# Predicting Price of Used Cars

**Saurabh Kurjekar, Rishabh Mohan, Jingda Zhou, Siyu Zhu**

Purdue University, Krannert School of Management, 403 W. State Street, West Lafayette, IN 47907

## **skurjeka@purdue.edu, mohan35@purdue.edu, zhou854@purdue.edu, zhu554@purdue.edu**

## **Abstract**

This DSS aims to help private sellers and buyers decide the market value of a used vehicle. By understanding the influencing factors of the used car price, the DSS build and evaluate a multi-regression model that predict the price of used car sold online. This is important because purchasing used vehicles online is an important question people commonly make. And private sellers and buyers risk losing profit or increasing cost, respectively, without sufficient market information. In order to help them to decide on the price, the DSS build a shiny app that identify key predicting variables and the according market value.

## **Keywords:** prediction, price, DSS, used

## **Business Problem**

* Brief description and stakeholders

The core business problem our decision support system (DSS) is to determine the market price of used vehicles. In order to decide the price, our DSS identified key factors that influence the price of the car, and help predict the price given certain parameters. The key stakeholders are divided into two types: a) private used car sellers, and b) private used car buyers.

* Discussion whether the problem is amenable to an analytics solution

Because the business problem is straight-forward, it is fairly easy to convert it into analytics problem. The major challenge is to identify the key influencing factors in this problem, and avoid overfitting while maintaining comprehensiveness of the model.

maintaining comprehensiveness of the model.

* Refinement of the problem to identify any delineate constraints

Time is a primary constraint of our business problem. In order to build a functional DSS within the limited amount of time, we needed to quickly decide the set of most useful variables. There can be other variables that can contribute to the accuracy of the model, but for the time being, we only selected the most significant ones.

* Define the initial set of business benefits

The significance of the DSS lays in helping private sellers and buyers decide the market value of a certain vehicle. Because Used car market place is influenced by various factors, and without sufficient information, sellers can easily overprice or underprice the vehicles, and thus lose profit in the transaction. More specifically, the DSS can inform sellers of the market price based on the other sellers from the past.

* The business initial benefits concern:

a) To help private sellers decide the “right” selling price; and

b) To assist private buyers, estimate the cost of a used car within certain parameters.

* The agreement on the business problem statement has been determined by the stakeholders.

**Analytics problem**

The analytics problem in the DSS is to select the most relevant predictors, and to develop a model that achieves reasonable prediction accuracy. Specifically, the key analytics problems include selecting variables as well as building, evaluating, and refining models.

## Develop a proposed set of drivers and relationships to outputs

In order to select the qualified set of drivers, our team checked the content of each potential predictors in the dataset, and removed those that are not directly relevant to the business problem. developed a correlation matrix for all numerical variables and dummy coded factor variables. As a preliminary step, we fed the variables into the linear model, and checked significance of the overall model and individual predictors. The model turned out to be significant, hence all the predictors are kept at this step.

## The key set of assumptions related to the problem

There are two assumptions related to the problem.

Firstly, the outcome variable—price is not affected by inflation factor during year 2016, which is the selected year when all transactions occurred.

Secondly, we assume that the relationship between predictors and the outcome is linear.

* Key metrics of success are defined as:

1. User-friendliness of the shiny app
2. Business relevance of the predictors
3. Accuracy of the model

## Statement claiming stakeholder agreement on the approach

Agreement on the approach has been determined by the shareholders.

## **Data**

## Acquiring, cleaning, reshaping data, and preliminary analysis

The dataset used to develop the current model is downloaded from Kaggle and provided by eBay (originally 370,000 rows and 9 columns). Initial data cleaning includes removing null values, reshaping columns After cleaning the data, we did a set of descriptive analyses to get better understanding of each variable. Boxplots, tables and correlation matrices were created to help us recognize existence of any influencing factors of vehicle price, count of records in each level of factor variable, as well as correlations between individual variables. Because no significant correlation is found between any numerical variables, we decided to dummy code categorical variables and examine the effect of the dummy variables on price. A primary data check is performed, and all rows containing null values are removed.

**Methodology Selection**

* Identify a few problem-solving approaches (methods) and use one or two

The problem can be solved by creating a predictive and an optimization model. For our project, we chose to create a prediction model with some descriptive statistics. The reason behind using prediction model is that it is more important to predict the price and then optimize it. The seller would know exactly what to sell, what car type, model, mileage, and so here prediction model works best, and optimization might not be very useful.

