# CMPT 3830: Project Proposal

## 1. Project Title:

**"Optimized Vehicle Mileage Classification Using Machine Learning: A strategic Solution for Go Auto"**

## 2. Project Overview:

### - Objective:

To develop a sophisticated machine learning classification model that segments vehicle into distinct mileage bands (low, medium, high) to enable Go Auto dealerships to strategically align marketing and sales efforts with customer preferences. This initiative aims to empower Go Auto with actionable insights that enhance operational efficiency, improve customer engagement, and drive revenue growth.

### - Background:

As the automotive industry becomes increasingly data driven, Go Auto seeks to optimize its inventory management and marketing strategies. Mileage is a critical factor influencing customer decisions, yet current methods of categorization lack scalability and precision. Leveraging the power of machine learning, this project aims to streamline the classification process, enhance the accuracy of mileage-based grouping, and provide visual insights that facilitate informed decision-making.

### - Scope:

This project includes:

- Comprehensive data preprocessing to ensure dataset quality.
- Exploratory data analysis (EDA) to uncover pattern and insights.
- Development, evaluation, and optimization of a multi-class classification model **using state-of-the-art machine learning algorithms.**
- Creation of visuals dashboards for stakeholders to interpret results effectively.

Exclusions:

- Advanced customer segmentation beyond mileage preferences.
- Recommendations regarding dealership operations or hardware improvements.

## 3. Project Deliverables:

1. **Data Preparation Report:** Includes data cleaning methodologies, summary statistics, and resolved quality issues.

2. **EDA Insights Document:** Provides visualization and key findings about trends and patterns in the data.

3. **Trained Model and Codebase:** A well-documented machine learning model that categorizes vehicle into mileage groups with associated code and reproducibility guidelines.

4. **Visualization Dashboard:** Interactive dashboards summarizing model performance, classification outputs, and insights for non-technical stakeholders.

5. **Final Presentation:** A professional presentation delivered to stakeholders highlighting the methodology, results, and recommendations.

## 4. Project Timeline:

| Milestone | Completion date | Details |
|---|---|---|
| Dataset Acquisition | January 16, 2025 | Receive the dataset from Go Auto. Perform and initial review to understand structure and contents, identifying challenges. |
| Problem Understanding | January 20, 2025 | Define the problem, identify objectives and the target variable, and understand business implications. |
| Team charter submission | January 23, 2025 | Submit the team charter outlining roles, responsibilities, and the overall project workflow. |
| EDA (Exploratory Data Analysis) | January 30, 2025 | -Perform summary statistics.<br>-Visualize trends, patterns, and anomalies.<br>-Analyze the target variable's class distribution.<br>-Document insights. |
| | | |

| | | |
|---|---|---|
| Data Processing | February 3, 2025 | -Handle missing values<br>-split the dataset into training and testing sets<br>-encode categorical data<br>-scale or normalize numerical features |
| Feature Engineering | February 7, 2025 | Select relevant features and create new features using domain knowledge or statistical methods to enhance data readiness and improve model performance. |
| Demo 1: EDA and feature engineering presentation | February 13, 2025 | Select relevant features and create new features using domain knowledge or statistical methods to enhance data readiness and improve model performance. |
| Model Selection | February 17, 2025 | Evaluate and select classification algorithms (e.g., Logistic regression, random forest, SVM) based on interpretability, complexity, and computational requirements. |
| Model Training | February 20, 2025 | Train selected models using the preprocessed and engineering data. Perform cross-validation to ensure robust results. |
| Model Deployment | February 25, 2025 | Deploy the trained model on a platform like flask, Django, or a cloud services. Test deployment for scalability and ensure security. |
| Phase 1 Report submission | February 27, 2025 | Submit a detailed report covering EDA, preprocessing, feature engineering, model training, and deployment |
| Model Optimization | March 5, 2025 | Perform hyperparameter tuning using grid search. |

| | | |
|---|---|---|
| | | Experiment with ensemble methods to improve performance. |
| Demo 2: ML Modeling and deployment presentation | March 13, 2025 | Present optimized model results, deployment process, and classification insights. Include challenges and solutions encountered during these phases. |
| Data visualization Using Power BI | March 16, 2025 | Develop interactive dashboards to present key EDA insights, feature importance, and model performance. |
| Phase 2 Report Submission | March 19, 2025 | Submit a comprehensive report summarizing modeling optimization, deployment, and insights derived from the classification task. |
| ML Application for prediction/classification | March 22, 2025 | Test the deployed ML model for real-world classification task. Ensure it performs accurately and aligns with business requirements. |
| Final presentation | April 7, 2025 | Present the complete project lifecycle, including all phases: EDA, preprocessing, modeling, deployment, and Power BI dashboard. Submit all deliverables. |

