

How much is my car worth?
A methodology for predicting
used cars prices using
Random Forest

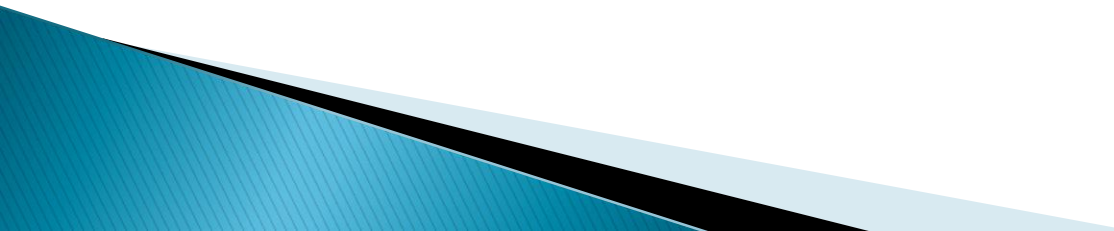
Published By

- ▶ Nabarun Pal
- ▶ Puneet Kohli
- ▶ Sai Sumanth Palakurthy
- ▶ Dhanasekar Sundararaman
- ▶ Priya Arora

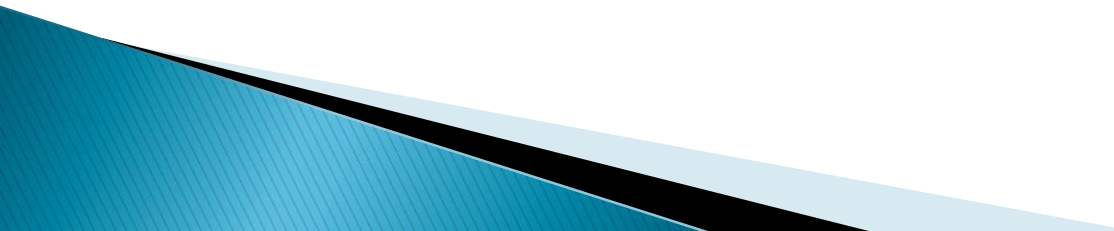
Published In : Future of Information and
Communications Conference (FICC) 2018




Supervisor & Team Members

- ▶ Supervisor: Dr. Sunil Agarwal
 - ▶ Pradeep Kumar 2015181
 - ▶ Rahul Gupta 2015196
 - ▶ Vipin Dhonkaria 2015274
- 

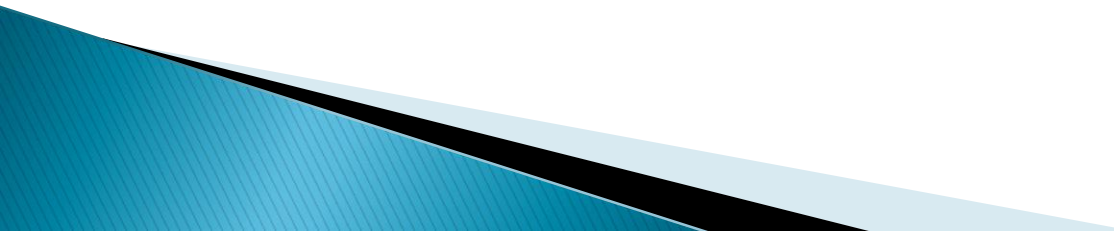
Abstract

- ▶ The rise of used cars sales is exponentially increasing. Car sellers sometimes take advantage of this scenario by listing unrealistic prices owing to the demand.
 - ▶ Therefore, arises a need for a model that can assign a price for a vehicle by evaluating its features taking the prices of other cars into consideration.
 - ▶ The model has been chosen after careful exploratory data analysis to determine the impact of each feature on price.
- 


Introduction

- ▶ The prices of new cars in the industry is fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes.
 - ▶ So customers buying a new car can be assured of the money they invest to be worthy.
 - ▶ But predicting the prices of used cars is an interesting and much-needed problem to be addressed.
 - ▶ Customers can be widely exploited by fixing unrealistic prices for the used cars and many falls into this trap.
 - ▶ Therefore, rises an absolute necessity of a used car price prediction system to effectively determine the worthiness of the car using a variety of features.
- 

