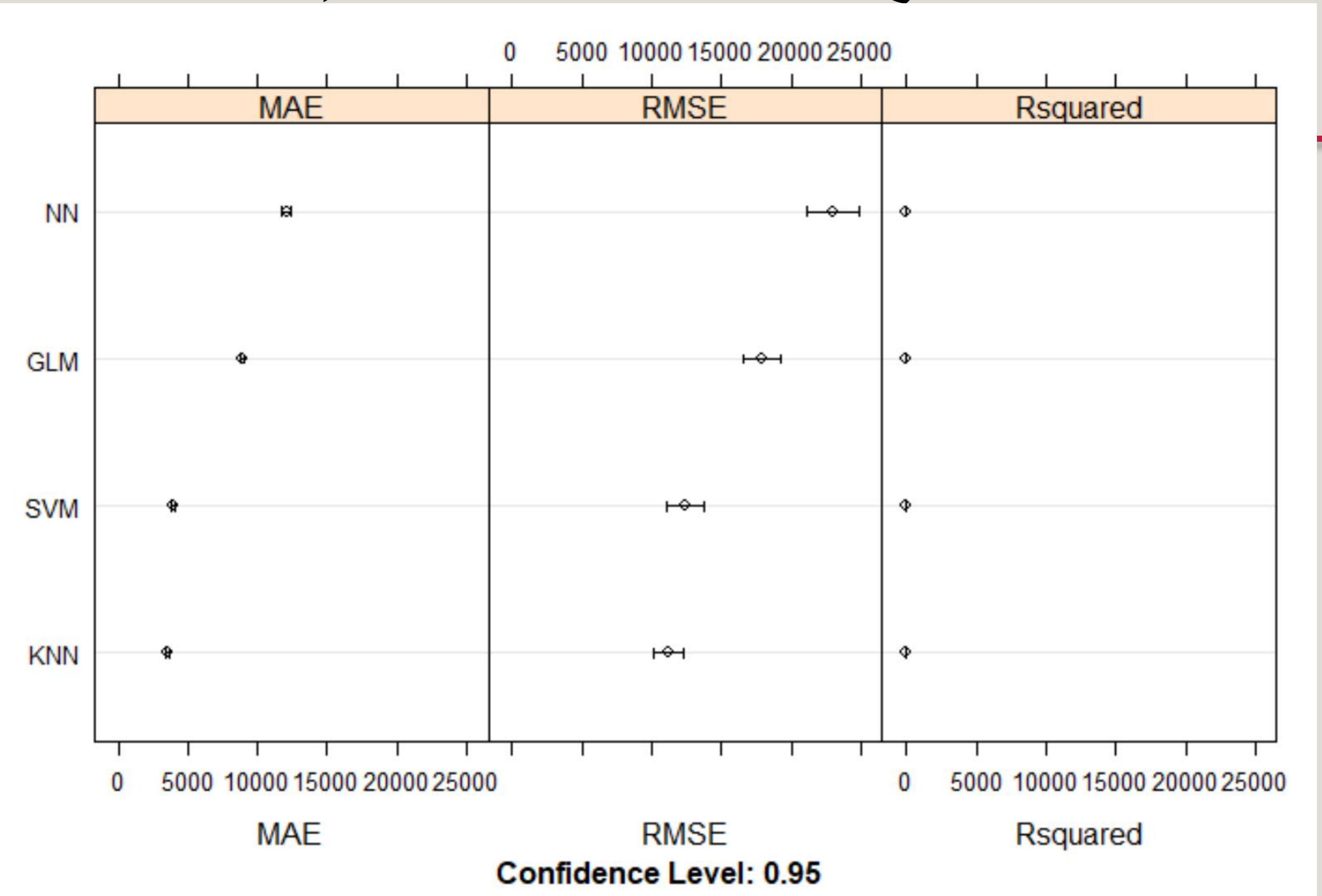
- Like other regressors it also tries to fit a line.
- SVR, you can deploy a non-linear kernel (here – radial basis function) and end up making non-linear regression, i.e. fitting a curve rather than a line.
- Make non-separable separable



