**DATA MINING FOR BUSINESS**

**PROJECT REPORT**

Group: Team 2

**Team Member:**

Ziyue Zhong

Shuangjie Ma

Wenxuan Tan

Qinruo Wu

Abstract

In the second-hand car market, customers to spend a lot of time on choosing a car with reasonable price. The purpose of this project is to find out the important factors that influence the price of a used car and build a price predict model. The conclusion of this project could be deployed to provide suggestions to both used car buyers and car sellers. Two datasets include used car dataset and demographic dataset are merged together. There are three models used in this project including regression model, neural network model and random tree model to do the analysis. ‘Time to sell’ is a potential indicator for price, gaussian weighting function is applied to calculate analytical weight which is added to analyzing models. We compare the results, model accuracy and get the conclusion. We found out the most important attribute that affect the price and used the coefficient get from the regression model to build the used car price predict model.

Introduction

Nowadays, instead of a brand-new car, second-hand car has become more and more popular in the market. The key reason people choose to buy used cars is always its low price. However, due to the variety of car brand and other detailed information, customers always have to spend a lot of time and money doing research about the used car market, compare different cars and find the most reasonable price. In order to provide an easier way to know the proper price for each used car in the market, we first find out the attributes that affect the price most and build a price prediction model based on the used car database containing car attributes and its sales information crawled from eBay website. We are trying to help car sellers sell their car in the desired way they want, they may choose to sell their car in a different location , a different time, ect. We also trying to help the second car buyers to decide which kind of car worth to buy and make the most of their money.

**Data Description**

There are 2 databases are used in project.

1. **Autos: Used Car Database over 370000 used cars scraped from Ebay**

**Data source**: Kaggle

**URL**: https://www.kaggle.com/orgesleka/used-cars-database

**Describe**: There are 26 attributes in the database, like Date Crawled, "name" of the car, seller, offer Type, the price on the ad to sell the car, vehicle type, year of Registration, Gearbox is manual or automatic or null, power PS model, kilometers the car has driven, month Of Registration, fuel Type, brand, not Repaired Damage, the date for which the ad at e-bay was created, number of pictures in the ad, postal Code, last Seen Online.

1. **Germany cities demographic information**

**Resource**:<https://www.destatis.de/DE/Publikationen/Thematisch/Preise/Verbraucherpreise/VerbraucherpreisindexJahresberichtPDF_5611104.pdf?__blob=publicationFile>

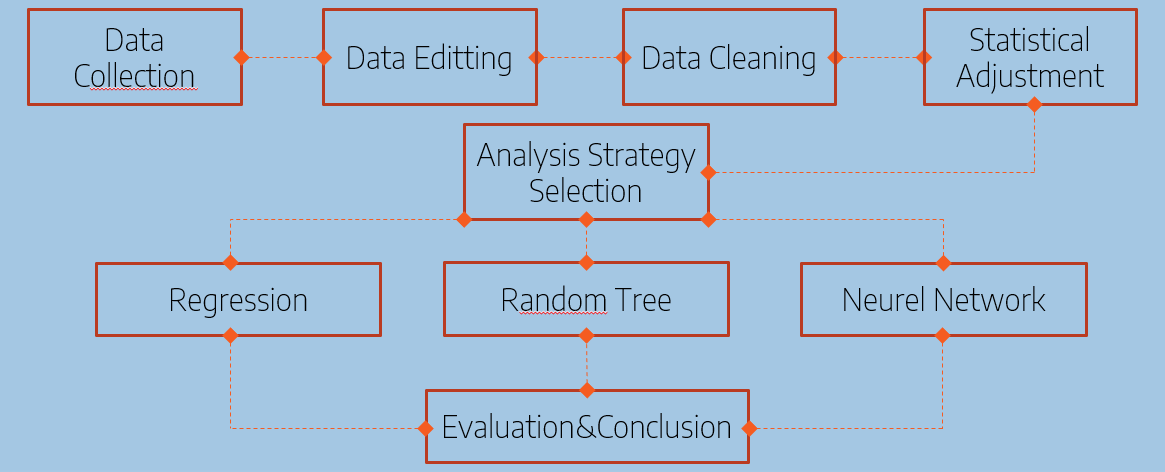
**Describe**: There are 12 attributes in the database: PostalCode, PlaceName ,State ,City, Latitude ,Longtitude ,Population ,Area ,NOofInhabitants

