Federal Urdu University
OF, ARTS, SCIENCE AND TECHNOLOGY

Data Mining

Term Project Report

Group Members

Name: Faraz Jahangir
Father Name: Jahangir Alam
Seat No: 19122019

Name: Ezan Adil
Father Name: Adil Hussain
Seat No: 19122190

Semester: 6th (A)
Department: Computer Science
Submitted to: Miss Uzma
Topic: Suzuki Used Cars Price Prediction
Date: 28-Dec-2023

Objective

The objective of this project is to develop a predictive model for suzuki used car pricing. The targeted models are Mehran, Alto, Swift, Cultus, WagonR.

The primary goal is to create a model that accurately estimates the price of a used car based on various features such as.

1. Model
2. Year
3. Mileage
4. Fuel Type
5. Transmission
6. Color
7. Assembly
8. Engine Capacity
9. City
10. Price

Introduction and Background

Used car pricing is a complex task influenced by various factors such as model, year, mileage etc. This project employs machine learning, combining supervised and unsupervised techniques, to enhance accuracy. By leveraging historical data, the model aims to provide reliable estimates, catering to the dynamic nature of the used car market.

Data Collection

The dataset for this project was gathered from PakWheels using web scraping techniques. PakWheels is a reputable online platform that provides comprehensive information about used cars, including their specifications, prices, and other relevant details.

Data preprocessing

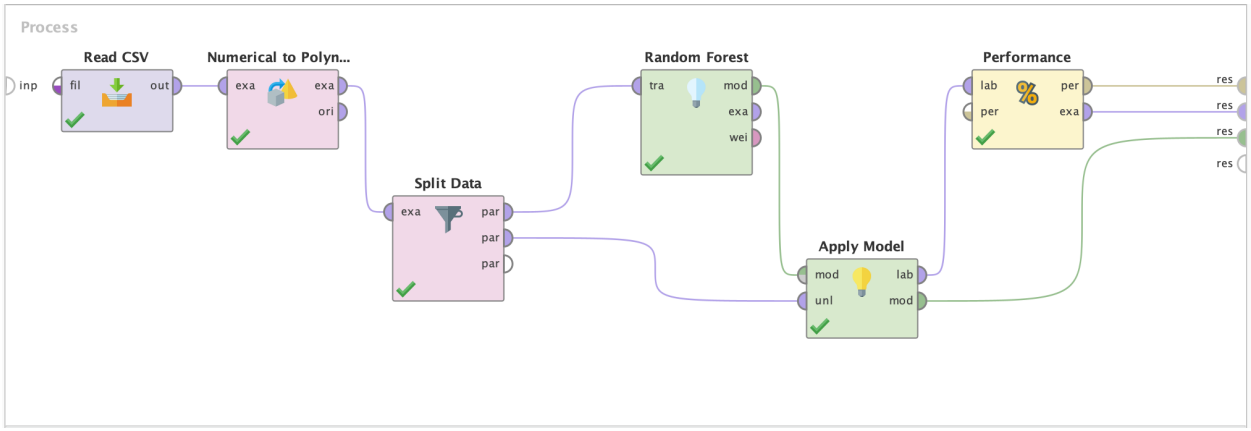
The dataset underwent thorough preprocessing using both Excel and SPSS.

- **Excel**
  - Used to fill missing values, ensuring a complete dataset.
  - Defined specific ranges for price and mileage.
- **SPSS**
  - Utilized to eliminate duplicates, ensuring the dataset's integrity.
  - Visualized and inspected values to identify and handle any out-of-domain entries.

Modelling

Two primary algorithms, Random Forest and K-Means clustering, were employed to develop and assess the predictive model.

- **Random Forest**
  Utilized the Random Forest algorithm to create a model for predicting used car prices. This ensemble learning method leverages multiple decision trees to enhance prediction accuracy.The dataset was split into training and testing sets to train the model and evaluate its performance accurately.



Design

Table View Plot View

accuracy: 84.14%

	true 1850000 – 2...	true 3530000 – 4...	true 2690000 – 3...	true 1010000 – 1...	true 1700000 – 1...	true 4370000 – 5...	class precision
pred. 1850000 – ...	481	0	40	25	0	1	87.93%
pred. 3530000 – ...	1	21	4	0	0	2	75.00%
pred. 2690000 – ...	43	12	138	0	0	0	71.50%
pred. 1010000 – ...	19	0	1	350	25	0	88.61%
pred. 1700000 – ...	0	0	0	19	34	0	64.15%
pred. 4370000 – ...	0	1	0	0	0	0	0.00%
class recall	88.42%	61.76%	75.41%	88.83%	57.63%	0.00%	

Performance Metrics

- **K-Means Clustering**
  Implemented the K-Means clustering algorithm to identify patterns and group used cars based on similar features.Determined the optimal number of clusters to maximize the effectiveness of grouping. Used k value 4 and measure type is MixedMeasures

Design

Cluster Model

```
Cluster 0: 1607 items
Cluster 1: 1085 items
Cluster 2: 523 items
Cluster 3: 845 items
Total number of items: 4060
```

Clusters

Conclusion

The Random Forest model exhibits a commendable accuracy of 84%, showcasing its proficiency in predicting used car prices.The model's predictions are based on influential features such as, model, year, mileage, etc. Contributing to an in-depth understanding of pricing determinants.

The K-Means clustering analysis unveiled four distinct clusters in the dataset, offering valuable segmentation insights. Here's a simplified breakdown:

- Cluster 0: 1607 items
- Cluster 1: 1085 items
- Cluster 2: 523 items
- Cluster 3: 845 items

In summary, the project effectively used machine learning to uncover valuable insights in the used car dataset. The high accuracy of the predictive models holds potential for real-world applications in the dynamic used car market. The segmentation from

K-Means clustering enhances strategic decision-making, and the Random Forest model proves valuable for accurately estimating individual car prices.

The dataset can be find on below github repository:
https://github.com/Farazjahangir/Suzuki-Used-Car-Dataset/tree/main