MODELING

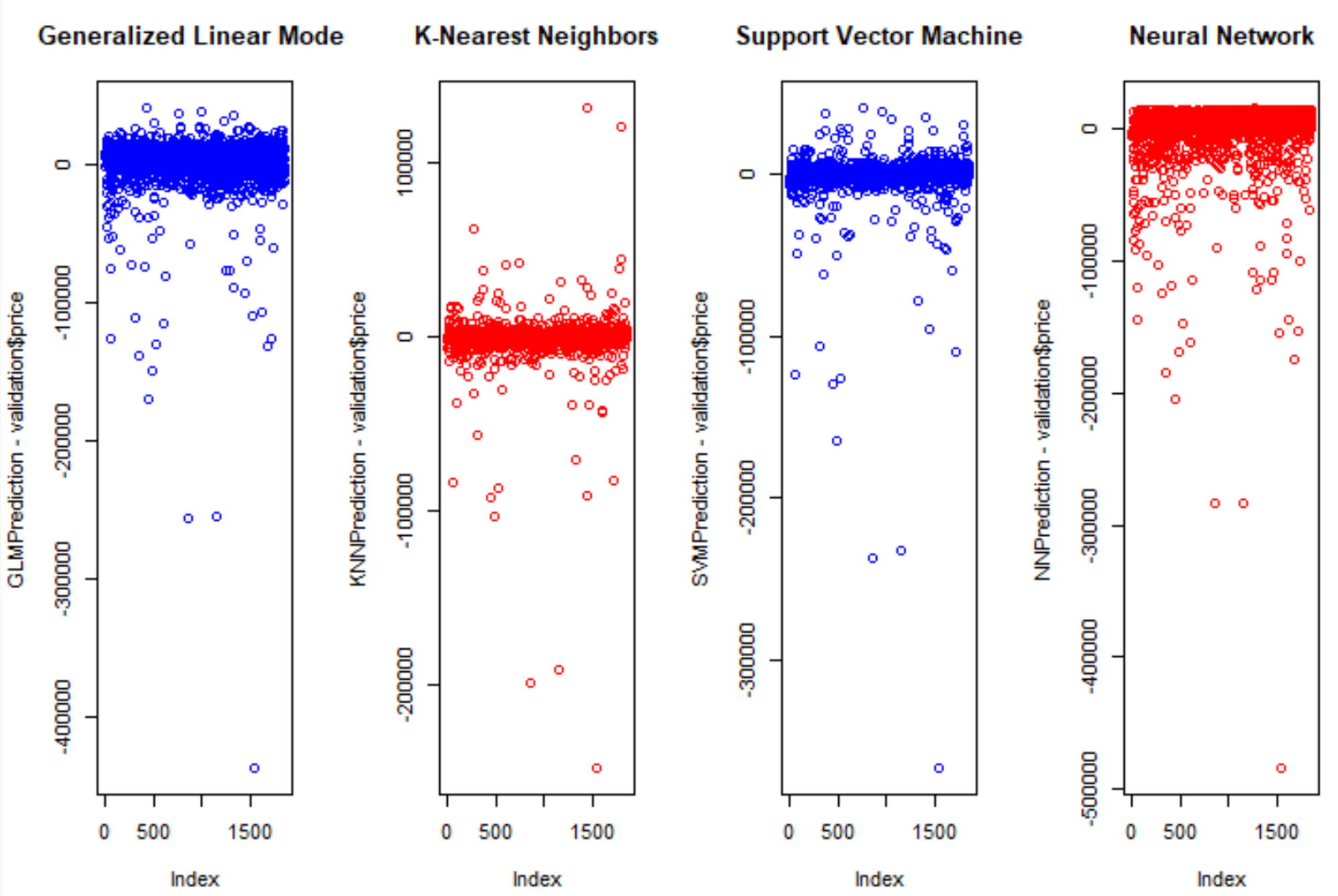
- Training Datasets = 80%
- Validation Dataset = 20%
- **10- fold cross validation on training datasets to generate model** - Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample(training data).
- Cross-validation is used in applied machine learning to estimate the skill of a machine learning model on unseen data.
- That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

