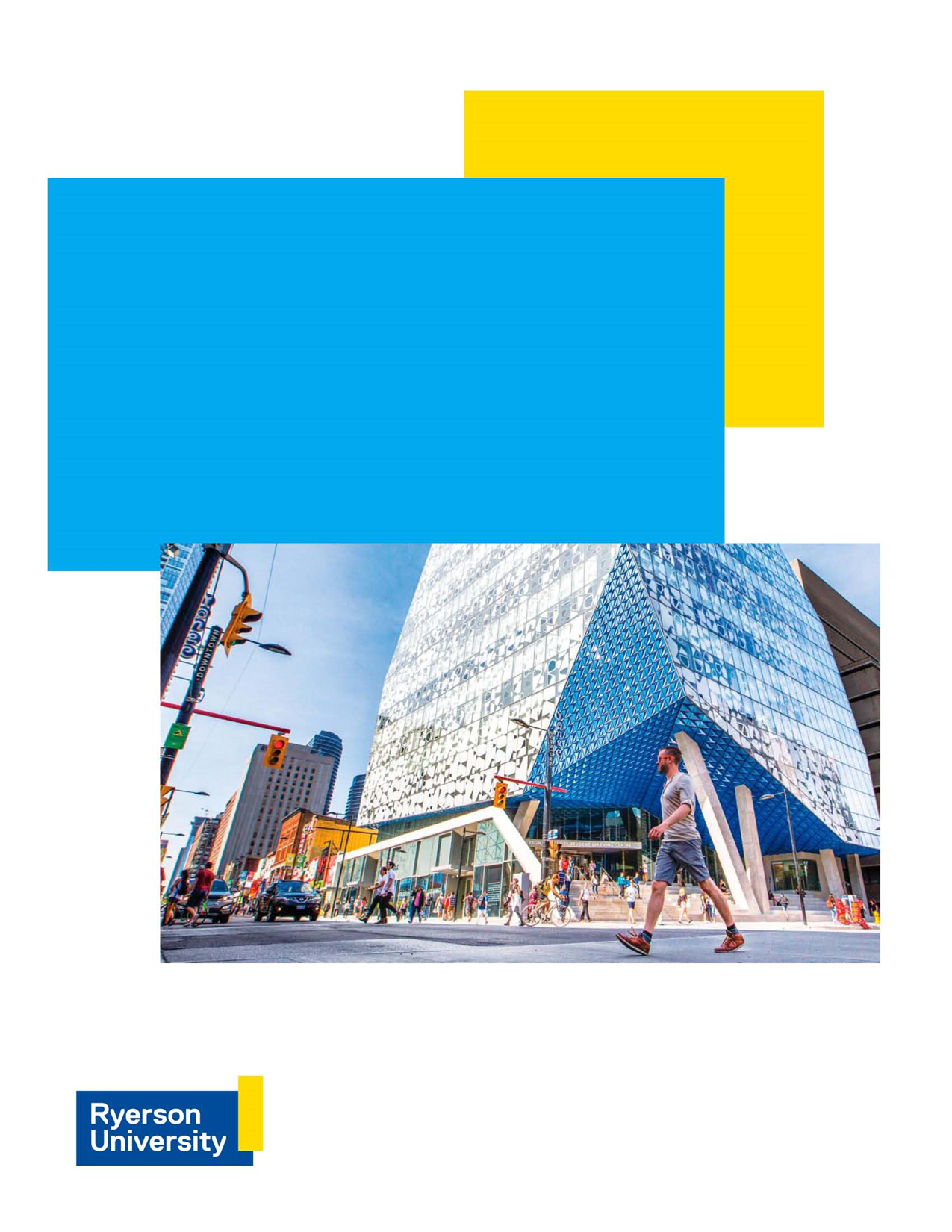
****

CIND 820 Advance Data Analytcs Project

Machine Learning (ML) Techniques To Predcit

Price of Cars on Kijiji Dataset

**Advisor:DR.** Ceni Babaoglu

**Prepared by:** Yoseph Abdi

**Student ID**:501147225

April 3, 2023

**Contents**

**Part One:- Introduction and background of the study .....**……………………………………….**2 –6**

* 1. Introduction
  2. Background of the study

1.3 Problem statement

**Part Two: - Literature review and methodology**……………………………………….……………**7-15**

2.1 Literature review

2.2 Methodology

2.2.1 Dataset Description

2.2.2 Correlation, Duplicates, Imbalance and Missing Values

2.2.3 Data Preparation

**Part Three: Machine learning regression model…………………………….…………………15-18**

3.1 Linear regression model

3.2 Decision tree mode

3.3 Random Forest model

**Part Four**…………………………………………………………………………………………………………..18-20

4.1 Conclusion

4. 2 Reference

**Part One**

**Introduction and background of the study**

**1.1 Introduction**

The prices of new cars in the industry are 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. Due to the increasing price of new cars customers are shifting to buy used cars. Therefore, having appropriate information about the price of used cars is relevant for the buyers and sellers [1].

In this research paper Kijiji used cars dataset is used([Kijiji Car Sales Ads | Kaggle](https://www.kaggle.com/datasets/hossaingh/kijiji-car-sales-ads)), which I found from Kaggle public open data source where I have account and authorized to use. Kijiji in one the famous web site in Canada to buy and sale used items.

This research paper contains background of the study, problem statements, literature review, methodology. Machine learning, linear regression, decision tree and random forest regression.

**1.2 Background of the study**

Kijiji car sales dataset is publicly available at [**Kijiji Car Sales Ads | Kaggle**](https://www.kaggle.com/datasets/hossaingh/kijiji-car-sales-ads).[8] [Kijiji Car Sales Ads | Kaggle](https://www.kaggle.com/datasets/hossaingh/kijiji-car-sales-ads). The dataset contains 51343 observations and twenty-eight attributes. The attributes are both categorical and numerical, this study used linear regression, decision tree and random forest methods to predict the price. The dataset contains categorical and numerical data, for safety and privacy purpose locations and VIN attributes are excluded. The study used fifteen attributes of the 28 these are city, price, sale by, make, year, condition, kilometers, body, fuel type, color, visit, no of doors, no of seats, transmission and sunroof. It has missing values of maximum of 77% sunroof and minimum 1.50 % of visit. It has different data types floating, int, object data types.**.**



