PREDICTING CAR PRICES BASED ON SECONDARY MARKET DATA IN ARMENIA

A Undergraduate Project Report

Presented to the

Zaven & Sonia Akian College of Science and Engineering

American University of Armenia

Yerevan, Armenia

In Fulfillment

of the Requirements for the Degree

BACHELOR OF ARTS IN DATA SCIENCE

in the

Zaven & Sona Akian College of Science and Engineering

By

LEVON A. YESAYAN

Supervisor

ARMAN ASRYAN

May 9, 2024

TABLE OF CONTENTS

ABSTRACT

The main objective of this capstone work is to create and implement a car price prediction system based on the analysis of information from the leading site for selling cars in Armenia. The project's problem lies in the high volatility of prices, and data is non-standardized, making it hard to buy and sell cars in Armenia. We propose an approach based on machine learning techniques for analyzing and processing collected data. To achieve this, wo pipelines were developed and tested for the study: one without car images and another using the Resnet18 model, deep learning technology for photo analysis. Interestingly, models trained with image features yielded similar accuracy levels on predictions. Henceforth, an online platform was built on Flask with features enabling users to evaluate their vehicle worth, such as an interactive user interface for entering car attributes and getting price estimates and a statistical graphics page including market analytics. This project contributes to improving the transparency and efficiency of the Armenian car market and opens up new directions for further research in machine learning in the automotive industry.

Keywords: Car Price Forecasting System, Armenian automobiles, Deep learning technologies, Photo Analysis

Github repository link: Capstone/Car Price Prediction

Google Drive archived data link: DATA

Chapter 1

INTRODUCTION

The modern car market in Armenia is highly dynamic and significantly price volatile, creating severe difficulties for buyers and sellers. Because of the absence of unified and accessible databases that could provide up-to-date information on car costs, the buying and selling process becomes not only less efficient but also less transparent. Research on the applications of machine learning methods used for analyzing car markets shows significant potential in optimizing and automating cost estimation processes; however, such technologies still need to be improved for use in the Armenian market.

This project aims to develop and test car price forecasting models based on a complex analysis of data from an Armenian automobile sales portal. Special attention is paid to the approach of analyzing ad images using convolutional neural networks. The application of this approach allows not only to increase the accuracy of predictions but also to better understand the influence of cars' visual aspects on their market value.

The main research question is whether integrating car images into the data processing process will significantly increase the accuracy of car price forecasts. The research hypothesis suggests that models that include analysis of visual characteristics and traditional car parameters can demonstrate higher efficiency in price prediction.

This project aims to improve understanding of price formation mechanisms in the Armenian automobile market and provide tools for more informed decision-making for private individuals and businesses. The project results will contribute to optimizing the buying and selling process in the domestic market.

DATA

This work used data collected from the Auto.am website [4] is one of the leading Armenian Internet resources for selling cars. Data collection was carried out by web scraping using the Selenium tool, which enabled automating the process of extracting information about vehicles for sale. This data was then structured in tabular form and included many attributes, each describing specific car characteristics.

The dataset contains the following key attributes for each car:

- Listing ID: Unique identifier for each ad.
- Make and Model: Vehicle make and model.
- Year: Manufacturing year.
- Price: The listing price of the vehicle.
- Mileage: Kilometers driven.
- Body type: Shape of the car.
- Fuel Type: The type of fuel the vehicle uses.
- Engine displacement: Engine size in liters.
- Horsepower: Vehicle engine power.
- Cylinders: Number of cylinders in the engine.
- Wheel Drive: Vehicle drive type (front, rear, all-wheel drive).
- Transmission: Gearbox type (automatic, manual).
- Wheel Displacement: Wheel location (left-drive, right-drive)
- Color: Car color.

- Listing Date: The date the ad was posted on the site.
- ListingLink: The URL of the ad page on the site.
- Image Features: "image_feature_1" to" image_feature_15" are visual features extracted and quantified using CNN.

The provided data is examined to evaluate Armenia's used automobile market dynamics and current trends. Particular attention is paid to studying the relationships between price offers and car technical parameters, which can contribute to more accurate forecasting of price fluctuations and consumer preferences.

Exploratory data analysis

The summary statistics provide insights into various aspects of the cars in the dataset (Figure 1). Here are some key observations:

1. **Year**: The cars range from 1978 to 2023, with a median year of 2017, indicating that the dataset predominantly contains relatively newer cars.
2. **Price ($)**: Prices range from $1,500 to $58,500 with a mean of $21,750.
3. **Mileage (km)**: Mileage ranges from 10 km to 470,000 km, with a median of 100,000 km..
4. **Engine Size**: Engine sizes range from 0.2 to 6.2 liters, with a mean size of 2.58 liters, indicating a diverse set of engine capacities.
5. **Horsepower (HP)**: Horsepower varies significantly from 70 to 600 HP, with a median of 188 HP.
6. **Cylinders**: Most cars have between 1 and 16 cylinders, with a median of 4 cylinders.
7. **Image features**: (image_feature_1 to image_feature_15) have a range of values between 0 and .

	Year	Price ($)	Mileage (km)	Engine Size	HP	Cylinders	image_feature_1	image_feature_2	image_feature_3	image_feature_4	...
count	672.0	672.000000	672.000000	672.000000	672.0	672.0	672.000000	672.000000	672.000000	672.000000	...
mean	2015.028274	21750.985119	118132.443452	2.581696	214.858631	4.797619	0.315755	0.314654	0.294619	0.347164	...
std	6.163261	12175.301612	84895.929126	0.887254	79.041101	1.489394	0.287496	0.282979	0.264013	0.313964	...
min	1978.0	1500.000000	10.000000	0.200000	70.0	1.0	0.000000	0.000000	0.000000	0.000000	...
25%	2011.0	13000.000000	51874.500000	2.000000	166.0	4.0	0.087857	0.088535	0.086638	0.074909	...
50%	2017.0	18400.000000	100000.000000	2.500000	188.0	4.0	0.240192	0.242664	0.240605	0.274975	...
75%	2019.0	28550.000000	165000.000000	3.000000	252.0	6.0	0.473407	0.477796	0.421913	0.556666	...
max	2023.0	58500.000000	470000.000000	6.200000	600.0	16.0	1.000000	1.000000	1.000000	1.000000	...