METRICS ON TRAINING DATASET - MAE, RMSE AND R-SQUARED



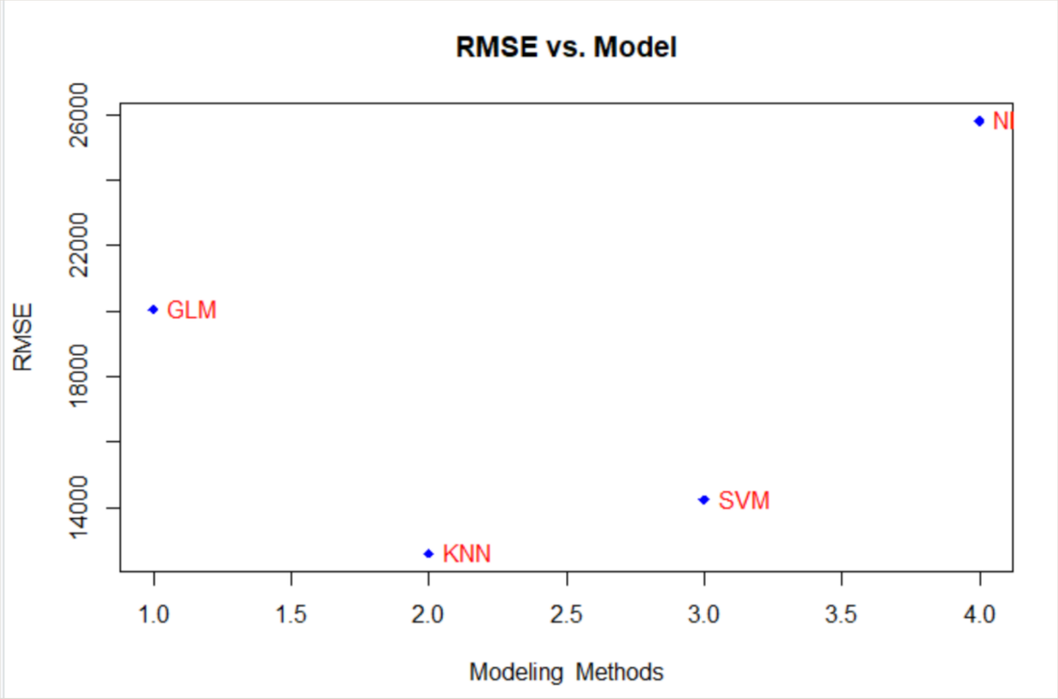
PREDICTION ON VALIDATION DATASET (ACTUAL – PREDICTED)

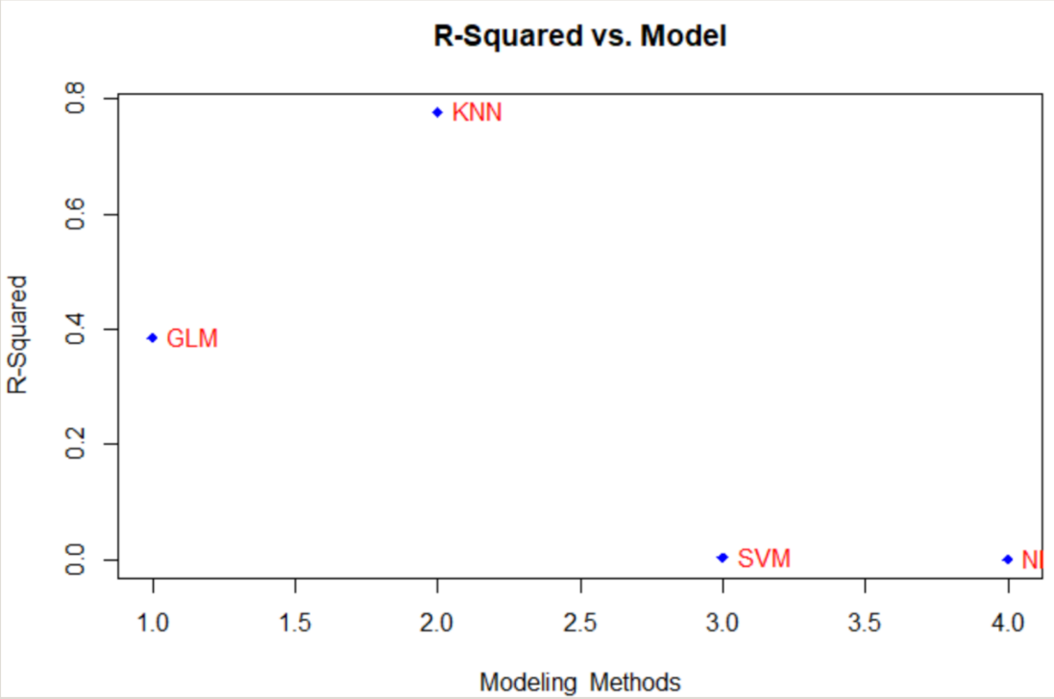
- We then applied the models on validation data and predicted the predicted price of cars
- Plotting the **difference** of Predicted Price and Actual Price.
- We could notice the most distribution/variance is in model GLM and NN.