***Table 1: Kijiji Used Car dataset attributes***

**1.3 Problem Statement**

The main objective of the research paper is to predict the price of used cars in Kijiji website. And to help buyers and sellers by providing appropriate information so that they can make the right decision. And, to develop hands on experience on big data analytics which covers all courses taken in Data Analytics, Big Data and Data Predictive program.

The three main objective of the study are -

* To predict the selling price of used cars and to know which car types are sold most using Kijiji dataset.
* To analysis the dataset using different machine learning techniques and to identify the most relevant attributes for the sales of used cars in Kijiji.
* To examine the relationship of the dependent attributes to the other independent attributes and to recommend Kijiji Car sales to improve their sales.

**Chapter Two**

**2. Literature review and methodology**

**2.1. Literature review**

Many research paper predicting the price of a products based on their features. Hossain Ghahraloud has used the same dataset to analyze car sales data on Kijiji website [9]. However, Hassain used the dataset just did exploratory analysis. He didn’t make any price prediction and didn’t used any machine leaning models. His aim was to analyze the data based on locations mainly in Toronto, Calgary, Ottawa, and Vancouver regions.

Chejarla Venkat Narayana, Chinta Lakshmin Likhitha, Syed Bademiya, Karre Kusumanjali have done a machine learning technique to predict the price of used cars [10]. The major drawback of their paper is the data they used. The data has only limited number if attribute of 4000 and the research is focused used car sales in India.

Enes Gokce, has done a research paper on predicting used car price with Machine Learning Techniques [11]. Generally, he has done an excellent work and used huge dataset. However, there are some drawbacks from 550313 rows Enes only used 380962 which is 69% of the dataset. Whereas 31% of the dataset is considered missing values and cleaned. The other drawbacks I found in this research is that Enes used ffill method to fill the missing values, ffill method is not the best method to fill missing values of numerical data types.

