CMPT 3830: Machine Learning Work Integrated Learning-1

Project Report: Phase 1
Predictive Modeling for Optimal Vehicle Pricing

In collaboration with

Submitted By:

Name	ID	Email
Bharat Bajaj	3092681	bbajaj@norquest.ca
Geetika LNU	3094334	ggeetika@norquest.ca
McLaren Packard	3089831	mpackard@norquest.ca
Renata Aline Moura Saccon	3061327	rsaccon@norquest.ca

Submission: Date: October 15, 2024
Fall 2024

Contents

List of Figures:

Figures in each page (if any) MUST be listed in this section.

List of Tables:

Tables in each page (if any) MUST be listed in this section.

1. Project Phase:

It is a summary of the phase you are currently working on. In this section, you are required to provide information about your accomplishments such as Exploring different Machine Learning Models, choosing a model, Apply ML model on Go Auto’s dataset, Model Evaluation and Model Optimization

1. Initiation and Planning

- This phase set the foundation for the project by establishing a clear team structure, defining the problem, and creating a roadmap for execution.
- Key deliverables, such as the team charter, project proposal, and a detailed project timeline, were successfully completed on schedule.
- The team aligned on objectives, responsibilities, and expectations, ensuring smooth coordination and a clear vision moving forward.

2. Data Preparation

- The team preprocessed the dataset, addressing missing values, cleaning, and encoding features to ensure data quality.
- New features, such as vehicle age, were engineered to enhance model accuracy

and performance.

3. **Exploratory Data Analysis (EDA)**
   - A thorough EDA was conducted to uncover trends, patterns, and outliers within the dataset.
   - Key insights were visualized, forming the analytical foundation for the predictive model.

4. **Insights and Recommendations**
   - Actionable insights derived from the analysis will guide Go Auto in optimizing vehicle pricing strategies and decision-making processes.

5. **Demo 1: Presentation and Feedback**
   - The team presented initial findings from the EDA to stakeholders, incorporating feedback to refine the project's direction and model development.

6. **Phase 1 Report**
   - Submission of the Phase 1 report, summarizing EDA results, insights, and next steps.

7. **Model Development (Upcoming)**
   - The team will begin constructing and training the machine learning model. This phase will focus on selecting appropriate algorithms and testing the initial performance of the model.

8. **Model Optimization (Upcoming)**
   - The team will optimize the model by fine-tuning hyperparameters and validating the model's accuracy to ensure high-quality predictions for vehicle pricing.

9. **Insights and Recommendations II (Upcoming)**
   - The team will analyze the final model's output to provide actionable insights and recommendations, helping Go Auto refine their pricing strategies.

10. **Demo 2: Presentation and Feedback (Upcoming)**
    - The team will deliver a second presentation, showcasing the model's performance and receiving feedback to further improve outcomes.

1. **Phase 2 Report (Upcoming)**
   - The Phase 2 report will be submitted, detailing the model development, optimization process, and actionable insights.

2. **Final Deliverables (Upcoming)**
   - The final phase will include the submission of a comprehensive presentation summarizing the project from data analysis to model recommendations.

Project tasks:

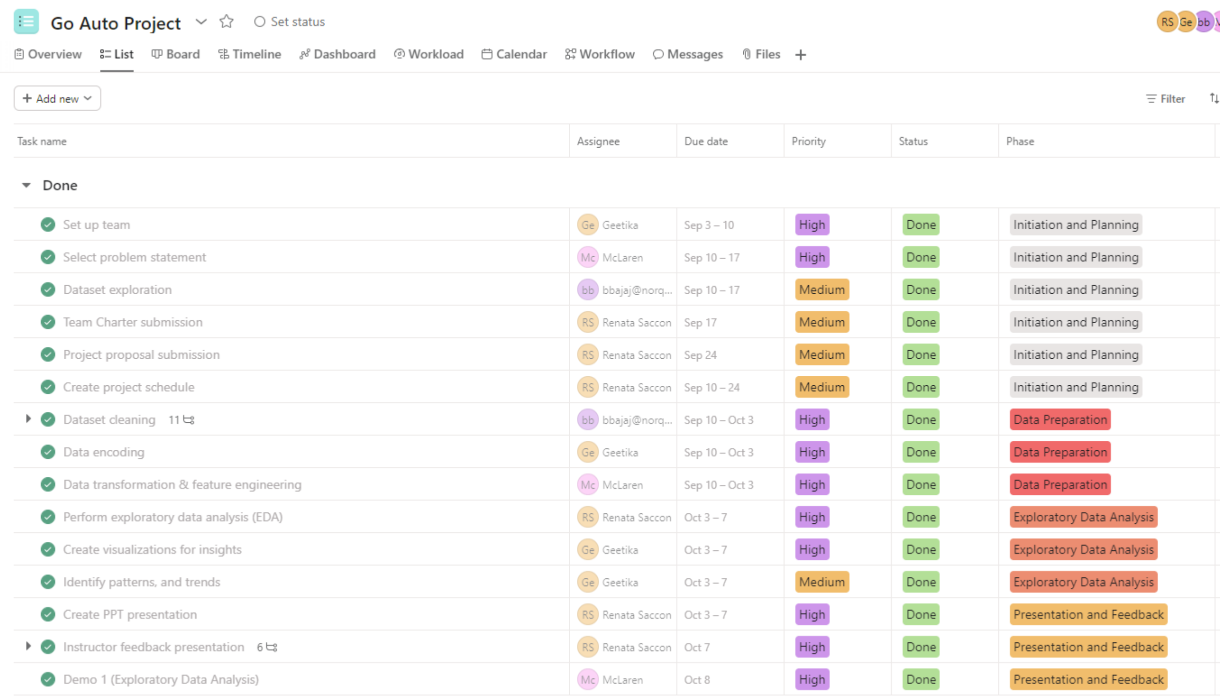
MLDrive team is using Asana to manage our project, and you can access the website here: https://asana.com. Please note that access will need to be requested. For now, we are including a screenshot below for your reference.

Please keep in mind that as we move forward, the steps, tasks, and deliverable dates may change, so adjustments may be required.

