

CMPT 3830: Machine Learning

Work Integrated Learning-1

**Project Report: Phase 1**

**Project Title Optimal Vehicle Pricing**

**In collaboration with**



**Submitted By:**

|  |  |  |
| --- | --- | --- |
| Name | ID | Email |
| Emem Antia | 3108445 | eantia@norquest.ca |
| Ayo Pedro | 3115351 | apedro@norquest.ca |
| Bube Ohaneje | 3110527 | bohaneje@norquest.ca |
| Monsurat Adejumo | 3027131 | madejumo118@norquest.ca |

**Submission: Date: 26 February 2025**

**Winter 2025**

**Table of Contents:**

A table of content is required with section numbers, names with correct page numbers. An example is shown below.

Contents

[1. Project Phase: 3](#_Toc191571756)

[2. Team Members’ Name with specific roles 3](#_Toc191571757)

[3. Reporting Period: 3](#_Toc191571758)

[4. Project Overview: 3](#_Toc191571759)

[Problem Statement: 3](#_Toc191571760)

[Solution Approach: 4](#_Toc191571761)

[5. Dataset 4](#_Toc191571762)

[5.1 Exploratory Data Analysis (EDA) Highlights: 4](#_Toc191571763)

[5.2 Visualization: 5](#_Toc191571764)

[ 5](#_Toc191571765)

[Types of Visualizations Used 5](#_Toc191571766)

[Ensuring Intuitive & Actionable Insights 6](#_Toc191571767)

[6. Challenges Encountered: 6](#_Toc191571768)

[7. Stakeholder Engagement: 6](#_Toc191571769)

[8. Lessons Learned: 7](#_Toc191571770)

[What Worked Well 7](#_Toc191571771)

[9. Future Recommendations: 7](#_Toc191571772)

[10. Impact on the Community: 8](#_Toc191571773)

[**Potential & Realized Benefits** 8](#_Toc191571774)

[11. Project Conclusion: 9](#_Toc191571775)

[12. Acknowledgments: 10](#_Toc191571776)

[13. Appendices: 10](#_Toc191571777)

[14. References 10](#_Toc191571778)

# Project Phase:

This phase focuses on **data exploration and preprocessing** for predicting vehicle pricing and analyzing sales trends. The key accomplishments include:

* Conducting **Exploratory Data Analysis (EDA)** to identify trends, outliers, and correlations.
* Conducted **EDA** to identify key trends, anomalies, and missing values.
* Performed **data cleaning** to remove inconsistencies.
* Applied **outlier detection & removal** using **IQR and Z-score methods**.
* Encoded **categorical variables** using **Label Encoding**.
* Split data into **features (X) and target (y)** to prepare for modeling.
* Understanding **pricing behavior and demand** based on vehicle features such as mileage, model year, and brand.
* Developing **data visualizations** to support insights.

# Team Members’ Name with specific roles

|  |  |
| --- | --- |
| Name | Role |
| Emem Ayo, Bube, Monsurat | Data Preprocessing & EDA |
| Emem, Ayo, Bube, Monsurat | Data Cleaning & Handling Outliers |
| Monsurat Ayo, Bube, Emem | Feature Engineering & Encoding |

# Reporting Period:

* **Week 1-2:** Data collection & initial analysis.
* **Week 3:** Data cleaning & handling missing values.
* **Week 4:** Outlier detection & removal.
* **Week 5:** Feature selection & encoding.
* **Week 6:** Preparing dataset for modeling.

# Project Overview:

# Problem Statement:

The goal of this project is to develop a machine learning model to predict vehicle prices based on features such as year, make, model, mileage, and other vehicle attributes. This will enable Go Auto dealerships to optimize pricing strategies and maximize sales.

# Solution Approach:

* EDA & Cleaning: Identify trends, missing values, and inconsistencies.
* Outlier Treatment: Remove extreme values to improve model performance.
* Feature Engineering: Transform categorical variables & extract key insights.
* Data Preprocessing: Encode categorical variables & normalize numerical ones.
* Modeling & Optimization: To be covered in the next phase

# Dataset

* Dataset Source: Go Auto Business Intelligence Team (via Canadian Black Book APIs)
* Data Size: 145,114 records, 46 features

**Types of Data:**

* Numerical Features: Mileage, Price, MSRP, Days on Market, Engine Size
* Categorical Features: Make, Model, Series, Drivetrain, Fuel Type, Exterior Color
* Target Variable: Price.

# 5.1 Exploratory Data Analysis (EDA) Highlights:

* Conducted summary statistics and distribution analysis to understand numerical variables.
* Identified missing values, which were significant in columns like engine\_size, series, and price\_history\_delimited.
* Analyzed price distribution, revealing extreme outliers and incorrect entries.
* Used correlation analysis to understand the relationship between key variables such as mileage, msrp, model\_year, and price.

**Key Findings & Patterns**

***price\_history\_delimited Analysis:***

* Parsed price\_history\_delimited to extract historical price trends, including first price, last price, number of price changes, and days between price changes.
* Created a price stability score to analyze frequent price changes over time.
* Found that some vehicles experienced multiple price changes within short periods, indicating market-driven adjustments.