The other research paper I reviewed is Nivitus. The research title is car price prediction using Machine Leaning [12]. The main drawback of this paper is the dataset used. Nivitus used only 301 records and 9 attributed, which is too small to make appropriate car price prediction. Moreover, the researcher mentioned there is no missing or null values which lucks to show data cleaning methods.

Ravi Shastri, Dr A Rengarajan, is the final research paper which I reviewed. The research paper is titled prediction of car price using linear regression [13]. The dataset for this study was taken from e-commerce site called Quirk and processed using python programming language. The main shortcoming of the research is that it has only 240 row and 8 attributes. Which is too small to make the appropriate price prediction [14].

**2.2. Methodology**

To build a car predication model, the dataset is collected from Kaggle open-source website, which I have an account and authorized to use. To analysis the dataset Jupyter notebook is used. The data is imported with python pandas and then some private and locational columns were dropped. Exploratory analysis was made to get the summary of the dataset statistics using profile. To get information about the number of attributes and variables, interactions, correlations, the duplicate data, missing values, and sample. The dataset has 51343 observations and 28 attributes.

Then, preprocessing will be made tasks to handle missing data, categorical variables, and scaling of features. And then feature engineering tasks such as handling the outliers, dividing the dataset into train, test and validation splits, and analyzing the importance of feature in building the model. At last linear regression model, Decision tree and Random Forest machine leaning algorithms are used and evaluate the accuracy test.

**2.2.1 Dataset Description**

After dropping the locations, private attributes, and color and fuel type which are to correlate to price the research is made on 13 attributes. The attributes are city, price, sale by, year, condition kilometers, body, visit, doors, seats, transmission, and sunroof. See Figure 1 dataset head. The data presented in price, year, kilometers, door, seats, sunroof are numeric, and the others are categorical.

1. **City:** The categorical attribute, indicates the name of the cities where the car is found. The dataset has 12 unique cities.
2. **Title:** It represents the name of car in the dataset for sale.
3. **Price:** it is the dependent variable which going to be predicated. The dataset has a price range from mean from 37,100 to 104,468.00.
4. **Sale by:** It shows by who the car is soled. The two sales by are owner and dealer.
5. **Make:** It shows as who make the car. There are 50 car manufactures in the dataset such as Ford, Ram, Toyota, Jeep, Chevrolet.
6. **Year:** It represent the year the car was manufactured. It ranges from1900 to 2023.
7. **Condition:** The status of the car wither it is new, used, salvage, damaged and lease takeover.
8. **Kilometers:** The number of kilometers that a car has driven so far. The dataset covers a range of 0 to 278,000.
9. **Body:** The body type of the car like SUV crossover, sedan, pickup truck. There are 9 types of body of the car in the dataset.
10. **Fuel Type:** It represent the type of fuel the car use such as diesel, gas, electric, hybrid and other.
11. **Visit:** The number of times the car is visited by the buyer.
12. **Doors:** It shows the number of doors the car has, a minimum of 2 doors and a max of 6 doors are in the dataset.
13. **Seats:** the number of seats the car has, the dataset seats ranges from 1 to 44.
14. **Transmission:** -It shows the type of car, like manual or automatic.

***Graphical user interface, text, application

Description automatically generated***

***Figure 1. The Pandas’ datasets describe.***

As per the below dataset statistics the dataset has 86929 or (11.30%) of missing cells and 258 or (0.50%) duplicate rows as shown in Table 2 below.

**Table 2: Dataset Statistics**

*Source: Kijiji Car sales profile report generated from pandas’ profile Report.*

**2.2.2 Exploratory Analysis**

Exploratory Data analysis (EDA) is used to understand the data to gather all possible insights, perform initial investigations, discover anomalies and patterns, test the hypothesis by using various statistics and visualization techniques.