- Done:

Figure 1

Go Auto Project Management Finished Tasks



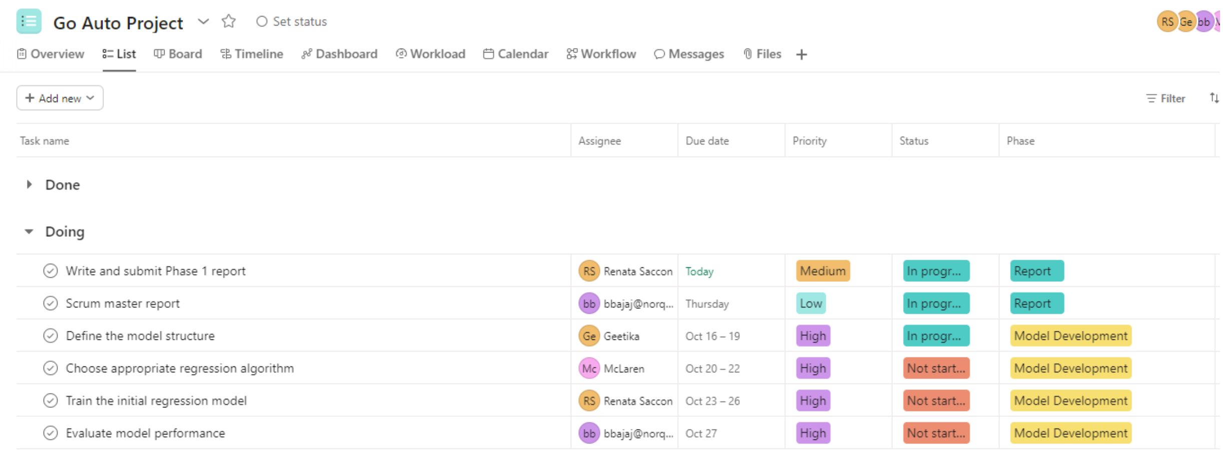
Task name	Assignee	Due date	Priority	Status	Phase
Done					
Set up team	Ge Geetika	Sep 3 – 10	High	Done	Initiation and Planning
Select problem statement	Mc McLaren	Sep 10 – 17	High	Done	Initiation and Planning
Dataset exploration	bb bbajaj@norq...	Sep 10 – 17	Medium	Done	Initiation and Planning
Team Charter submission	RS Renata Saccon	Sep 17	Medium	Done	Initiation and Planning
Project proposal submission	RS Renata Saccon	Sep 24	Medium	Done	Initiation and Planning
Create project schedule	RS Renata Saccon	Sep 10 – 24	Medium	Done	Initiation and Planning
Dataset cleaning 11	bb bbajaj@norq...	Sep 10 – Oct 3	High	Done	Data Preparation
Data encoding	Ge Geetika	Sep 10 – Oct 3	High	Done	Data Preparation
Data transformation & feature engineering	Mc McLaren	Sep 10 – Oct 3	High	Done	Data Preparation
Perform exploratory data analysis (EDA)	RS Renata Saccon	Oct 3 – 7	High	Done	Exploratory Data Analysis
Create visualizations for insights	Ge Geetika	Oct 3 – 7	High	Done	Exploratory Data Analysis
Identify patterns, and trends	Ge Geetika	Oct 3 – 7	Medium	Done	Exploratory Data Analysis
Create PPT presentation	RS Renata Saccon	Oct 3 – 7	High	Done	Presentation and Feedback
Instructor feedback presentation 6	RS Renata Saccon	Oct 7	High	Done	Presentation and Feedback
Demo 1 (Exploratory Data Analysis)	Mc McLaren	Oct 8	High	Done	Presentation and Feedback

Screenshot of tasks finished

○ Doing:

Figure 2

Go Auto Project Management Current Tasks



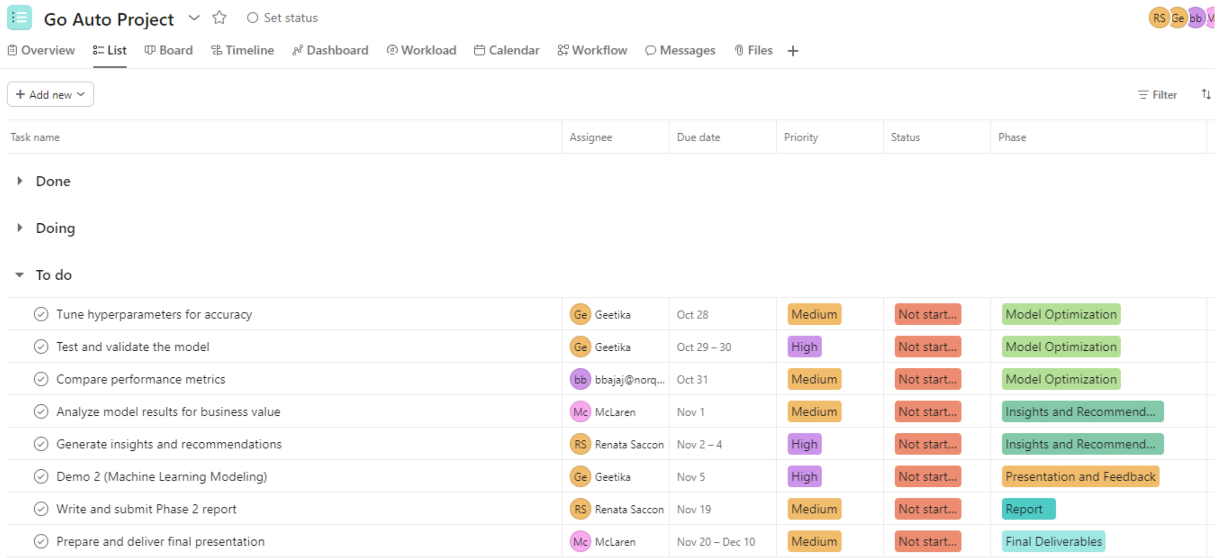
Task name	Assignee	Due date	Priority	Status	Phase
Done					
Doing					
Write and submit Phase 1 report	RS Renata Saccon	Today	Medium	In progr...	Report
Scrum master report	bb bbajaj@norq...	Thursday	Low	In progr...	Report
Define the model structure	Ge Geetika	Oct 16 – 19	High	In progr...	Model Development
Choose appropriate regression algorithm	Mc McLaren	Oct 20 – 22	High	Not start...	Model Development
Train the initial regression model	RS Renata Saccon	Oct 23 – 26	High	Not start...	Model Development
Evaluate model performance	bb bbajaj@norq...	Oct 27	High	Not start...	Model Development