8 rows x 21 columns

Figure 1: Summary statistics

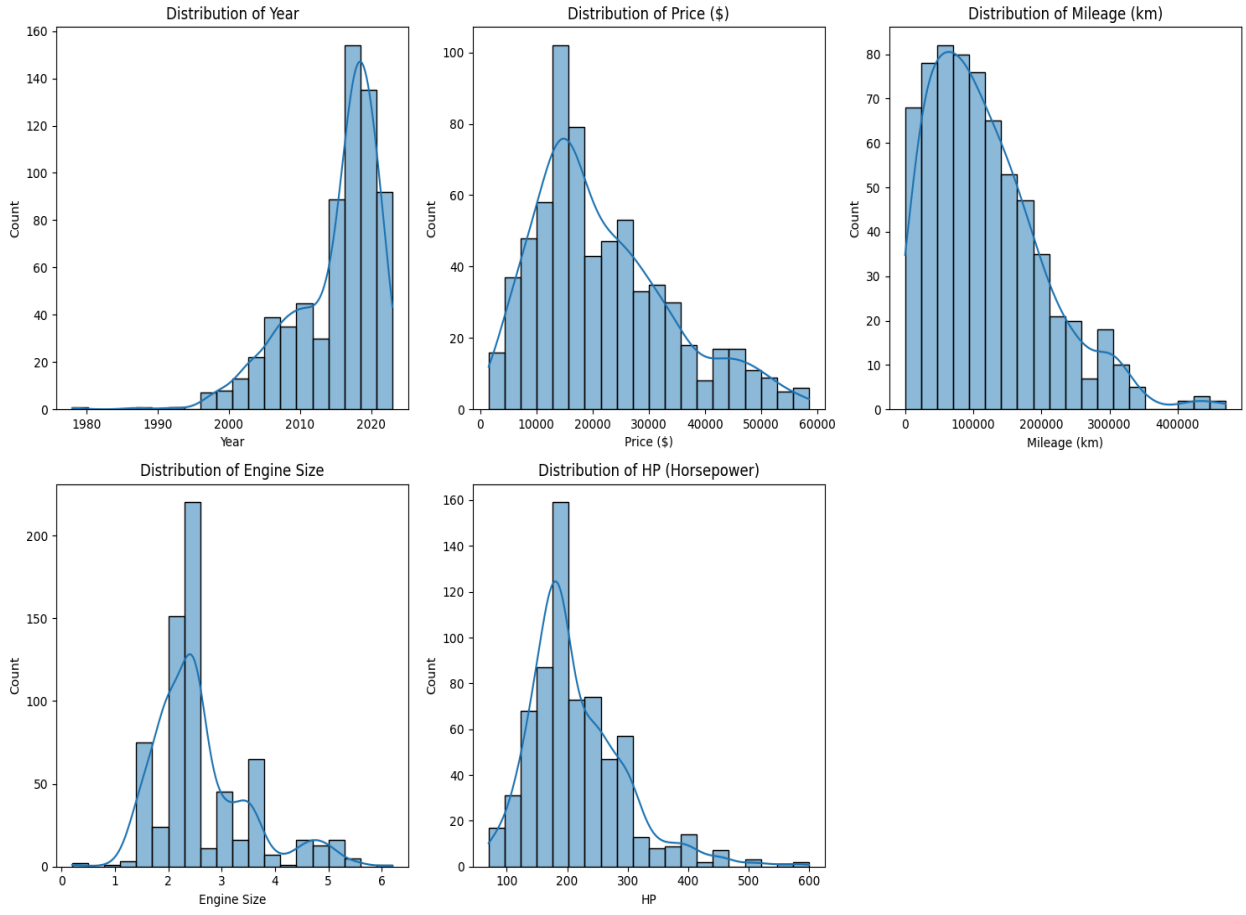

Figure 2: Distribution of numerical variables

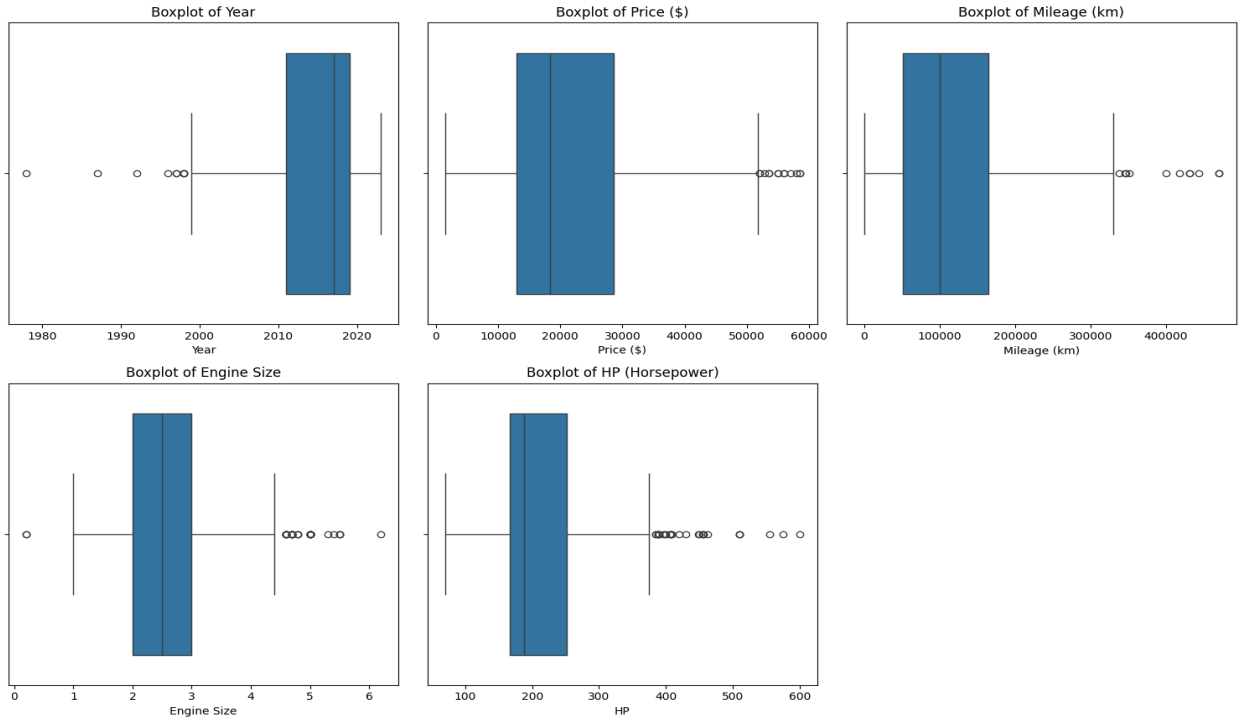

Figure 3: Boxplots of numerical variables

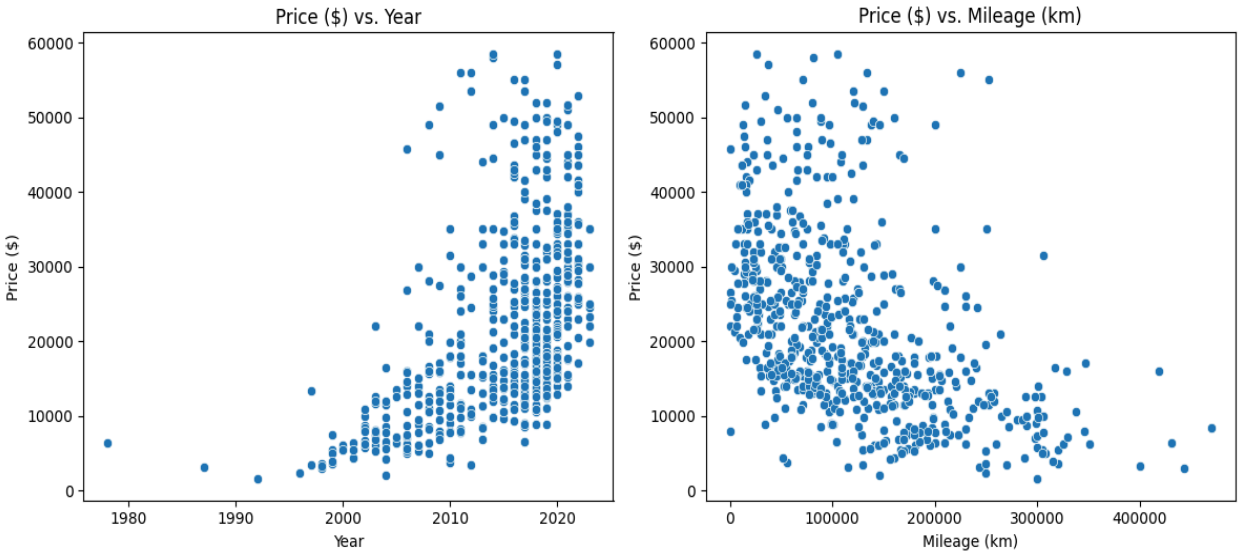

Figure 4: Price vs Year and Mileage

Histograms (Figure 2)

- **Year**: The distribution shows a concentration of cars from recent years, indicating a prevalence of relatively new cars in the dataset.
- **Price ($)**: Prices are right-skewed, meaning that while the majority of cars fall within the lower to mid-range price bracket, there are a few high-priced outliers.
- **Mileage (km)**: Similar to price, mileage is right-skewed. Most cars have moderate mileage, but there are cars with very high mileage, indicating that some older or more heavily used cars are included.

Scatter Plots (Figure 4)

- **Price ($) vs. Year**: There's a positive trend indicating that newer cars tend to be priced higher, which is expected as newer models generally come with more advanced features and less wear and tear.
- **Price ($) vs. Mileage (km)**: As expected, there's a negative trend showing that cars with higher mileage tend to have lower prices. This trend reflects the depreciation of car value with increased usage.

Chapter 2

METHODS

1. Data Preparation

Data Processing

To ensure the accuracy and reliability of our analysis, a series of data processing steps were undertaken on the car characteristics dataset. This section outlines the methodologies employed for data cleaning, structuring, and handling missing values.

The raw dataset underwent a rigorous cleaning process to rectify inconsistencies or errors. This involved identifying and removing duplicate entries, correcting typographical errors, and standardizing data formats across all variables.

Furthermore, as part of the dataset contained prices in AMD and mileage in miles, a crucial step was taken to convert all prices to United States Dollars (USD) and miles to kilometers (km). This conversion ensures uniformity and facilitates meaningful comparisons across different car models.

To handle missing values, we employed a method known as "padding," where missing values were replaced with the most frequent value of the corresponding car's make, model, and year. So far, approximately 4,750 listings have been scrapped for analysis. After data processing and cleaning procedures, 727 clean observations were obtained for further study.