We could have used multiple prediction models like non-linear models or random forest to predict the price of the used cars. But we chose to use linear multiple regression, as we are very familiar with this model, and also assuming that our variables have a linear correlation with price.

With various bar graphs, most popular brand by quantity sold, and most preferred brand by price of the vehicle were studied. Also, the most preferred category of vehicle out of various types like SUV, sedan, coupe etc. was observed.

For all the categorical variables, table function was used to count the number of records. Then, a box and whisker plot was created for the brands with maximum counts.

To predict the value of used cars, which would be useful for both buyers as well as sellers, a predictive model was built using multiple linear regression.

* Discuss why R is a viable tool to use

R is an excellent tool for data analysis as it is an open source platform. With its specialized packages, it is a great tool for solving various analytical and statistical problems which saves a lot of time. With GGplot2, the data can be visualized aesthetically with various kind of plots.

* Test and select an approach or approaches you believe might work

We analyzed the problem with correlation matrix and correlation plot first. We found very good correlation between some variables. Year of registration had a positive correlation and mileage (no. of miles driven) had negative correlation with price. We also built a linear multiple regression prediction model for small dataset just to test the model, and got a high R sq value.

**Model building**

* Run and evaluate the model(s)

With running descriptive analytics as discussed in the methodology section, we were able to get an idea about our variables. Various variables were dropped from the dataset which were not adding any significant value. Once the dataset was cleaned, multiple linear model was built with all variables included, and the p-values and T statistical values were observed. It was found that all the variables after first data cleaning were significant with p = 0. Adjusted R sq. was found to be 0.56. Next, the data was split into 2 sets, training (75%) and test (25%).

The mean of the price difference between the predicted values and the real values was -16.85$ with a standard deviation of $3484.62.

* Discussion of integrating the model back to the problem

With our multiple linear regression results, we could predict the value of any used car. We found that gearbox type as in automatic or manual would make a significant difference in the value of car. A manual gearbox car would decrease the price of the car significantly. Year of registration is the other variable which directly affects the price of the car. A relatively new car would be much costlier than an older car. One of the independent variables which doesn’t make much difference is the kilometers it has been driven.

* Discussion of any findings (including assumptions, limitations, and constraints)

It was assumed that there are no other significant variables while building the model. All the models which were included were significant by the p-values. Although Our car dataset from eBay is very popular and had about 300000 rows of data, it had some variables missing variables like miles per gallon, condition of the car, vehicle style etc. We are assuming that these variables would not make a significant difference in the value of the used cars. The model is only good for cars after year 1990 with at least $400 value.

**Functionality**

* Discuss what the DSS can do

Our DSS can predict the value of any used car which was registered after year 1990. This will be useful to both seller as well as buyer. The predicted value will be based on various factors like years of registration, kilometers it has been driven, vehicle type, brand, fuel type as well as on the gearbox

Our bar plots would also be able to tell which brand is the most popular by the number of counts and by the price. The box and whisker plot would tell the full statistics for the most popular 5 brands.

The linear multiple regression model could tell what variables are the most important while buying or selling a used car. If a buyer is on a tight budget, he or she could go for a relatively old manual gearbox vehicle. As, the kilometers the car has been driven is not making a significant difference, the buyer can look for relatively less driven cars but still older.

* Discuss any R packages you found useful

GGplot2 by Hadley Wickham was the most useful package apart from shiny package used to create the shiny app. With GGplot2, we could visualize data in the form of bar graphs with a one line code. With GGplot2, we could create complex graphs in an elegant manner.

* Did you have to write any conditional logic?

No, we did not have to use any conditional logic. The R shiny videos really helped us to create the app without using any conditional logic in the server.