Screenshot of tasks currently being worked on

○ To do:

Figure 2

Go Auto Project Management To do Tasks



Task name	Assignee	Due date	Priority	Status	Phase
Done					
Doing					
To do					
Tune hyperparameters for accuracy	Ge Geetika	Oct 28	Medium	Not start...	Model Optimization
Test and validate the model	Ge Geetika	Oct 29 – 30	High	Not start...	Model Optimization
Compare performance metrics	bb bbajaj@norq...	Oct 31	Medium	Not start...	Model Optimization
Analyze model results for business value	Mc McLaren	Nov 1	Medium	Not start...	Insights and Recommend...
Generate insights and recommendations	RS Renata Saccon	Nov 2 – 4	High	Not start...	Insights and Recommend...
Demo 2 (Machine Learning Modeling)	Ge Geetika	Nov 5	High	Not start...	Presentation and Feedback
Write and submit Phase 2 report	RS Renata Saccon	Nov 19	Medium	Not start...	Report
Prepare and deliver final presentation	Mc McLaren	Nov 20 – Dec 10	Medium	Not start...	Final Deliverables

Screenshot of tasks to be done

2. Team Members' Name with specific roles

At MLDrive, we recently went through a restructuring to better position ourselves for success. We took a close look at the skills within our team and the needs of the project, and we created new roles to make sure everyone is in the right spot.

Team Member	Role	Key Responsibilities
Geetika LNU	Team Lead & CEO	- Oversees the project and makes key decisions.
		- **Data Cleaning** (Sep 30 - Oct 3): Prepares the dataset, handles missing values, removes duplicates.
		- **Define Model Structure** (Oct 16-19): Leads the development of the model.
		- **Choose Regression Algorithm** (Oct 20-22): Selects the appropriate algorithm.
		- **Tune Hyperparameters** (Oct 28): Ensures the model is fine-tuned for accuracy.
		- **Test & Validate Model** (Oct 29-30): Ensures the model functions accurately.
Renata Aline Moura Saccon	Project Manager & Data Engineer	- **Team & Project Charter** (Sep 17 & Sep 24): Creates and submits charters.
		- Manages project schedule, ensuring deadlines are met.
		- **Exploratory Data Analysis (EDA)** (Oct 3-7): Leads data analysis to uncover trends.

		- **Create Visualizations** (Oct 3-7): Develops data visualizations. - **Generate Insights & Recommendations** (Nov 2-4): Provides actionable business recommendations. - **Submit Phase 1 & 2 Reports** (Oct 15 & Nov 19): Documents findings. - Prepares the final presentation.
Bharat Bajaj	Scrum Master & Model Performance Analyst	- **Scrum Master Report** (Oct 17): Updates team progress. - **Evaluate Model Performance** (Oct 27): Tests the model's effectiveness. - **Test & Validate the Model** (Oct 29-30): Ensures accurate model validation. - **Compare Metrics** (Oct 31): Compares performance metrics to evaluate the model.
McLaren Packard	Data Scientist	- **Dataset Exploration** (Sep 10-17): Leads exploration of the dataset. - **Data Encoding & Feature Engineering** (Sep 28 - Oct 3): Prepares data by encoding variables and creating features. - **Train Regression Model** (Oct 23-26): Builds and trains the model. - **Analyze Model Results** (Oct 31): Interprets results for business value. - **Final Presentation** (Nov 20-Dec 10): Prepares technical components of the final presentation.

3. Reporting Period: [Specify the reporting period, e.g., Month/Year to Month/Year]Key breakdown of each part submission.

Reporting Period	Key Tasks & Deliverables
September 10, 2024 - September 17, 2024	- **Team Charter Submission**: Defined team roles, responsibilities, and overall project goals.
September 17, 2024 - September 24, 2024	- **Project Charter Submission**: Formalized the project scope, objectives, deliverables, and risk management plan.
September 24, 2024 - October 15, 2024	- **Phase 1 Report (EDA)**: Completed Exploratory Data Analysis (EDA), created visualizations, and highlighted trends and insights.
October 15, 2024 - November 19, 2024	- **Phase 2 Report (ML Modeling)**: Developed machine learning models, fine-tuned them, and provided business insights.

November 20, 2024 - December 10, 2024	- **Final Project Submission**: Delivered the final presentation, summarizing the EDA, ML models, and recommendations.

4. Project Overview: Overview with the problem statement and solution approach you followed.

Problem Statement:
The project focuses on developing a Pricing Model (Regression) for vehicle sales optimization at Go Auto. The goal is to create a machine learning regression model that predicts the optimal pricing of vehicles based on their key attributes, such as year, make, model, and mileage. This model will help Go Auto enhance its pricing strategy, ensuring competitiveness in the market while maximizing sales and profitability.

Approach:
The team has conducted an exploratory data analysis (EDA) to identify patterns and trends within the dataset, followed by developing and fine-tuning the regression model. By using vehicle attributes, the model will generate price estimates or ranges, enabling Go Auto to make data-driven decisions regarding vehicle pricing.

So far, our work has focused on preparing the dataset for analysis and future model development. We have made significant progress in cleaning the data, handling missing values, encoding categorical variables, and performing initial exploration analysis.

1. Data Cleaning

We began by thoroughly inspecting the dataset using standard pandas functions like df.head() and df.info() to get an understanding of its structure. After identifying issues such as missing values, duplicates, and inconsistent formats, we applied the following techniques:

- **Handling Missing Values**:
  - For numerical columns, we filled in missing data using the mean of the column. The mean was chosen to reduce the influence of outliers.
  - For categorical columns, we filled in missing values using the mode (the most frequent value). This ensures that the missing values are filled in with realistic, common values that reflect the overall data distribution.
- **Removing Duplicates**:
  - We used the df.drop_duplicates() function to remove any duplicate rows that could skew our analysis, ensuring that each row represented a unique vehicle.
- **Outlier Detection and Handling**:
  - Outliers were detected using the Interquartile Range (IQR) method. This method helped us identify data points that fell significantly outside the expected range. We either removed these outliers or manually edited them where the IQR method did not apply well.
- **Handling Zero Values**:
  - In cases where zero values were invalid (e.g., mileage or price fields), we replaced them with NaN and treated them using the same methods as missing values.