**Problem Statement**

1. Find out the most important attributes that effect the price of the used car, including demographic attributes
2. Find out the influence of “time sold out” on the price of used carHow CPI (Consumer Price Index) in different cities influences Used Car Market
3. Find out the correlation coefficient for each attribute and built a used car price predict model

Methodology

We will use SPSS and Python to find some rules and associations between attributes.



1. **Data collection**

Our project explores Kaggle’s ‘Used Car Database’ scraped from Ebay-Kleinanzeigen (in German) and merge a dataset which contains demographic information.

1. **Data Editting**

Reduced the number of variables

we reduce the number of variables from 36 to 25 to explore the dataset. The variables have no significant impact on explaining the data are eliminated.

1. **Data Cleaning**

We will clean the dataset by forming upper and lower limits, Removing Empty Cells and setting the blank rows as NA.

1. **Statistical Adjustment**

We refine and Standardize the variables which include date and time. Eg, ‘Date Crawled’, ‘Date Created’, ‘Last Seen’ would be transferred to standard data type.

1. **Analysis strategy selection**

We choose three models in total including random tree, neural network and regression.

We use ‘time-to-sale’ as a potential predictor for price. we had some notion of which cars sold in a "normal" time, which sold "fast" and which sold "slow." We calculate the analytical with Gaussian weighting function:

𝑊𝑒𝑖𝑔ℎ𝑡 = 〖𝐸𝑥𝑝(−(((𝐷𝑎𝑦\_𝑓𝑜𝑟\_𝑠𝑒𝑙𝑙\_𝑀𝑒𝑎𝑛(𝐷𝑎𝑦\_𝑓𝑜𝑟\_𝑆𝑒𝑙𝑙))/𝑆𝐷𝐸𝑉(𝐷𝑎𝑦\_𝑓𝑜𝑟\_𝑠𝑒𝑙𝑙))〗^2)/2

We ignore the time-to-sale then include it is to see what changes. That is, for each major model we use, we have one version where time\_to\_sell has role Input and one version where time\_to\_sell has role None. For each major model you use time\_to\_sale to determine analytical weight, have one version where the weight is enabled and one version where the weight is disabled.

we compare the accuracy of different models and get the conclusion.

1. **Data analysing**

We will use SPSS and Python accordingly to analyze the dataset. Two dataset would be merged together by zipcode to multi aspect analyzing.

1. **variable selection**
2. Variables Keep

we will keep 25 attributes which we consider have impacts on price. These attributes include:

Price, VehicleType, YearofRegistration, GearBox, PowerPS, Model, Kilometer, MonthofRegistration, FuelType, Brand, NotRepairedDamage ,DateCreated, PostalCode, DateLastSeen, DayforSale, Weight, PostalCode, PlaceName ,State, City, Latitude, Longtitude, Population, Area, NOofInhabitants

1. Variables New

“Age” Variable will be Created by Using Year of Registration and “Day for sale” Variable will be created by using Date Created and Last Seen for analyzing convenience.

1. **Model Building**
2. Regression analysis: regression model

We use regression analysis to estimate the relationship between price and other featured variables. We want to understand how the typical value of the Price changes when any one of the independent variables is varied, while the other independent variables are held fixed. This would help us to predict the price of any used car based on specific values of attributes.

1. Relevance Analysis: neural network model & random tree model

Attribute relevance analysis is an addition important part of analysis in our project which operates with target variable. We use relevance analysis to see if there is a relevance between the price of the used car and city’s demographic information.