***Price & Mileage Errors:***

* Discovered new cars with high mileage, which was incorrect.
* Found used cars with zero mileage, likely due to data entry errors.
* Addressed errors by implementing an expected max mileage per model year filter and logical correction strategies.

**Outliers in Pricing:**

* Detected illogical price values such as cars priced at $11 or above $1 million.
* Applied a reasonable pricing threshold to filter out extreme outliers while retaining valuable pricing data.

**Certified Vehicles Pricing Anomaly:**

* Found that certified vehicles had lower median prices than non-certified ones, contradicting expectations.
* This suggested potential dealer-specific pricing strategies or data inconsistencies.

**Fuel Type & Engine Size Impact on Pricing:**

* Electric and hybrid vehicles showed higher price retention compared to gasoline vehicles.
* Larger engine sizes were generally associated with higher prices, but some luxury models with smaller engines retained high value due to brand perception.

**Time-Series Price Trends:**

* Vehicles with frequent price adjustments indicated dealerships adjusting pricing dynamically based on demand.
* Some models showed seasonal price variations, suggesting trends influenced by market conditions.

# Visualization:

Developing Interactive Visualizations

* Created multiple interactive visualizations using Matplotlib, Seaborn, and Plotly to represent EDA findings effectively.
* Used dynamic filtering to allow deeper analysis based on make, model\_year, fuel\_type, and dealer location.
* Developed interactive price trend charts to analyze how vehicle prices change over time.

# 

# Types of Visualizations Used

**Bar Charts:**

* Compared average prices across vehicle makes and models.
* Highlighted the distribution of certified vs. non-certified vehicle prices.
* Showcased frequency of price changes per dealer.
* Heatmaps:
* Analyzed correlation between mileage, msrp, engine\_size, and price.
* Identified high-impact features that influence pricing.

**Time-Series Plots:**

* Used price\_history\_delimited to visualize price changes over time.
* Identified trends in pricing behavior, including frequent markdowns and seasonal pricing variations.

**Boxplots & Histograms:**

* Displayed price distribution and outlier analysis before and after cleaning.
* Visualized mileage patterns for new vs. used vehicles.
* Geographical Maps:
* Mapped dealership locations to analyze pricing differences across regions.
* Visualized high-demand vehicle models based on location.

# Ensuring Intuitive & Actionable Insights

* Ensured all visualizations had clear labels, legends, and interactive elements to improve usability.
* Color schemes were optimized for clarity, distinguishing between different categories (e.g., fuel type, stock type, dealer regions).
* Integrated tooltips and hover interactions for deeper insights in time-series and geographical visualizations.

# Challenges Encountered:

* **Inconsistent price history formatting: The price\_history\_delimited column contained multiple formats and delimiters, requiring extensive parsing and restructuring.**
* **Handling extreme price outliers: Some vehicle listings showed unrealistic price values (e.g., $11 or above $1 million). Setting appropriate thresholds and filtering criteria without removing valuable data was a key challenge.**
* **Missing & inconsistent mileage values: Some new vehicles had abnormally high mileage, while used cars had 0 mileage recorded. This required setting logical rules to impute missing mileage values based on expected max mileage for each model year.**
* **Data encoding: Many categorical variables had high cardinality (e.g., make, model, series), making encoding challenging. We opted for a label encoding**

# Stakeholder Engagement:

**Summary of the Last Session with Stakeholders**

A stakeholder session was conducted to present key EDA findings and gatherinsights from Go Auto’s Business Intelligence Team. The session included:

**Presentation of Data Trends:**

* Analyzed price fluctuations using *price\_history\_delimited*.
* Identified inconsistencies in mileage, highlighting unrealistic values.
* Showcased the impact of model year, fuel type, and engine size on price trends.

**Feedback and Additional Information Requested by Stakeholders:**

**Understanding dealer\_type Values (“I” and “F”):**

* Stakeholders explained that these codes represent **different dealership categories** but suggested more documentation is needed.
* **More Data on Used Vehicles:**
* Requested to compare **depreciation rates** by brand and fuel type to validate assumptions.

**Key Takeaways from the Discussion:**

* Mileage and price errors were confirmed as data entry issues, and logical imputation strategies were approved.
* More work needs to be done to analyze analysis on historical price changes**.**

# Lessons Learned:

### **What Worked Well**

**Thorough Exploratory Data Analysis (EDA):**

* Conducting in-depth EDA helped uncover hidden data inconsistencies, missing values, and pricing anomalies.
* Visualizing trends using interactive charts made it easier to communicate findings to stakeholders.

**Effective Data Cleaning Strategies:**

* Addressing price\_history\_delimited inconsistencies provided valuable insights into pricing patterns.
* Using logical rules for mileage correction helped improve data quality significantly.
* Filling missing engine\_size values using a structured approach (per make/model) improved completeness.

**Stakeholder Engagement and Feedback:**

* Regular check-ins with the Go Auto Business Intelligence Team helped clarify dataset uncertainties.
* Gathering real-world insights (e.g., dealership pricing strategies)
* improved understanding of pricing trends beyond data alone.