2. Data Encoding

With the dataset cleaned, we moved on to encoding categorical variables to make them suitable for the machine learning algorithms we plan to use. Here’s what we did:

- **Label Encoding**:
  - For columns with only two categories (e.g., Stock Type, where values are either "new" or "used"), we applied label encoding. We converted 'new' to 1 and 'used' to 0, simplifying these values for model training.
- **One-Hot Encoding**:
  - For columns with multiple categories (like make and model), we used one-hot encoding. This method created binary columns for each category. For instance, instead of having a single "Make" column with multiple car brands, we now have separate columns for each brand (e.g., "Ford", "Toyota", "Honda"), ensuring that the model treats these categories equally without implying any rank or order.

3. Feature Engineering

To enhance the dataset and make it more informative for the model, we created new features:

- **Vehicle Age**:
  - We created a new feature called vehicle age, calculated by subtracting the car’s manufacturing year from the current year. This feature helps capture the depreciation of cars over time and is a critical factor for pricing models.

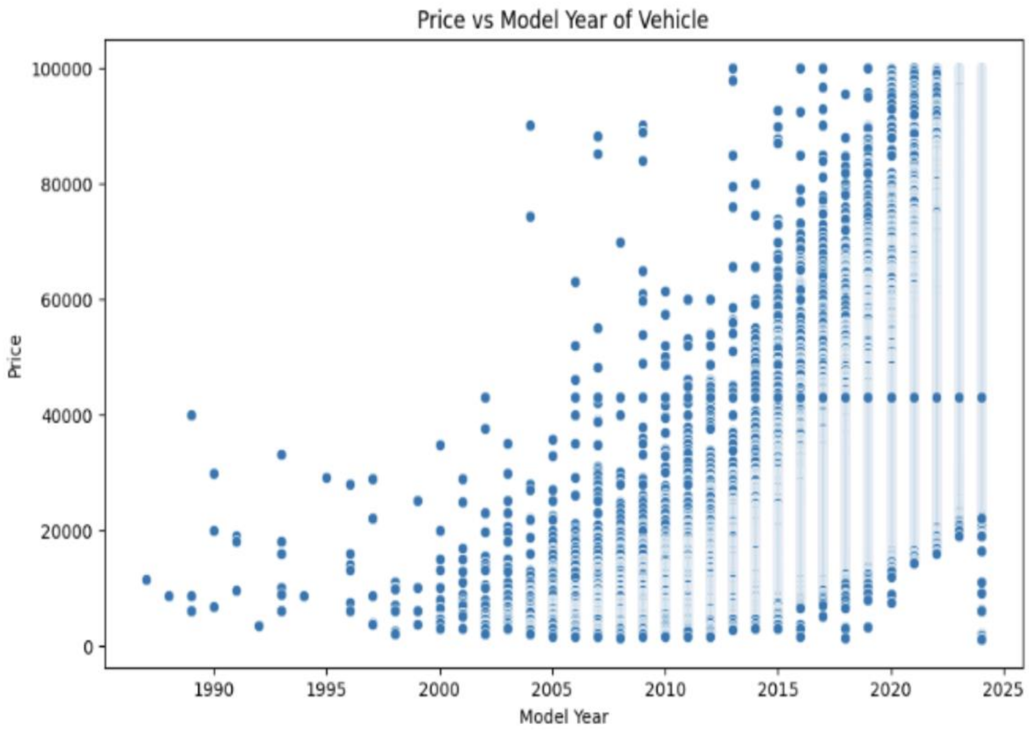
4. Initial Insights from Data

During our Exploratory Data Analysis (EDA), we uncovered some interesting trends:

- **Price vs. Model Year**:
  - We observed that older vehicles (pre-2000) had a wide range of prices, while cars manufactured after 2010 displayed a more consistent trend of increasing prices. This suggests that newer models, especially those released after 2015, tend to hold more stable pricing.