The dependent variable price is highly correlated with the independent variables fuel type, color, doors and transmission. There is a duplication of 0.5%, imbalance and missing. In addition to the pandas profiling, I will use other seaborn function to check the visualization and distribution. Zero values in price, kilometers and visit attributes will be replaced with mean values. Price is highly skewed (y1=224.9220245) and Kilometers highly skewed with (y1=31.39418225). ***Figure 2 shows Alerts of the dataset.***

Chart

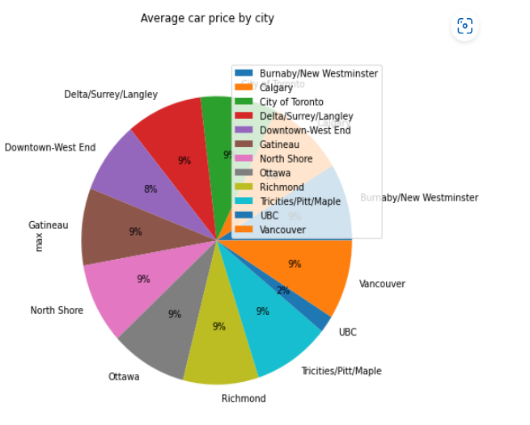
Description automatically generated

Chart, bar chart, histogram

Description automatically generated Chart, bar chart, histogram

Description automatically generated

Chart, pie chart

Description automatically generated



**Table 3: Correlation and alerts of attributes**

**2.2.3 Data Preparation**

It is the most important step in processing of predication. In addition to pandas profiling, different functions used to clean, analysing each attribute. Missing values and duplicates are identified.

**Missing Values:** Imputation method is used to replace the missing values for numeric data types. After analyzing the statistics central tendency (mean) is used to fill the missing values. As sunroof attribute has the highest percentage of missing values 77.19% or (39633) records it dropped from the study. Table 2 below shows the percentage of missing values.

**Table 4: Percentage of missing values**

Table 4: Kijiji car sales missing value percentage generated using Jupiter pandas’ function.

**Outliers: -**The dataset has an outlier of numerical variables and analyzing the outlier’s quantile methods of 25% and 75% is used to remove the upper and lower outliers. Boxplots graph is used to check the outliers.

**Encoding categorical features:** To hand the categorical datatypeOne hot encoder is used on each categorical attribute which converts each categorical value to 0 and 1.

**Feature Scaling: -**Huge data value can dominate the fields of small values and many hide values insights. Scale or normalize the data present in the fields so that every filed can contribute equally to build the model.

**Splitting the dataset:** The dataset is split in to test, train and cross-validation such that 60% train, 20% test, and 20% validation.

**Model Building:** Linear regression model is used to predict the target variable. Decision tree technique and Random Forestis used to access the result.

**Packages/library used:** The famous Jupyter notebook, pandas sklearn and other packages listed in table 4 used to analyze the dataset.

**Table 5: Library packages**



**Part three**

**Machine Learning Regression Model**

**3.1 Linear Regression model:** It helps to predicate the value of dependent variable based on given collection of independent features. Linear regression shows the linear relationship between a target variable “Y and group of inputs features like X1, X2, X3………….Xn.

The relationship between predictor variables and target variable is described by regression coefficients. The direction of association between a predictor variable and target variable is shown by the sigh of each coefficient. A positive sigh means that as the predictor variable rises, the target variable also rises. A negative sign means that the target variable drops as the predicator variables rises.

Figure 1 shows the linear regression R 2 score of 1 and mean absolute error of 3.7265. Which mean the model predict the price with best accuracy. And the most car sold is SUV crossover is the top car sold.

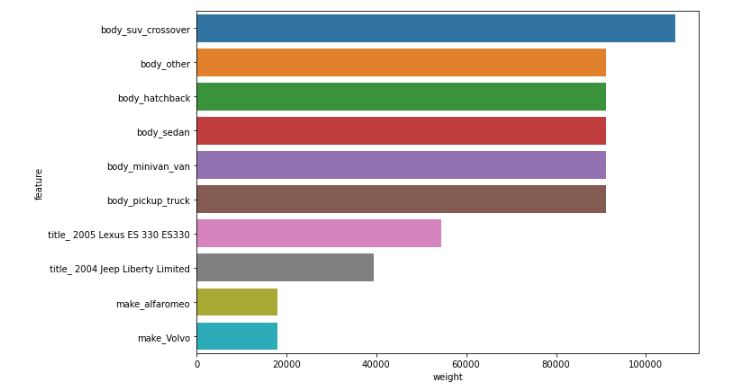


Figure 3, Linear Regression model; source from coding

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**3.2 Decision Tree Regression model: -** This technique uses a decision tree to go from the observations to final decisions. In these tree structure, leaf nodes represent the targets class and labels or the values to be predicated and intermediate nodes represent combination of feature that led to final decisions or values.