Image Processing and Feature Extraction

Processing car pictures is the second step in data cleansing. Each image is reduced to a standard size and format that satisfies the specifications of the ResNet model to guarantee the effectiveness of the procedure and the analysis that follows. This includes color channel balancing, image scaling to specific dimensions (such as 224 by 224 pixels [1, p. 2]), and transforming image data into a format that training models can understand.

Visual features are extracted from the processed listing images using the ResNet18 model. These features include details about the car's color, shape, textures, and other visual elements that can significantly impact its market value. ResNet extracts and encodes the characteristics into high-level abstractions, which enables its application for analytical models.

ResNet has a unique feature, residual blocks, which avoid the vanishing gradient problem in deep neural networks. In these blocks, input data is passed through one or more convolutional layers and then summed with the layer's output, allowing the original information signals to be transmitted over a greater distance across the network. This allows deep networks to train efficiently while maintaining high accuracy [2].

After visual features are extracted, feature selection is performed. As ResNet outputs multiple image features for each image, the most significant ones were selected from each image to create a necessary structure for the dataset for further ML model input. Afterward, they are combined with other collected data, such as make, model, year of manufacture, mileage, body type, and other technical parameters of the car. This combined information forms the final dataset used to train and test machine learning models. Integrating different types of data allows you to create a more complete picture of each vehicle, which helps improve the accuracy and reliability of price predictions.

2. Data Engineering

This section describes the data engineering processes applied to the course project. Once the data collection and processing stages have been completed, the data must be effectively managed to ensure subsequent analysis and use. The first step after the final processing of data is to archive it for later access and preservation. Google Drive is used as a platform for archiving the data. The final clean data is stored in Google Drive for access and preservation. Next, we use BigQuery for data warehousing. BigQuery is used for processing and analyzing data in the cloud. It provides scalable and performant data processing, making it an ideal choice for our purposes.

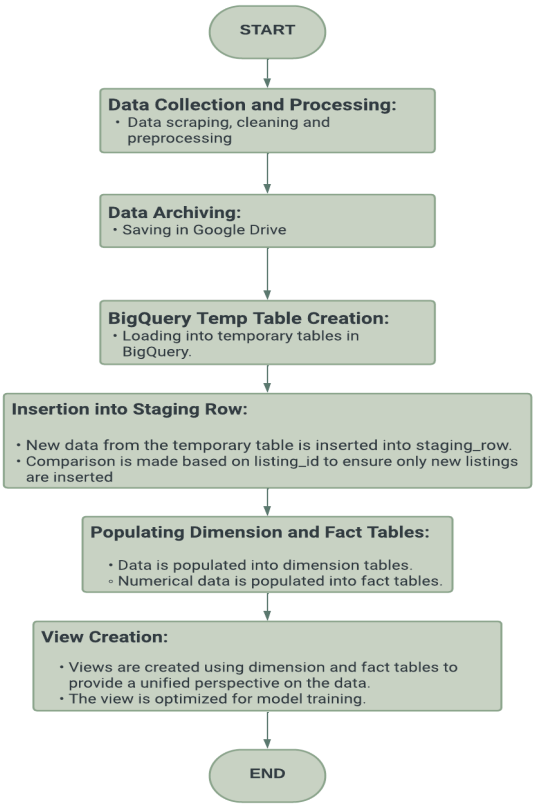

Figure 5: Data Storage Pipeline

The final data is loaded into BigQuery temporary tables for later analysis and use. We then create the staging row table, which serves as an intermediate store for the data before loading it into the final tables. We use this table to insert new data from temporary tables whenever available. This allows us to manage data updates and avoid duplication of information efficiently.

Data is inserted into the staging row table table by comparing listing identifiers (listing_id) with existing records. Only new or updated listings are ingested into the staging row table, ensuring that the data in our system is up-to-date and clean.

After the data is loaded into the staging row table, we populate dimension and fact tables. Dimensions include information about dates, vehicle images, vehicle characteristics (such as make, model, body type, fuel type, etc.), and other attributes.

The date dimension is constructed externally and integrated into the database. Facts represent relationships between dimensions and include vehicle write-off data, including write-off date, ad ID, ad link, and other attributes. The database diagram of dimension and fact tables can be observed in Figure 6.

Figure 6: Database Diagram

We then create a view based on these dim and fact tables, which provides a convenient way to analyze the data and prepare it for training models. The view contains grouped and combined data from various tables, making it easier to interpret and use later.

Finally, we connect to this new robust view and use it to train models and develop the application. This view provides access to up-to-date and structured data, allowing one to use it for a variety of purposes effectively, including model training and application development.

3.Development and comparison of car price forecasting models

As part of the fundamental work, two methodological approaches were implemented to build predictive models of car prices using data collected from the Auto.am website [4]. The first approach was limited to traditional technical and categorical vehicle features, while the second approach additionally included visual features extracted from vehicle images.

For both approaches, data preprocessing was integrated to improve modeling quality: categorical variables were transformed using the One-Hot Encoding technique, and numerical features were scaled to normalize ranges of values. Various machine learning algorithms were used to build predictive models.

These include linear regression as benchmark, Decision Tree, Random Forest,

KNN, Gradient Boosting, and XGBoost [3]. Grid search was used to fine-tune the hyperparameters for each algorithm to maximize the performance of the machine learning models. Models were evaluated based on metrics such as mean squared error (MSE), coefficient of determination (R^2), mean absolute error (MAE), and mean absolute percentage error (MAPE).

4. Web application

- Convenient tool designed to estimate car prices based on car features.
- Visualize market trends and statistics.