Figure 4

Price vs Model Year of Vehicle Visual

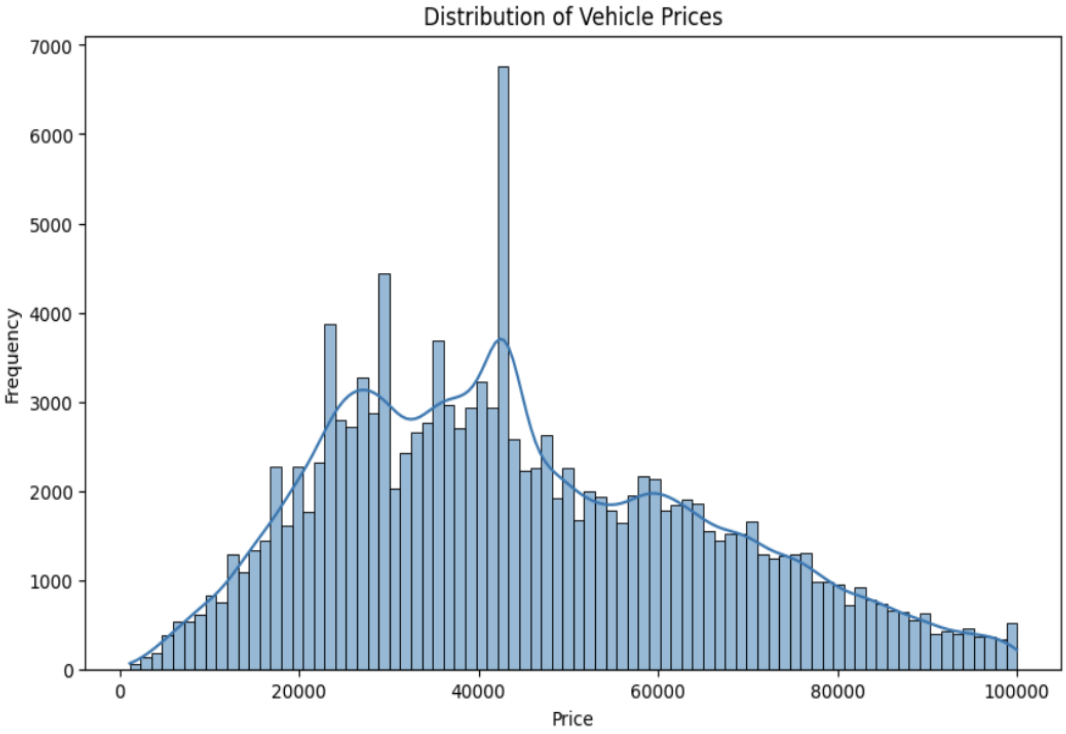


Visual demonstrating the prices compared to the model year of vehicles

- **Distribution of Vehicle Prices**:
  - A large concentration of vehicles is priced around $40,000, indicating a strong market preference for mid-range vehicles. There is still, however, a notable demand for luxury vehicles, even though fewer cars fall into this price bracket.

Figure 5

Distribution of Vehicle Prices Visual

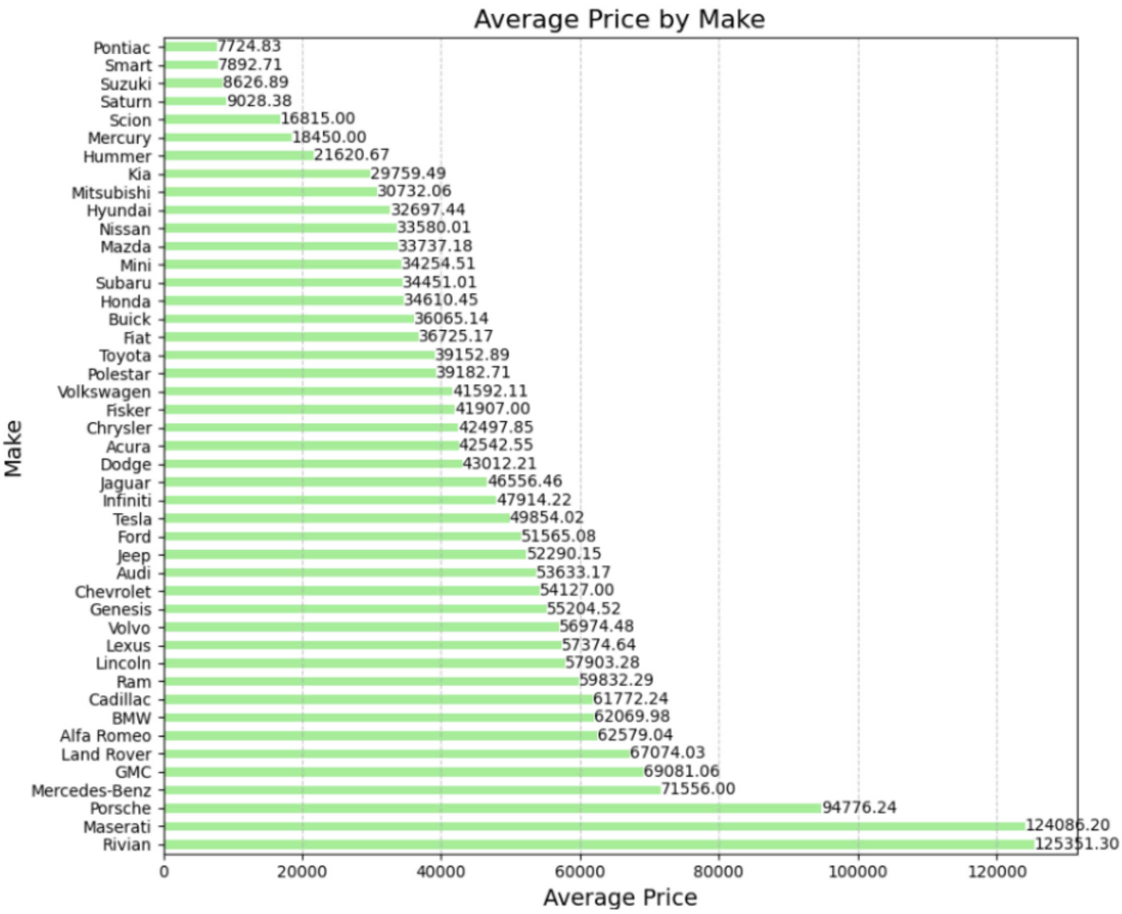


Visual showcasing the distribution of vehicle prices

- **Average Price by Make**:
  - Brands like Rivian were observed to have the highest average prices, surpassing even well-established luxury brands like Mercedes-Benz and Maserati. On the other hand, Kia, traditionally seen as a budget brand, has shown a surprising shift toward higher-priced vehicles, reflecting its move into the mid-to-upper market range.

Figure 6

Average Price by Make Visual



Bar chart demonstrating the average price of a vehicle based on its make.

Next Steps

With the data cleaned and encoded, our next focus will be on building and refining the machine learning models. Here's what's coming up:

1. **Building the Model**:
   - We will train and test a regression model to predict vehicle prices based on the key features we have engineered.
2. **Testing and Validation**:
   - We will validate the model using unseen data to ensure its reliability and adjust hyperparameters for better accuracy.

3. **Visualization and Insights**:
   - Once the model is ready, we'll present pricing predictions through graphs and charts, making recommendations on the best pricing strategies for Go Auto.

This groundwork will ensure that the final model is built on a solid, clean dataset, maximizing the chances of delivering accurate and meaningful predictions.

5. Dataset

5.1 Visualization:

Explain how you:

- Developed interactive visualizations to represent EDA findings.

  - ➢ To graphically depict the main conclusions from the dataset, I used bar charts and a heat map. These illustrations made it simple to understand how car characteristics like manufacture, transmission type, and fuel type affect cost. The interactive price pattern exploration offered by the charts helped dealerships uncover useful information.

- Visualizations include bar charts, heat maps, and geographical maps showcasing donationhotspots and high-traffic routes.