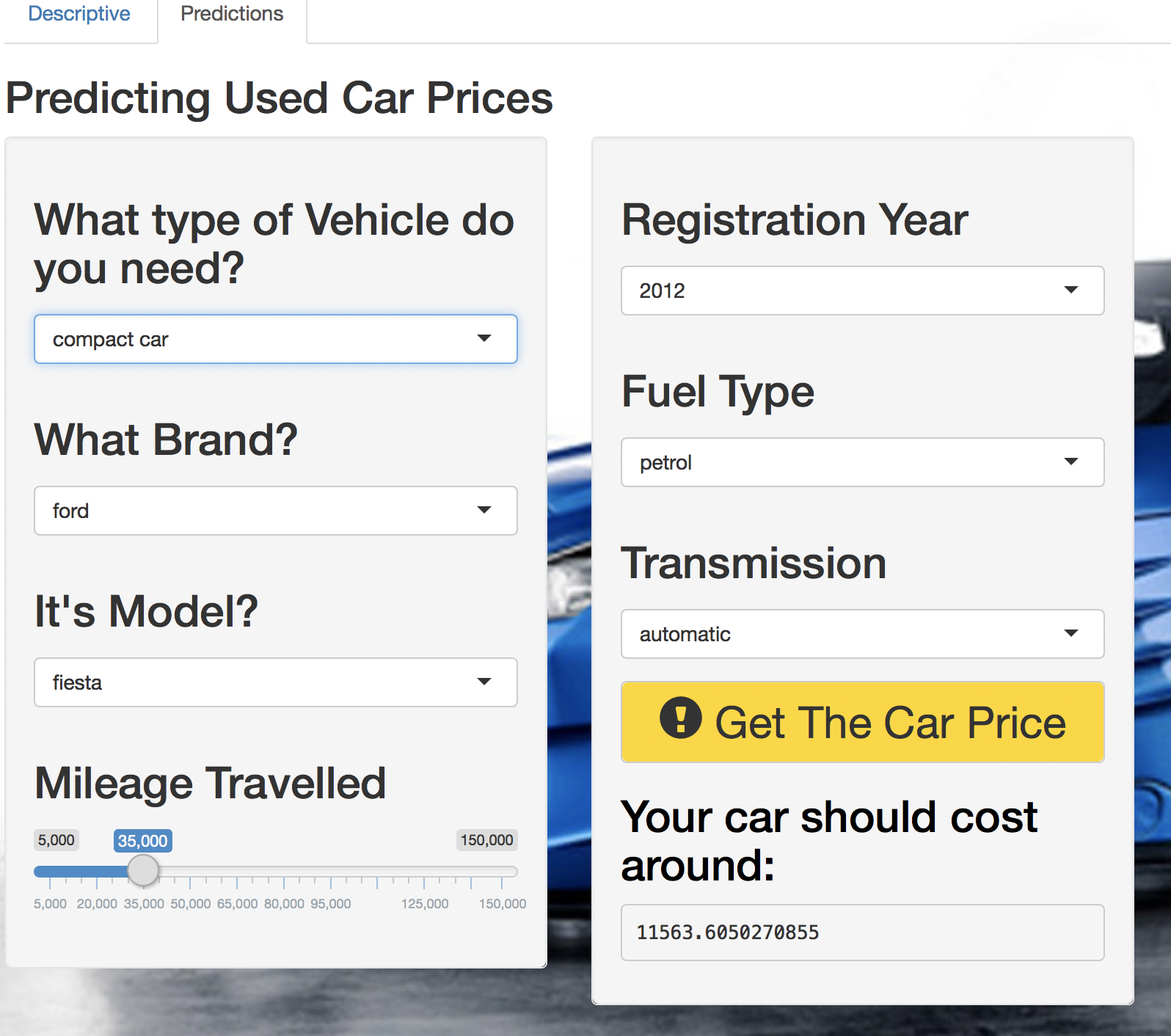
* If you had more time or experience, what other enhancements might you like to add?

If we had more experience as well as time, we would have built multiple models for prediction like non-linear model, k-nearest neighbor algorithm, random forests etc. We would have created reduced models or done partial F –testing, and would have dropped variables step by step, looking for the most significant variables. We would also have also taken multicollinearity and heteroscedasticity in account to see if there is any correlation between various independent variables.

**GUI Design & Quality**

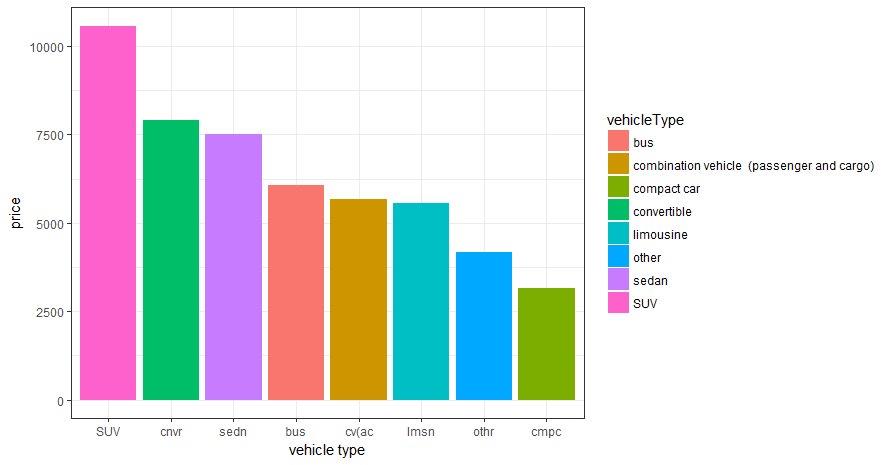
* Does the tool work without errors?