1. **Evaluation**

We will divide the dataset into two subsets:

1. Training set is a subset of the dataset used to build predictive models.
2. Validation set is a subset of the dataset used to assess the performance of model built in the training phase.

RESULTS AND DISCUSISON

**Regression Model**

**1. Model Building**

In order to create a price prediction model for both car sellers and buyers, we build two different regression models to find the correlation between car price and other featured variables.

The first regression model targets to car buyers. ‘Year Of Registration’; ‘gearbox’; ‘powerPS’ ; ‘kilometer’; ‘brand’; ‘notRepariedDamage’ are set as inputs and ‘car price’ is set as target. Also, the day for sales variable is used to calculate the weight according to the Gaussian weight function. By applying the weight function to our regression model, we adjust the price to its real market value. Then, we run the regression model in the SPSS modeler.

For the second regression model, its potential users are car sellers. Besides ‘Year Of Registration’; ‘gearbox’; ‘powerPS’ ; ‘kilometer’; ‘brand’; ‘notRepariedDamage’, we add one variable’, ‘day for sale’ as inputs and set ‘car price’ as target variable. Meanwhile, the weight function is removed. Finally, we run the regression model in the SPSS modeler

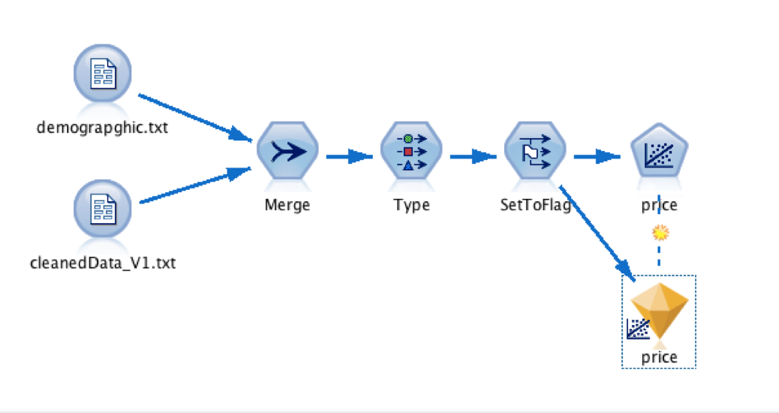


Figure1.Stream of Regression Model

**2. Result**

For the first regression model for car purchasers, the result shows that the top important predictors are ‘brand’, ‘kilometer’ and ‘yearOfRegression’. Also, the coefficients generated can be used to predict real market price for car buyers.

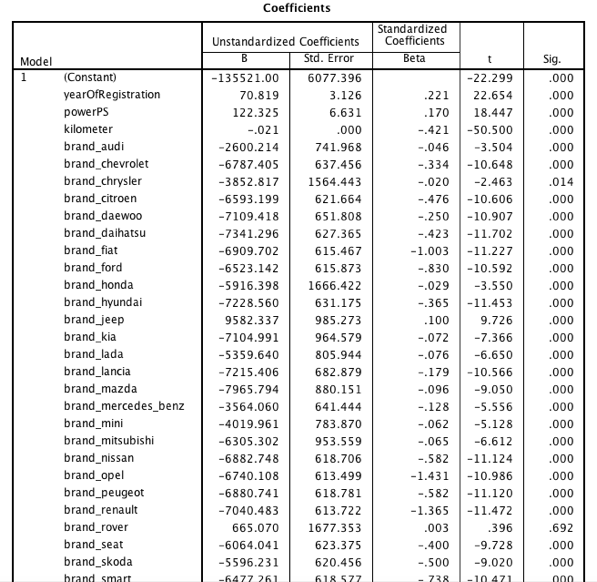
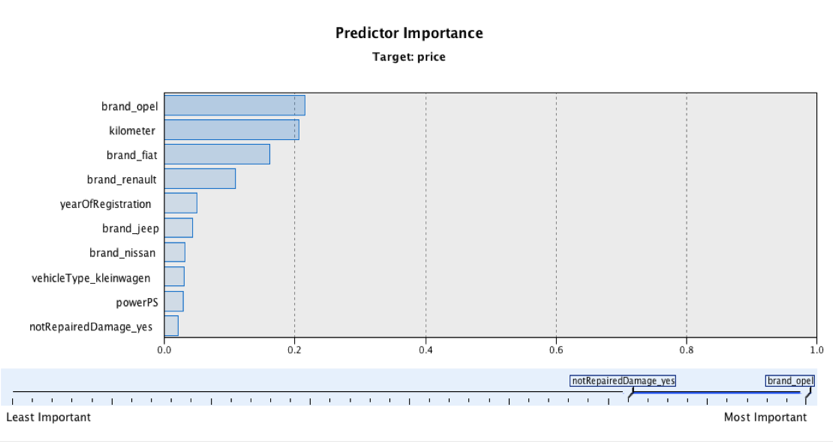