  - ➢ **Average Price by Make**: This bar chart shows the average price across different vehicle brands. Luxury brands like BMW and Mercedes are priced higher compared to brands like Ford and Honda, aligning with their market positioning.
  - ➢ **Average Price by Transmission Type**: This chart illustrates how vehicles with automatic transmissions tend to be priced higher than manual vehicles, reflecting consumer preferences.
  - ➢ **Price Stock Distribution by Price Type**: The chart highlights how new cars are generally priced higher than used and certified pre-owned vehicles, giving dealerships insights into price ranges across stock types.
  - ➢ **Correlation Heat Map**: The heat map revealed relationships between numerical attributes like price, mileage, and model year. It confirmed that newer vehicles with lower mileage are typically priced higher.
  - ➢ **Average Price by Fuel Type**: This chart shows the average price by fuel type, with electric vehicles tending to have higher prices than gasoline and diesel cars.

- Ensured visualizations were intuitive and conveyed actionable insights.

  - ➢ Clear labels, standardized color schemes, and tooltips for simple interpretation were all incorporated into the visualizations' design. The heat map revealed significant associations, while bar charts simplified category comparisons, enabling stakeholders to swiftly extract useful information from the data.

6. Challenges Encountered:

Be transparent about challenges encountered during the project phase and how they were addressed. This could include technical challenges, resource constraints, unexpected issues,model performance, real-life context, etc.

- ➢ One significant challenge was dealing with **missing or inconsistent values** in key features, such as the `leather` and `navigation` attributes, which were entirely filled with zeros. Since these attributes lacked meaningful data, we treated them as missing values. However, due to their absence across entire rows, we decided not to consider them as relevant attributes for the model.

- The Secondary challenge we faced during the project was related to **outliers in the dataset**, particularly in the `price` attribute. Some vehicles, especially **unbranded or uncommon cars**, showed extremely unreasonable prices that were not justifiable based on their attributes (e.g., mileage, model year). These outliers distorted the overall pricing patterns and could have led to inaccurate predictions.

  Addressing the Outliers:

  → **Manual Inspection**: After identifying that coding techniques like automated outlier detection were not sufficient, we manually inspected the outliers. In many cases, the inflated prices were due to data entry errors or anomalies in the listings.

  → **Filtering**: To improve the accuracy of the model, I filtered out these extreme values for certain vehicles and excluded them from the training set. This ensured the model focused on realistic pricing ranges.

7. Stakeholder Engagement:

Given our limited direct interaction with Go Auto, we rely on structured engagements through our professor and focused client interactions during demos. Here are the strategies we have used to maintain effective communication and align with the client's needs:

1. **Initial Meeting (Completed):**

   - In the initial meeting, Go Auto outlined their business objectives, particularly around optimizing vehicle pricing strategies. We gathered key insights into their expectations and challenges, which helped shape our project's direction.
   - The meeting provided a foundation for aligning our efforts with Go Auto's goals, ensuring our work remained focused on solving their specific business problems.

2. **Professor as the Key Communication Link:**

   - Since the professor serves as our main communication channel with Go Auto, we engage with him regularly to ensure that we remain aligned with the client's evolving needs.
   - We provide project updates to the professor, ensuring that any feedback from the client is integrated into our work.

3. **Demo 1: Presentation and Feedback (Completed):**

   - After completing the Exploratory Data Analysis (EDA), we presented our findings to the client during Demo 1. Key insights and feedback from this session included:

- **Use of Median and Mean for Missing Values:** Tiago, our client, was particularly interested in understanding why we chose to replace NaN values using median and mean. We provided a rationale based on the nature of the data distribution and the impact of outliers.
- **Challenges Faced:** Tiago asked about the most challenging aspects of the project so far. We shared that handling outliers and dealing with columns like has_leather and color-related attributes were difficult, especially since some colors were also described as having leather in them.
- **Next Steps - Feature Selection:** Tiago suggested that we focus on using our EDA results to determine which features are most important to include in the machine learning model. This feedback will guide our feature selection process in the model development phase.

4. **Preparing for Future Demo Presentations:**

   - Moving forward, we will incorporate Tiago's suggestions by using our EDA findings to justify feature selection in our model. We will also prepare specific questions for the client during demos to ensure we maximize the feedback opportunities.
   - Future demos will emphasize how our model aligns with Go Auto's pricing strategies and the business objectives set during the initial meeting.

5. **Continuous Feedback Loop via the Professor:**

   - We maintain regular communication with the professor to ensure that any changes in client expectations are communicated to us. This helps us stay aligned with Go Auto's needs, even between demos.

By integrating client feedback and maintaining regular touchpoints through the professor, we ensure that our project evolves in line with Go Auto's goals, with particular emphasis on addressing challenges and optimizing feature selection for the pricing model.

8. Lessons Learned:

- The data cleaning process, particularly handling missing values, went smoothly overall. Although some more efficient methods were unknown to us before the demo, everything was completed successfully. However, had we known these alternative techniques, the process could have been faster and more streamlined.
- Encoding the data presented significant challenges, and the various encoding choices were somewhat confusing. While we managed to encode the data successfully, some methods, such as using the label encoder, may not have been the most optimal choice. For future projects, a more thorough review of what needs to be encoded and a deeper understanding of available encoding methods would improve both efficiency and accuracy.
- The EDA process went well, and analyzing the data to find insights was relatively straightforward. However, the analysis did reveal the presence of outliers and indicated that some of our data cleaning methods were not fully effective. Moving forward, we should perform more frequent checks to ensure our cleaning processes are working as intended.
- While team communication was generally good, there were instances where it felt

like we were not all on the same page or were working at different stages. Improving our coordination and ensuring that all team members are kept informed and aligned will help streamline the project. Implementing better scrum practices could enhance organization and collaboration.