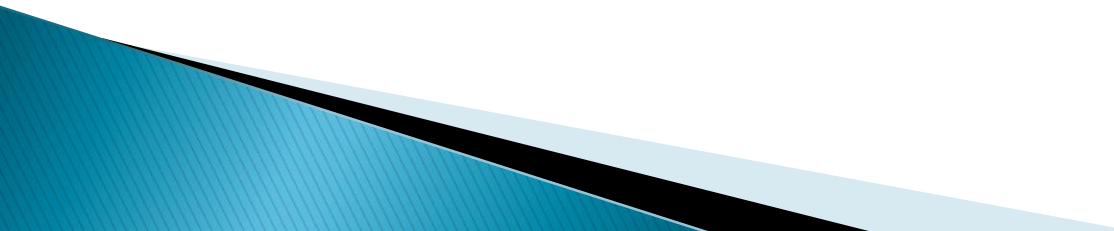
About Dataset

- ▶ “Used Car Database” from Kaggle which is scraped from eBay–Kleinanzeigen, the German subsidiary of eBay, a publicly listed online classified portal.
 - ▶ The dataset contains the prices and attributes of over 370,000 used cars sold on the website across 40 brands.
 - ▶ Our dataset contains 20 unique attributes of a car being sold.
- 


Features of dataset

- ▶ Price, Vehicle type, Age, Power PS, Model, Brand, Kilometre, Fuel type, Damage repaired and Is Automatic.
 - ▶ Out of these features, the most important for our prediction model are
 - ▶ 1. Price: The specified asking amount for the car
 - ▶ 2. kilometre: A number of Kilometres the car has driven
 - ▶ 3. Brand: The car's manufacturing company
 - ▶ 4. Vehicle type: Whether a small car, limousine, bus, etc.
- 

Related Work

- ▶ "Advanced data science systems and methods useful for auction pricing optimization over network." by Strauss, Oliver Thomas, and Morgan Scott Hansen.
 - ▶ "Model of Predicting the Price Range of Used Car" by Xinyuan Zhang , Zhiye Zhang and Changtong Qiu in 2017
 - ▶ "Predicting the price of used cars using machine learning techniques." by Pudaruth, Sameerchand in 2014
- 

Random Forest

- ▶ Random forests work as a large collection of decision trees. It is an ensemble learning model for classification and prediction.
 - ▶ In this technique, the given large training dataset is divided into many random subsets.
 - ▶ Since every data subset is randomly made, each subset is known as random tree and all random trees are collectively forming a random forest.
 - ▶ For each subset, a decision tree is being constructed at training time.
- 

Random Forest

- ▶ Polling is conducted among the decision trees to predict the class label for the given instance.
- ▶ Random Forest: Can be run efficiently on large databases.

Coefficient of Determination

- ▶ In statistics, the **coefficient of determination**, denoted R^2 or r^2 and pronounced "R squared".
- ▶ It is the proportion of the variance in the dependent variable that is predictable from the independent variable.
- ▶ The main purpose is either the prediction of future outcomes or the testing of hypotheses, on the basis of other related information.
- ▶ It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

Coefficient of Determination

A data set has n values marked y_1, \dots, y_n (collectively known as y_i or as a vector $y = [y_1, \dots, y_n]^T$), each associated with a predicted (or modeled) value f_1, \dots, f_n (known as f_i , or sometimes \hat{y}_i , as a vector f).

Define the **residuals** as $e_i = y_i - f_i$ (forming a vector e).

If \bar{y} is the mean of the observed data:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

then the variability of the data set can be measured using three **sums of squares** formulas:

- The **total sum of squares** (proportional to the **variance** of the data):

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

- The regression sum of squares, also called the **explained sum of squares**:

$$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

- The sum of squares of residuals, also called the **residual sum of squares**:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$