User Interface

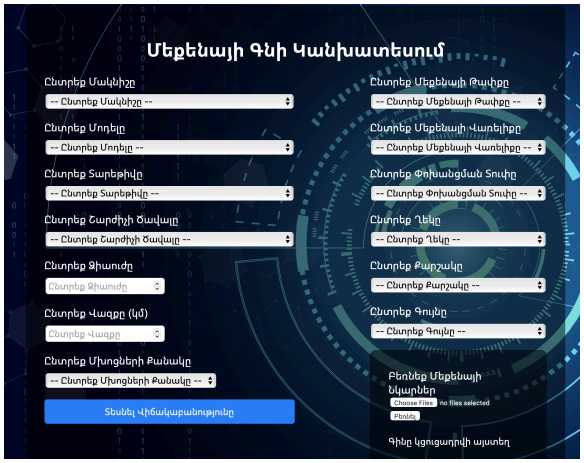

Figure 7: Web Application main page

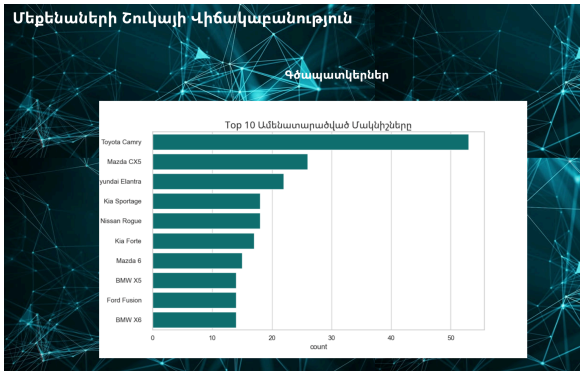

Figure 8: Web Application statistics page

The web application interface is designed to ensure maximum convenience and intuitiveness for users. The main page includes a form for entering data about the car, which is necessary to calculate the forecast price. The user inputs parameters

such as make, model, year, mileage, etc., and uploads vehicle images.

Input processing

After the user enters data, it is preprocessed. Categorical data, such as a car's make or model, is transformed using One-Hot Encoding so that it can be correctly interpreted by the model. Numerical data, such as mileage or engine size, is normalized to eliminate differences in value scales. As mentioned before, visual features are extracted from uploaded images using a pre-trained convolutional neural network.

The processed data is fed into the best-performing machine-learning model (Model Performance and Comparison) trained on the car sales data. Finally, the output prediction is rendered back into the web application for the user.

Statistical Analysis of the Market

The app's statistics page features graphs and charts that visualize various aspects of the car market. This may include the distribution of cars by age, mileage, fuel type, and other parameters. All data for visualization is taken from a database that is regularly updated to reflect the current state of the market.

5. Tools and Software

Various tools and technologies were used during the work. Programming, data manipulation, data analysis and development of Machine Learning models were performed in Python. SQL (Structured Query Language) was employed for data engineering tasks. To develop a web application, Python, JavaScript and the Flask framework are used for the backend, HTML/CSS for the front end.

Chapter 3

RESULTS and DISCUSSIONS

1. Model Performance and Comparison

In this section, we present the performance evaluation of the various machine learning models employed in our study, including image features. We assess the models based on their Mean Squared Error (MSE), R-squared (R2) score, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE). The metrics provide insights into each model's predictive capability and generalization ability. This analysis should guide the selection of a model based on both performance and the specific needs of the deployment environment.

Model	MAE	MAPE	MSE	R^2 Score
Random Forest	4,339	0.23	35,524,399	0.77
K-Nearest Neighbors	5,891	0.43	64,612,917	0.59
Decision Tree	6,027	0.3	69,364,489	0.55
XGBoost	4,153	0.22	31,932,216	0.79
Gradient Boosting	4,157	0.21	33,067,465	0.79
Linear Regression	4,955	0.36	45,561,756	0.70

Table 1: Model Performance of TEST

Model	MAE	MAPE	MSE	R^2 Score
Random Forest	1,707	0.09	5,776,176	0.96
K-Nearest Neighbors	0	0	0	1
Decision Tree	3,888	0.2	29,724,449	0.79
XGBoost	67	0.004	8,126	0.99
Gradient Boosting	1.3	0.2	0.06	0.99
Linear Regression	2010	0.12	8,529,085	0.94

Table 2: Model Performance of TRAIN

1. Random Forest:

Train: High R^2 (0.96) indicates a good fit on the training data.

Test: Lower R^2 (0.77) on the test data suggests some loss of generalization.

Conclusion: Moderately effective but shows signs of overfitting.

2. K-Nearest Neighbors:

Train: Perfect R^2 (1.0) suggests a perfect fit.

Test: Significantly lower R^2 (0.59) indicates substantial overfitting.

Conclusion: Overfits the training data severely, poor at generalizing.

3. Decision Tree:

Train: Decent R^2 (0.80), but not overly high.

Test: Further reduced R^2 (0.56) in testing, indicating some overfitting.

Conclusion: Fair performance but with visible overfitting.

4. XGBoost:

Train: Extremely high R^2 (0.9999), almost perfect fitting.

Test: High R^2 (0.80) in testing, best among all models.

Conclusion: Excellent performance with strong generalization capabilities.

5. Gradient Boosting:

Train: Practically perfect R^2 (1.0).

Test: High R^2 (0.79) but slightly less than XGBoost.

Conclusion: Very effective but slightly less optimal than XGBoost in test performance.

6. Linear Regression:

Train: High R^2 (0.94) showing a good fit.

Test: Lower R^2 (0.71), indicating loss in generalization.

Error Metrics (MSE, MAE, MAPE)

Mean Squared Error (MSE): Higher values in the test dataset for all models indicate variance in predictions. XGBoost and Gradient Boosting have lower MSE values in the test dataset, indicating more accurate predictions compared to others.

Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE): Lower values for these metrics in the test data for XGBoost and Gradient Boosting again highlight their superior performance in terms of error magnitude and percentage, respectively. The comparison of the metrics for all models also can be found in Figure 9.

Analysis of pipelines

By obtaining model performance results, we can indicate robust capability in capturing the underlying patterns and complexities of the dataset. Given these results, the XGBoost model has been decided to continue with, utilizing the best models for further development and implementation within the project.

After experimenting with both approaches, it was found that incorporating visual features did not significantly improve the performance of the models. This was confirmed by comparing the accuracy metrics of each model. While initial results did not show a substantial improvement, complex interactions between image features and other variables might yield benefits that are not immediately apparent but could be unlocked with more sophisticated modeling techniques or with larger datasets.

Based on the results and future vision, the second approach was chosen as the main one for further use in the project. Visual cues can be identified as key to improving the performance of vehicle price prediction models which can be confirmed by research [5], confirming the importance of an integrated approach to data analysis in today's automotive market.

Sticking to the second approach, the best ML model is picked and saved based on the performance metrics each time the pipeline is retrained.

Best Model

XGBoost emerges as the best model with an R^2 of 0.80 on the test set, indicating that it is capable of explaining 80% of the variance in the target variable. It also has lower error rates across MSE, MAE, and MAPE compared to other models, suggesting it makes fewer large errors in its predictions.

The scatter plot (Figure 10) shows the actual car prices on the x-axis against the predicted prices by the model on the y-axis. The red line represents the line of perfect prediction, where predicted values match the actual values. Most data points are clustered around the line, indicating a relatively good prediction accuracy, especially for lower-priced cars. The variance increases with higher car prices.

This residual plot (Figure 11) displays the residuals (the differences between the predicted and actual prices) on the y-axis against the predicted prices on the x-axis. The concentration of data points around the zero line for lower predicted prices suggests that the model is more accurate in this range. There does not appear to be any systematic pattern in the distribution, which indicates that the model does not suffer from heteroscedasticity.

This Distribution of Residuals plot (Figure 12) shows the frequency of various residuals occurring, along with a curve fitting these occurrences. The distribution is symmetric (bell shaped) and centered around zero, which is generally a good sign for model predictions, indicating that the model predictions tend to be close to actual ones. It appears to be slightly skewed to the right, meaning that there are more large positive residuals than large negative ones.