9. Future Recommendations:

- To improve the predictive accuracy of future models, it would be beneficial to include additional features that capture more detailed aspects of vehicle characteristics.
- Specific attributes like leather interiors, navigation systems, or other luxury features, which can significantly impact vehicle pricing, should be incorporated. However, this would likely require sourcing or creating a new dataset that captures these details, as the current dataset may not have sufficient granularity to support these additional features.
- The encoding methods used in this project, while functional, were not always the most efficient or accurate for the data at hand. For example, label encoding was used in cases where one-hot encoding or target encoding may have been more appropriate. In future projects, it would be beneficial to carefully evaluate the encoding needs for each variable before starting, ensuring the most suitable method is chosen for each feature.
- While the team worked effectively overall, there were moments where communication and collaboration could have been more streamlined. To enhance team cohesiveness, it is recommended that we adopt more structured collaboration tools (such as Trello or Asana) and enforce more consistent communication practices.
- Regularly scheduled check-ins, clear task delegation will ensure that all team members remain on the same page, avoiding instances where tasks are misaligned or work progress unevenly across the team.

10.Impact on the Community:

In this project, not only is Go Auto benefiting, but it also supports the broader community in several important ways:

- The project’s pricing model ensures that vehicles are priced appropriately for a broad range of consumers. By avoiding overpriced vehicles, more people can afford to purchase cars, improving access to reliable transportation and enhancing the overall quality of life in the community.
- By optimizing vehicle pricing through data-driven insights, this project helps ensure that consumers are paying fair prices for their vehicles. When pricing is based on accurate market data and vehicle features, customers can trust that they are receiving a competitive and reasonable price, enhancing consumer confidence in the dealership.
- As Go Auto maintains competitive pricing, it encourages more vehicle sales, which supports the local economy. More sales generate more business for the dealership, which can lead to job creation and economic growth within the community.
- Through this partnership, Go Auto is also contributing to the professional development of Norquest College students. By providing the opportunity to work on real-world business problems, they are helping students gain practical experience and preparing them for the challenges of the job market, further enriching the community’s workforce.

11. Project Conclusion:

Summarize the overall success of the project, emphasizing how it has met or exceeded initial goals and contributed to the betterment of the food donation process in Edmonton.

While we have just completed the initial phases of the project, this journey has already been a transformative learning experience for our team. The concepts we have been learning throughout our program, including machine learning, data analysis, and statistics, have come together in a practical, real-world application. We now have a much clearer understanding of how these foundational skills integrate to solve complex business problems, such as optimizing vehicle pricing.

The hands-on work we have done has allowed us to see the full picture. Concepts that once seemed abstract are now clear, as we have witnessed their direct impact on our analysis and model-building. This project has given us confidence in our ability to apply these skills moving forward.

Though we have only completed a few steps, we feel we are on the right path. Our progress so far has been encouraging, and we are excited to continue refining our model and delivering valuable insights. This project has not only strengthened our technical abilities but has also provided us with a holistic understanding of how data science can drive impactful business decisions.

Also, this project has contributed to the community in Edmonton, ensuring that the strategies we develop can be applied to real-world community challenges. By leveraging data-driven insights, we can help improve efficiency and resource allocation in processes beyond the automotive industry, offering broader societal benefits.

12. Acknowledgments:

Acknowledge the contributions of team members, stakeholders, sponsors, and any other parties involved in the project.

We would like to extend our heartfelt thanks to all those who contributed to the success of this project:

- **Go Auto**, to provide us with the opportunity and dataset to work on this impactful project. Your collaboration allowed us to apply our skills to a real-world challenge and gain invaluable experience.
- **Our client Thiago Valentin**, for offering valuable insights and feedback during the demo presentations, and for promptly answering our questions via email. We are especially grateful for your time spent coming in person and showing enthusiasm for our work. Your excitement and support have been incredibly motivating for the team.
- **Our professor Md Mahbub Mishu**, for serving as the essential link between our team and Go Auto, and for providing continuous guidance and encouragement. You allowed us the space to develop and grow within the classroom, understanding that

life's busy demands can sometimes impact on our work. The class time dedicated to this project enabled us to focus and put in the effort needed to learn and succeed.

- **The Machine Learning Department at Norquest College**, for creating a learning environment that encourages us to apply the concepts and techniques we have learned to practical business problems. Your hard work is helping this program grow, with students sharing their experiences and building a strong reputation for the program across Canada. Many students from other provinces are now applying, a testament to the program's quality.
- **Our Team Members**, for bringing together diverse skills that formed the complete puzzle. Every team member played a crucial role in driving this project forward, and our collaborative efforts have been key to our progress and success.

13.1Appendices:

Include any additional materials, charts, graphs, or data visualizations that support the information presented in the report.

Scrum Documentation for Go Auto Project: Vehicle Pricing Model

1. Project Overview

- **Project Name**: Vehicle Pricing Model
- **Product Owner**: Mehbub Mishu
- **Scrum Master / Team Leader**: Geetika
- **Development Team**:
    - **Data Analysts**: Bharat, McLaren, Geetika
    - **IT Programmer**: Renata
- **Stakeholders**: NorQuest College, Go Auto
- **Project Goals**:
    - Develop a machine learning model to predict vehicle prices based on features such as make, model, year, and mileage.

- Ensure thorough analysis, cleaning, encoding, and visualization of the dataset for effective modeling.
- Provide insights and visual representations to help in model development and decision-making.

2. Sprint Planning Documentation

Sprint Overview

Sprint Duration: 2 weeks (Sept 10, 2024 – Sept 24, 2024)
Sprint Goal: Analyze and clean the dataset to prepare it for encoding and model development in subsequent sprints.

Sprint Backlog

User Story	Story Points	Priority	Assigned To
US1: Analyze the Vehicle Dataset	9	High	Geetika, Bharat, McLaren, Renata
US2: Clean the dataset	12	High	Geetika, Bharat, McLaren

User Stories in Detail

User Story 1: Analize the Vehicle Dataset

Description: As data analysts, the whole team will perform a detailed analysis of the vehicle dataset to understand its structure, identify trends, and detect potential issues such as missing values or outliers.

Acceptance Criteria:

Generate summary statistics (mean, median, standard deviation) for key features like price, mileage, year

Detect and report missing values and null entries in the dataset.

Identify outliers and anomalies

.

Definition of Done: The dataset analysis is completed, with a report summarizing insights on data quality, missing values, and anomalies, ready for cleaning.

User Story 2: Clean the Dataset

Description: As data analysts, Geetika, Bharat, and Mcleran will clean the dataset by handling missing values, removing duplicates, and correcting any erroneous data.