Figure2.Result of Regression Model 1

For the second regression model for car sellers, the top important predictors are the same, proving that these three variables have a great influence on the car price. Also, the coefficients generated can be used to predict a reasonable price for sellers when they assign an expected sale time and other car attributes in our model.

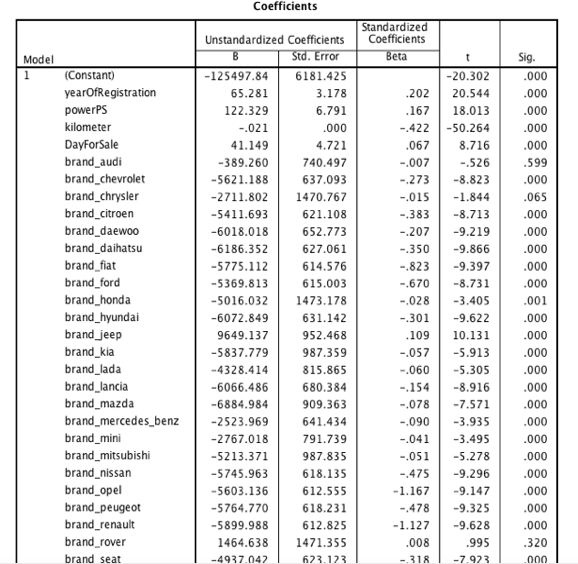
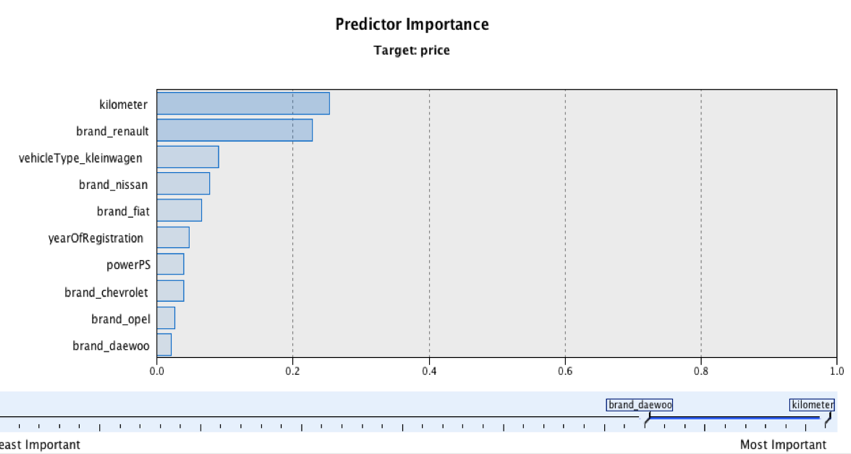


Figure3.Result of Regression Model 2

**3. Conclusion & Insights**

By using the price prediction regression model, both seller and buyer can gain accurate price suggestions which improve the overall business efficiency. One a change is made on one or several attributes, a price gap would be generated which is a significant price information for car sellers and buyers to adjust their strategy in order to optimize their benefits. Generally, the market can become more and more transparent and healthy with the open price information.