**Structured Outlier Handling:**

* + Implementing IQR-based outlier detection effectively removed extreme price values while preserving valid data.

**Enhancing Data Imputation Methods:**

Some missing values (e.g., engine\_size, series) lacked clear imputation strategies due to limited reference data.

* + Future projects could integrate external automotive datasets or use predictive modeling for better imputation.

**Handling Large Datasets More Efficiently:**

* + Computational constraints slowed down certain processes, such as correlation analysis and data filtering.
  + Optimizing memory usage and leveraging distributed computing tools would improve performance for larger datasets.

# Future Recommendations:

1. **Better Data Preprocessing & Automation:**
   * Automate **data cleaning, outlier detection, and imputation** to streamline preprocessing steps.
   * Leverage **cloud computing resources** to process larger datasets more efficiently.
2. **Real-time Price Prediction API:**
   * Develop an API-based system where dealerships can **input vehicle attributes and receive optimized price predictions in real-time**.
   * Implement a **dashboard for price monitoring** based on historical trends and dealership behavior.
3. **Incorporating Macroeconomic & Market Trends:**
   * Include factors like **inflation, interest rates, and fuel prices** to account for economic influences on car pricing.
   * Investigate how **electric vehicle adoption rates** impact traditional fuel-based car prices.

.

# Impact on the Community:

1. **Better Data Preprocessing & Automation:**
   * Automate data cleaning, outlier detection, and imputation to streamline preprocessing steps.
   * Leverage cloud computing resources to process larger datasets more efficiently.
2. **Real-time Price Prediction API:**
   * Develop an API-based system where dealerships can input vehicleattributes and receive optimized price predictions in real-time.
   * Implement a dashboard for price monitoring based on historical trends and dealership behavior.
3. **Incorporating Macroeconomic & Market Trends:**
   * Include factors like inflation, interest rates, and fuel prices to account for economic influences on car pricing.
   * Investigate how electric vehicle adoption rates impact traditional fuel-based car prices.

**Potential & Realized Benefits**

1. **Empowering Dealerships with Data-Driven Pricing:**
   * Helps Go Auto dealerships optimize pricing, leading to faster sales cycles and increased profitability.
   * Reduces reliance on manual pricing decisions and improves consistency across dealerships.
2. **Consumer Benefits & Market Transparency:**
   * Provides consumers with fair and competitive pricing, increasing trust in dealership pricing strategies.
   * Helps buyers understand price trends and make informed purchasing decisions.
3. **Optimizing Vehicle Resale Values:**
   * Identifies ideal resale price windows to help car owners maximize trade-in values.
   * Supports financial planning for buyers by predicting expected depreciation trends.
4. **Reducing Price Manipulation Risks:**
   * Detects and prevents unusual price fluctuations that could lead to market inefficiencies or unethical pricing practices.
   * Encourages fair market competition by ensuring that pricing aligns with true vehicle value.

# Project Conclusion:

The **Vehicle Pricing Optimization using Machine Learning** project has successfully met and, in some cases, exceeded its initial objectives.

The primary goal of this phase was to prepare the dataset for modeling by conducting **Exploratory Data Analysis (EDA), data cleaning, outlier detection, feature engineering, and encoding**. Through a structured and methodical approach,

we have:

**Developed a robust understanding of vehicle pricing trends** by analyzing *price\_history\_delimited* and identifying patterns in price changes.

**Ensured data quality and consistency** by handling missing values, addressing inconsistencies in mileage, and resolving outliers in price and other numerical features.

**Implemented effective feature engineering techniques** that transformed raw data into valuable inputs for predictive modeling.

**Engaged with stakeholders** to validate findings, refine data assumptions, and incorporate real-world dealership insights into the pricing strategy.

**Created insightful visualizations** that highlight price fluctuations, dealership

**Exceeding Initial Goals**

* The project we successfully parsed and structured historical price data, which was initially a significant challenge due to inconsistent formatting.
* Additional insights into dealership pricing behavior and seasonal fluctuations were extracted, providing deeper contextual value to the pricing model.
* The cleaned and preprocessed dataset is now ready for advanced machine learning models, setting a strong foundation for predictive analytics.

**Next Steps**

* With the data now structured and optimized, the next phase of the project will focus on developing and evaluating predictive models to accurately estimate vehicle prices.
* Future work will also explore time-series forecasting, model validation, and real-time pricing recommendations, strategies, and vehicle value trends.

.

# Acknowledgments:

Special thanks to the **Go Auto Business Intelligence Team** for providing data and insights and our Instructor for the Guidance.

# Appendices:

**Sprint 1: Project Setup & Data Understanding**

* **Sprint Goal:** Set up project structure, define objectives, and explore the dataset.
* **Key Tasks:**
  + Define problem statement and objectives.
  + Set up GitHub and Jira board.
  + Perform initial data exploration (EDA).
* **Outcomes:** Completed problem definition, dataset collection, and initial EDA.

.

# References

Cite your sources (MUST follow APA style).