The decision tree model has the best predicted of R square result 0.9986 and mean absolute error of 697.665 which is almost similar with other models.

***Figure 2****: Decision Tree model used to predict Kijiji car price.*

Graphical user interface, text, application, email

Description automatically generated

**3.3 Random Forest Regression Model: -** It is an ensemble learning technique, that does a classification, regression or other tasks by creating a large number of decision tree during training and output the values or decision by considering all those trees. As shown in figure 3, Random Forest regression model predict a best value of 0.999 R square value, which is similar with Decision tree and linear models.

**Graphical user interface, text, application

Description automatically generated**

***Figure 4****: Random Forest regression model, source Kijiji car price prediction code.*

**Part IV. Conclusion and Reference**

1. **Conclusion**

It is understood that building predication model is challenging task. It involves domain knowledge, data analysis, machine learning techniques. In this study I followed different methods to predict the price of cars price in Kijiji dataset. Data exploratory analysis has done to understand the contain and over all relation of the attributes. And different data preprocessing techniques, missing values, duplicates, removing of outliers, splitting the dataset to train, test and validation has been made before the modeling. Also encoding categorical variables, scaling of features, and finally linear regression, Decision three and Random Forest modeling were implemented. Even if all the three-model predicated similar output. In this case linear regression model out smart even Random Forest model and predicated 1:00 R squared error and 1:00 mean absolute error on train dataset. Therefore, linear regression model produced best car price predictor in this study.

**References**

1. Pal, Arora and Palakurthy, 2018.
2. [what is big data? Characteristics, types, and technologies | by The Educative Team | Dev Learning Daily (medium.com)](https://medium.com/educative/what-is-big-data-characteristics-types-and-technologies-67dc9f958659)
3. [What is Machine Learning? | IBM](https://www.ibm.com/topics/machine-learning)).
4. [REGRESSION ANALYSIS. In statistical modeling, regression… | by Sameer Maurya | Medium](https://mauryasameer.medium.com/regression-analysis-def876d5904b)
5. [Different Regression Analysis Models in Machine Learning | by Dr Roushanak Rahmat, PhD | Medium](https://medium.com/@roushanakrahmat/different-regression-analysis-models-in-machine-learning-e7eca835a5ee)
6. [Decision tree learning - WikiMili, The Best Wikipedia Reader](https://wikimili.com/en/Decision_tree_learning)
7. [Predictive Analysis in Data Mining - GeeksforGeeks](https://www.geeksforgeeks.org/predictive-analysis-in-data-mining/?ref=gcse)
8. [Kijiji Car Sales Ads | Kaggle](https://www.kaggle.com/datasets/hossaingh/kijiji-car-sales-ads)
9. [Analyzing car sales data on the Kijiji website | LinkedIn](https://www.linkedin.com/pulse/analyzing-car-sales-data-kijiji-website-hossain-ghahraloud/)
10. [Machine Learning Techniques to Predict the Price of Used Cars: Predictive Analytics in Retail Business | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9532845)
11. [Predicting Used Car Prices with Machine Learning Techniques | by Enes Gokce | Towards Data Science](https://towardsdatascience.com/predicting-used-car-prices-with-machine-learning-techniques-8a9d8313952)
12. [Car Price Prediction Using Machine Learning | by Nivitus | Medium](https://medium.com/@Nivitus./car-price-prediction-using-machine-learning-6d7678cdbf94)
13. Shastri,extension://elhekieabhbkpmcefcoobjddigjcaadp/https://www.ijtsrd.com/papers/ijtsrd42421.pdf