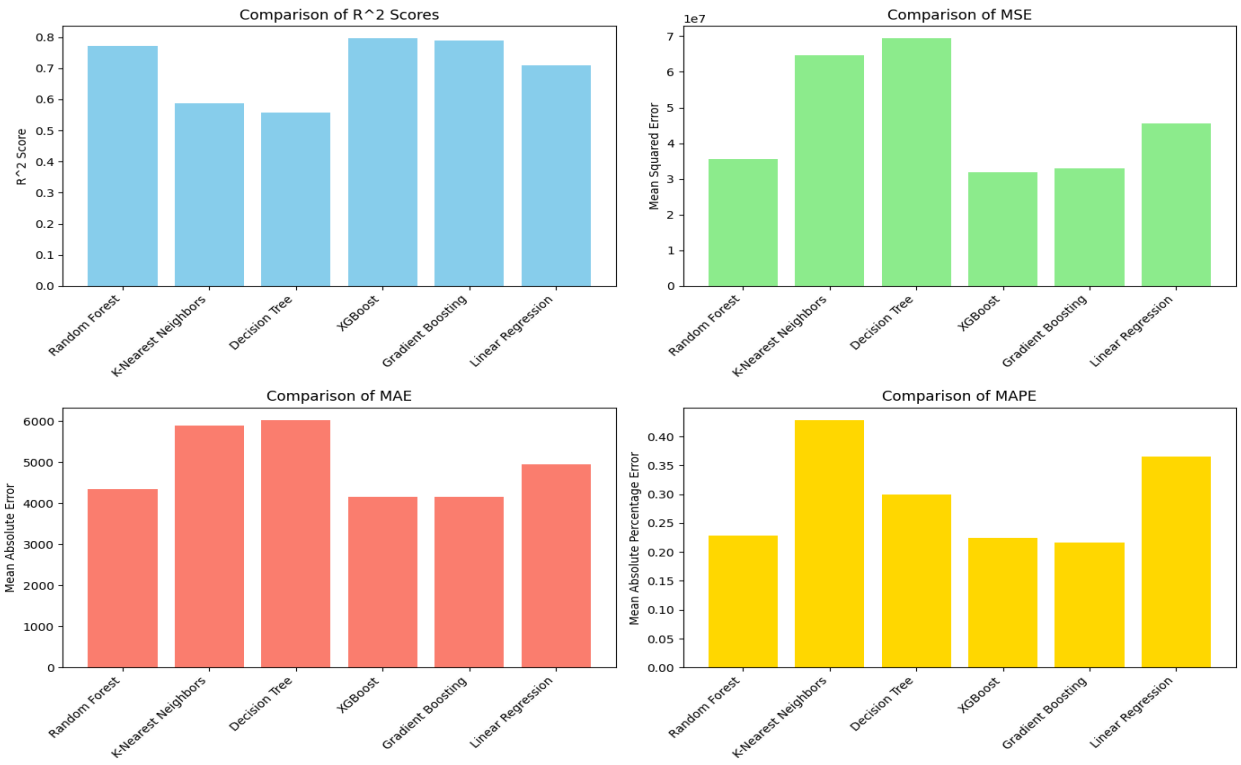

Figure 9: Comparison of metrics

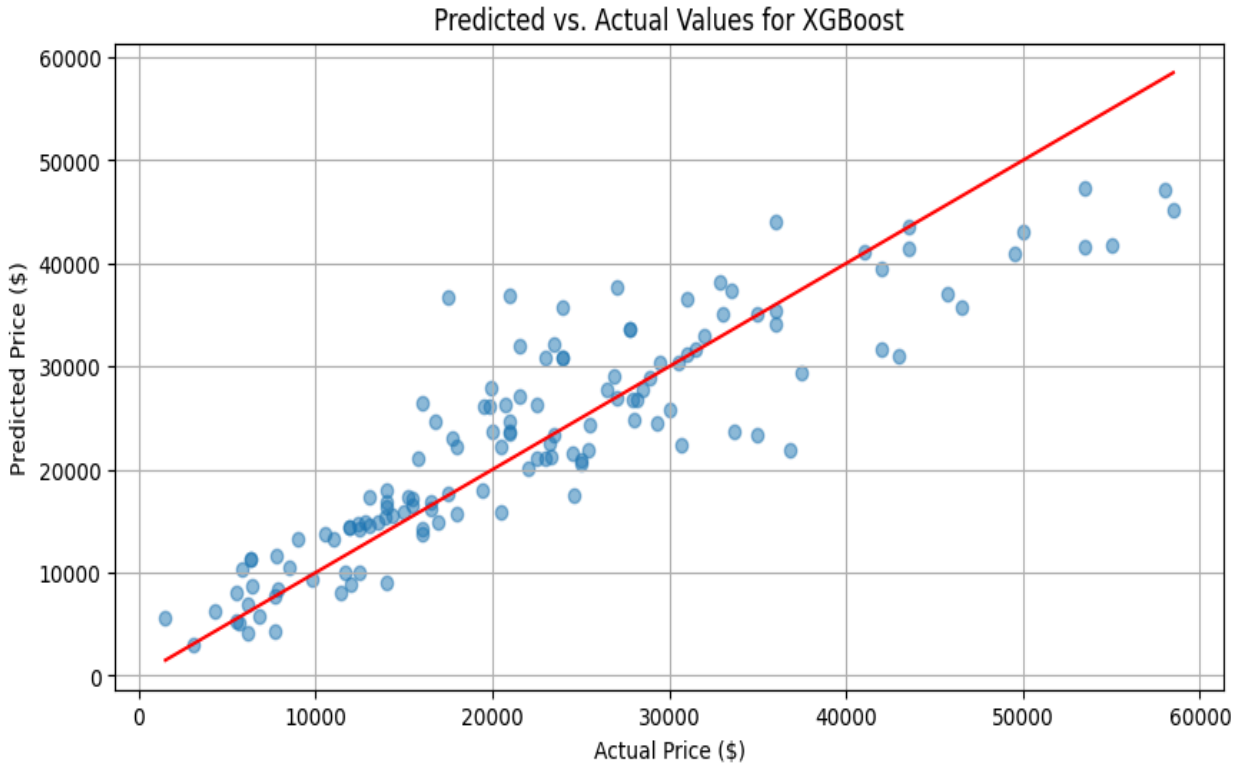

Figure 10: Predicted vs Actual values for best current best model

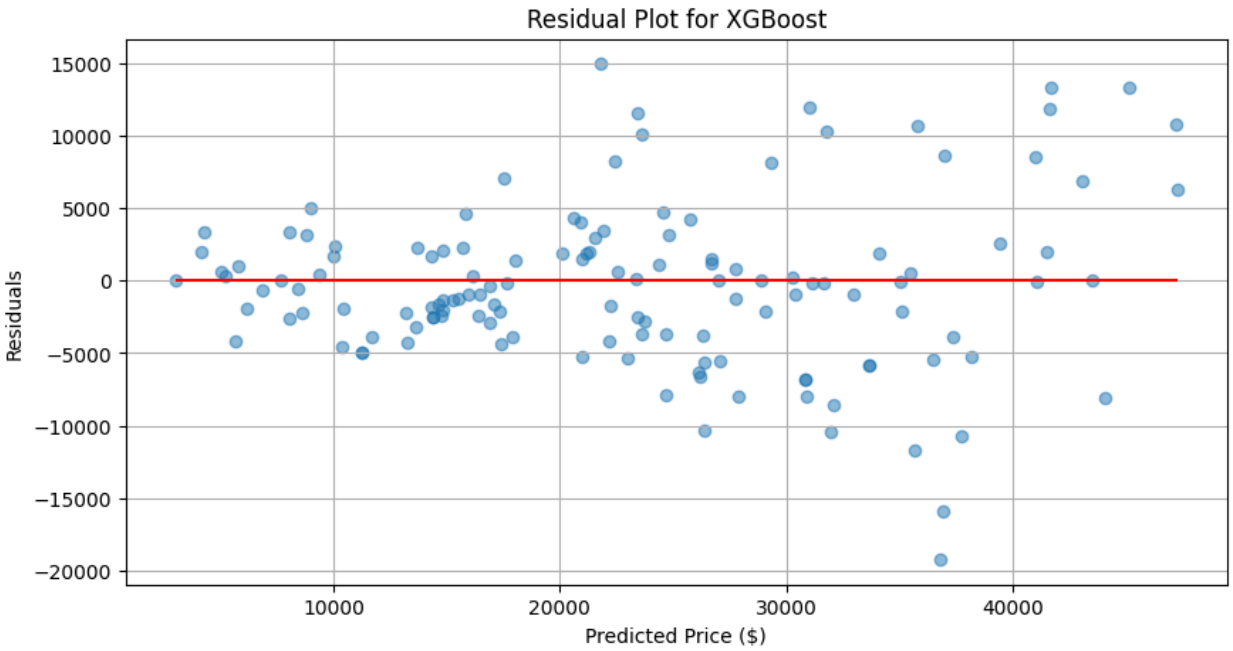

Figure 11: Residual plot of current best model

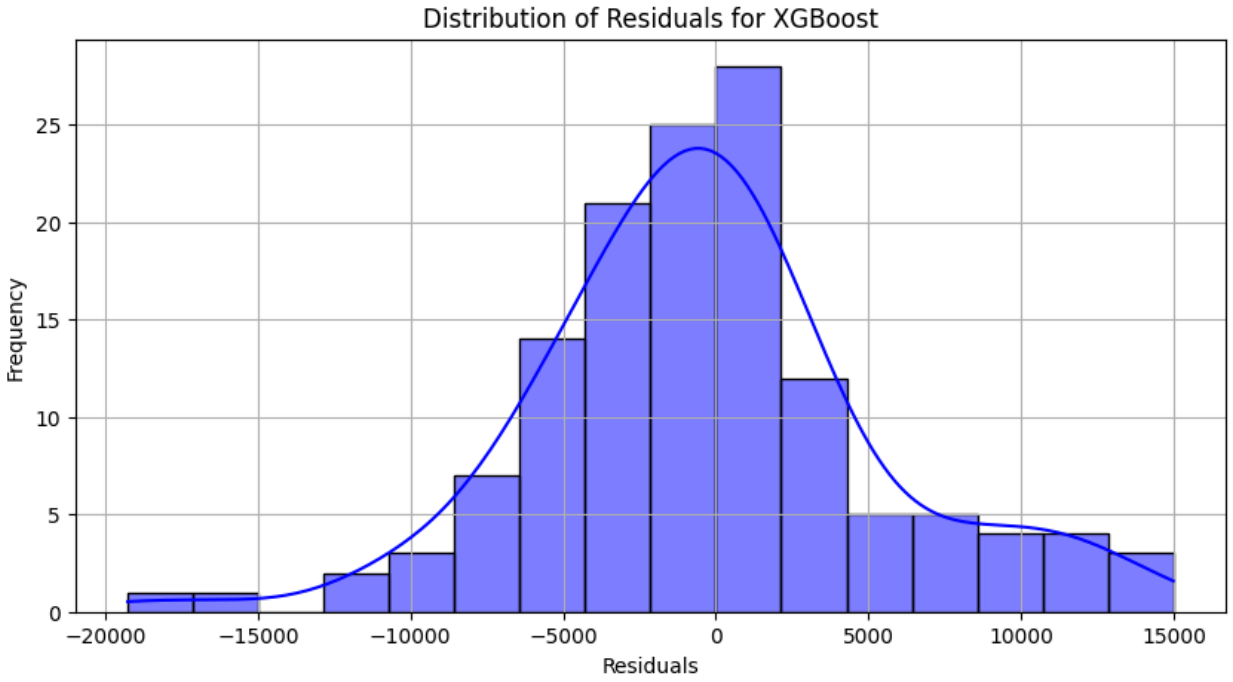

Figure 12: Distribution of residuals of current best mode

2. Discussions and Future Prospects

The results obtained from the model performances indicate modest predictive performance in estimating car prices based on the available dataset. However, it's crucial to recognize that these results are influenced by the relatively small size of the current dataset. The limited size of the dataset may constrain the models' ability to adequately capture the intricate relationships between car characteristics (especially image features) and prices.

With a more extensive and diverse dataset, the models could yield more accurate predictions by learning from a broader range of examples. Moving forward, it's imperative to address these challenges by collecting more comprehensive and accurately annotated data. By curating a larger dataset with cleaner and more representative samples, we can mitigate the impact of erroneous data entries and improve the overall performance of our predictive models.

Despite the current limitations, continuous data collection, rigorous preprocessing, and model refinement will enhance the models' performance. By iterative training and validating the models with new data, we anticipate achieving higher levels of prediction accuracy in estimating car prices.

Chapter 4

Conclusion

To conclude, we explored the potential of utilizing car characteristics to predict vehicle prices in the Armenian secondary market. Our key finding indicates that car images can aid in predicting car prices, provided an ample dataset is available. This insight underscores the value of visual data in enhancing the accuracy of predictive models in automotive markets. This approach presents several benefits, including