Acceptance Criteria:
All missing values are either filled appropriately or removed.

Remove duplicates from the dataset.

Correct any inconsistent data (e.g., negative values for price or mileage).

Definition of Done: The dataset is cleaned, with all issues addressed, and ready for encoding and further processing.

3. Sprint Execution Documentation (Sprint 1)

Daily Stand-ups

Daily stand-ups are conducted to track progress, identify blockers, and plan the day's work. The standard questions asked are:

From Day 3:

Geetika: "Completed an overview of the dataset, identified missing values and anomalies. Will begin working on detecting outliers today."

Bharat: "Reviewed summary statistics and checked data consistency. No blockers."

McLaren: "Analyzed vehicle make and model data, will collaborate with Bharat on cleaning tasks."

Sprint Burndown Chart

Date	Story Points Remaining
Day 1	21
Day 3	16
Day 5	12
Day 7	10
Day 9	5
Day 14	0

Sprint 2 Overview

Sprint Duration: 2 weeks (Sept 10, 2024 – Sept 24, 2024)
Sprint Goal: Perform data encoding and create visualizations to facilitate model development in future sprints.

Sprint Backlog

User Story	Story Points	Priority	Assigned To
US1: Perform Data Encoding	12	High	Geetika, Bharat, McLaren, Renata
US2: Create Data Visualisation	10	High	Geetika, Bharat, McLaren

User Stories in Detail

User Story 1: Perform Data Encoding

Description: As data analysts, Geetika, Bharat, and McLaren will perform encoding on categorical features in the dataset to prepare it for model training.

Acceptance Criteria:

Apply one-hot encoding for categorical features.

Apply label encoding for ordinal features.

Ensure all encoded data is compatible with the model requirements.

Definition of Done: All categorical data is encoded and ready for the next steps in the model development process.

User Story 2: Create Data Visualizations

Description: The team will generate visualizations to help stakeholders understand key patterns in the dataset and support the model development process.

Acceptance Criteria:

Generate visualizations for key features (e.g., price distribution by make and model, mileage over time).

Ensure visualizations are easy to interpret for stakeholders.

Definition of Done: Visualizations are created and presented in a clear format, ready for review by stakeholders.

Sprint Execution Documentation (Sprint 2)

Daily Stand-ups

Daily stand-ups are conducted to track progress, identify blockers, and plan the day's work. The standard questions asked are:

From Day 3:

Geetika: Completed the encoding for categorical variables. Will start working on visualizations.

Bharat: Finished one-hot encoding. I am working on label encoding now. No blockers.

McLaren: Creating visualizations for the data analysis phase. Collaborating with Bharat for insights.

Renata: Assisting with visualizations and ensuring data is properly formatted for display

Sprint Burndown Chart

Date	Story Points Remaining
Day 1	22

Day 3	17
Day 5	12
Day 7	10
Day 9	5
Day 14	0

4. Sprint Review

Sprint Review Agenda:

Date: Oct 8, 2024
Participants: Scrum team, product owner, key stakeholders (Go Auto, Norquest College)
Objective: Demonstrate the combined outcomes of Sprint 1 and Sprint 2, focusing on dataset analysis, cleaning, encoding, and visualizations.

Demo:

- The team presented comprehensive work completed across both sprints. The following key components were demonstrated:

Dataset Analysis and Cleaning:
The analysis performed in Sprint 1 provided a thorough understanding of the dataset's structure, highlighting key statistics and any anomalies present in the data.
The team showcased how missing values were handled and duplicates were removed, ensuring data quality and readiness for subsequent processing.

Data Encoding:
The encoding process for categorical variables was demonstrated, detailing the application of one-hot encoding for nominal features and label encoding for ordinal features.
The team emphasized how these transformations prepared the dataset for the machine learning model.

Data Visualizations:
A series of visualizations were presented, illustrating key trends and patterns within the vehicle dataset. This included visual representations of price distributions, vehicle features, and any identified correlations.
The visualizations aimed to provide stakeholders with clear insights to support decision-making and model development.

Feedback:

Mehbub Mishu (Product Owner):
No feedback given

Go Auto Stakeholder:
Plan out what features are going to be used for the model

NorQuest College Stakeholder:
No feedback given

Next Steps:

Incorporate feedback from stakeholders regarding visualizations and explore additional segmentation in the next sprint.

Proceed with the development of the vehicle price prediction model, building on the groundwork laid in the analysis, cleaning, encoding, and visualization phases.

5. Sprint Retrospective

What Went Well:

- The initial analyzing of the dataset went smoothly, and the team reached a consensus on what columns could be dropped and what had to be edited.
- Cleaning and EDA also went well with very minor issues
- Data visualization also went fairly well. No issues on that end.

What Didn't Go Well:

- Encoding was very difficult at the start. The team was unsure of what methods to use and how to implement them, but the issues were ironed out approaching the demo.
- Team wasn't the most cohesive at times. Communication and response times varied, and it felt like some of the team lagged or were too far ahead at times.

Improvements for Next Sprint:

- Better team cohesiveness and presence to avoid any issues that may result from the team not being on the same page.
- More research and testing with encoding to ensure the choices made were the best option and the team isn't unsure or second-guessing themselves.

6. Product Backlog

Backlog Item	Priority	Story Points	Status
Revisit data encoding techniques to ensure the best methods are used.	High	10	To Do
Begin model testing once encoding finishes.	High	20	To do

7. Definition of Done (DoD)

The Definition of Done outlines the criteria that must be met for a user story or backlog item to be considered complete.

- Code is written Code is complete and follows coding standards.
- Code is tested: Unit tests and integration tests have passed.
- Code is reviewed: Peer-reviewed and merged into the main branch.
- Feature is documented: Appropriate documentation for end users and stakeholders.
- Approved by Product Owner: Final sign-off from the Product Owner.

14.References

Cite your sources (MUST follow APA style).

Canadian Black Book. (2024). *CBB Listings [Data set]*. Go Auto.

Go Auto. (2024). *Project: Driving insights: Predictive modeling, market analysis, and data visualization for vehicle sales optimization for Go Auto* [PowerPoint slides]. CMPT 3830: Machine Learning Work Integrated Learning-1, Fall 2024, NorQuest College.