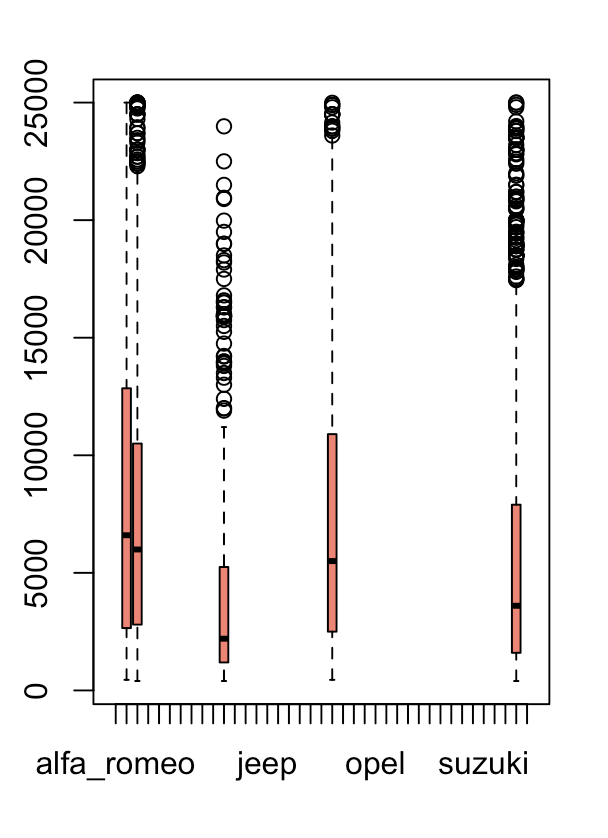
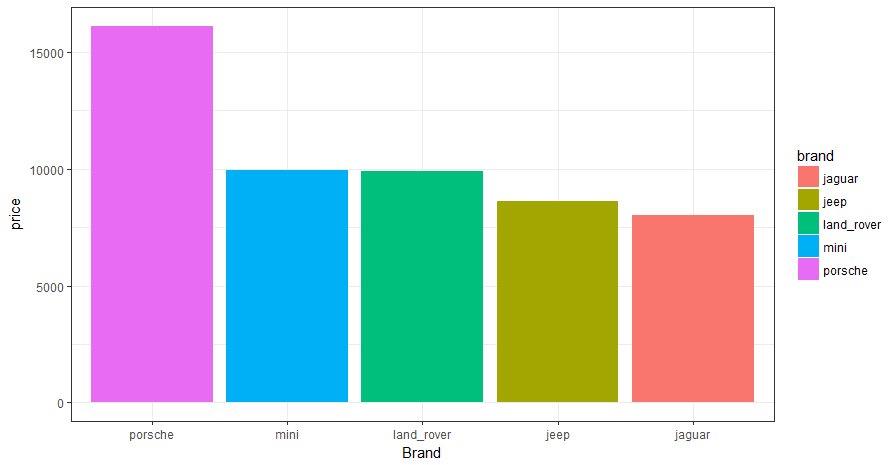
We got a lot of errors, but with step-by-step problem solving, we were able to get rid of all the errors, and our first shiny app works beautifully.



* Does it appear as good or better than the provided Shiny student examples from last year?

Although we did not create a very sophisticated application, our application is as good as the shiny examples provided by the last year students. Our app can predict the value of the used cars by taking various factors in account like model, car type, mileage etc. With this app, both buyers and sellers could predict the value of used cars without spending a lot of time on it.

****

****

**Conclusion**

In sum, we have created a wonderful application which can predict the value of the used cars by various factors like vehicle category, type, mileage, gearbox type and various other variables. We have used a big database consisting of about 370000 rows consisting of cars of carious brands and models. We have used descriptive statistics to know our variables better and then created a predictive model using multiple regression to predict the value of the used cars. The dataset was also split into training and test data, and test data (25%) was used to evaluate our predictive model. By putting this model in the shiny server, we are able to create a beautiful app, which can be easily used by the decision maker which is either buyer or seller in this case to accurately predict the value of the used cars. This app can be used as a standard benchmark in future to get the fair purchase or selling price based on what others have paid.