capturing subtle details not typically listed in traditional datasheets, such as the car's condition and aesthetic appeal. These details can influence buyer decisions and thus affect pricing. However, the effectiveness of this method is contingent upon the availability of high-quality and extensive image datasets. The lack of such data could limit the applicability and accuracy of the predictions.

The implications of our findings are substantial for stakeholders in the automotive industry, particularly in markets similar to Armenia's secondary market. By integrating image recognition technologies with existing prediction models, businesses can achieve more accurate pricing, tailored marketing strategies, and enhanced customer satisfaction.

For future research, we will continue expanding the dataset to include a wider range of car types and conditions, which could provide more insights into the diverse factors affecting car prices.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv.org*, Sep. 04, 2014. https://arxiv.org/abs/1409.1556

[2] P. Ruiz, "Understanding and visualizing ResNets," *Towards Data Science*, Apr. 23, 2019. [Online]. Available: https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8

[3] P. Gajera, A. Gondaliya, and J. Kavathiya, "Old car price prediction with machine learning," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 3, no. 3, pp. 284–290, Mar. 2021, doi: https://www.irjmets.com/uploadedfiles/paper/volume3/issue_3_march_2021/6681/1628083284.pdf

[4] "Ավտոմեքենաների վաճառք Հայաստանում," *Auto.am*. https://auto.am

[5] R. R. Yang, S. Chen, and E. Chou, "AI Blue Book: Vehicle Price Prediction using Visual Features," *arXiv.org*, Mar. 29, 2018. https://arxiv.org/abs/1803.11227