METRICS ON VALIDATION DATASETS RMSE AND R-SQUARED

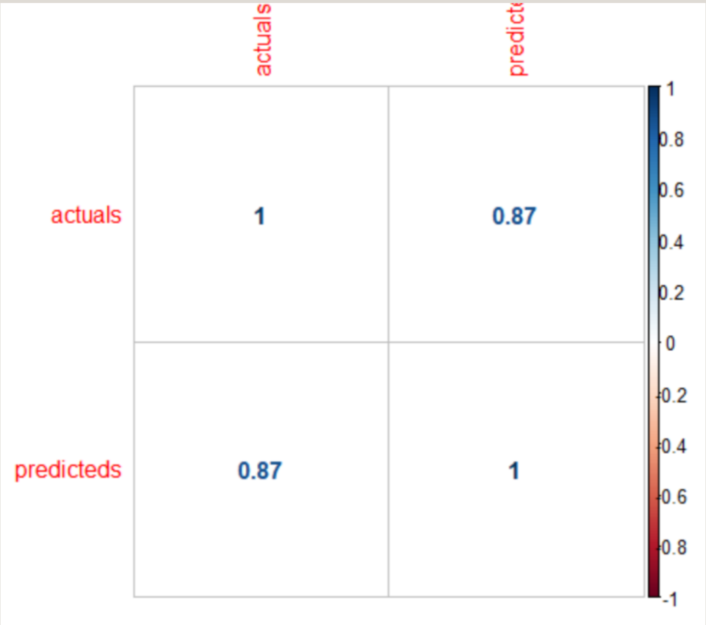
KNN has least RMSE value and highest R-Squared value. Hence the recommended model is KNN





CORRELATION ACCURACY - ACTUAL V/S PREDICTED

- The correlation accuracy obtained for KNN model is 87%.
- Whereas for GLM and SVM is 63% and 82% respectively.
- Thus KNN model best fit for the case compared to SVN and GLM.



	actuals	predicte
actuals	1	0.87
predicteds	0.87	1

HOW CAN IT BENEFIT YOU?

- If you are business, chances of losing sales reduces by making predictions about future pricing information for the item
- Remove human bias/prejudice/“gut instinct.”
- Helps smaller ecommerce merchants stay competitive.
- If you are customer, you know if asking price is reasonable or not.
- Improves customer engagement and increases revenue.

FUTURE SCOPE

- Look more in-depth at pricing strategies between different dealers, regions, seasons, etc.
- Investigate more thoroughly the factors that go into pricing. There are many other factors that can be used -
  - The braking system, Exterior and interior color
  - Safety features etc.
- Collect data over time and predict price trends for each vehicle
  - How fast will it depreciate?
  - What determines the rate of depreciation?
- Build in a way to make the output attractive and interactive

THANK YOU!