## 6. Project Plan:

### - Tasks and Activities:

| Task | Owner(s) | Due Date | Details |
|---|---|---|---|
| Dataset Acquisition and initial review | Aditya Mehta & Harvir Kaur | January 16, 2025 | Obtain the dataset, verify structures and contents, and identify potential challenges. |

| | | | |
|---|---|---|---|
| Problem Understanding and definition | Aditya Mehta & Inderjeet Singh | January 20, 2025 | Define the problem, set objectives, specify the target variable, and clarify real-world business implications. |
| Team Charter Submission | Inderjeet Singh | January 23, 2025 | Compile and submit the team charter outlining roles, responsibilities, and overall project workflow |
| Project Charter submission | Inderjeet Singh | January 30, 2025 | Document the problem definition, objectives and project scope. Submit the finalized project charter. |
| Data cleaning and preparation | Harvir Kaur & Ishmeet Singh | February 5, 2025 | Clean and preprocess raw data by handling missing values, duplicates, and inconsistencies. Ensure data quality. |
| Exploratory data Analysis | Aditya Mehta &Inderjeet Singh | February 10, 2025 | Perform summary statistics, visualize patterns, and analyze the target variable's distribution. Document findings. |
| Feature Engineering | Harvir Kaur & Inderjeet Singh | February 12, 2025 | Create and select relevant features to enhance predictive performance. Document rationale behind feature selection. |
| Demo 1: EDA and feature Engineering presentation | All Team members | February 13 2025 | Prepare and present insights, trends, and new features derived from EDA and feature engineering. |

|  |  |  |  |
|---|---|---|---|
| Model Selection | Aditya Mehta & Inderjeet Singh | February 17, 2025 | Evaluate and select classification algorithms based on complexity, interpretability, and computational requirements. |
| Model training | Aditya Mehta & Inderjeet Singh | February 20, 2025 | Train selected models using the preprocessed and engineered data. Document initial performance metrics. |
| Model Deployment | Inderjeet Singh & Harvir Kaur | February 25, 2025 | Deploy the trained model using Flask, Django, or cloud services. Test for scalability, reliability, and security. |
| Phase 1 Report compilation and submission | Inderjeet Singh | February 27, 2025 | Summarize EDA, feature engineering, model selection, and deployment findings for phase 1 report submission. |
| Model Optimization and validation | Aditya Mehta and Ishmeet Singh | March 5, 2025 | Tune hyperparameters and validate models, experiment with ensemble methods to improve performance. |
| Demo 2: ML Modelling and deployment presentation | All team members | March 13, 2025 | Present optimized model results, deployment process, and insights from classification tasks. |
| Data Visualization Using Power BI | Aditya Mehta & Harvir Kaur | March 16, 2025 | Build interactive dashboard to showcase EDA insights, feature |

| | | | |
|---|---|---|---|
| | | | importance, and model performance. |
| Phase 2 report compilation and submission | Inderjeet Singh | March 19, 2025 | Document modeling results, optimizations, and business recommendations for phase 2 submission. |
| ML Application for prediction/classification | Aditya Mehta & Inderjeet Singh | March 22, 2025 | Test the deployed model for real-world prediction and classification tasks. Validate its accuracy and usability. |
| Final presentation preparation | Team Collaboration | April 15, 2025 | Create a polished presentation summarizing the full project lifecycle, including Power BI dashboards. |
| Final Presentation | All team members | April 17, 2025 | Deliver a comprehensive presentation covering all phases, from problem understanding to deployment and insights. |

## 7. Resources Required:

| Resources | Description | purpose | Estimated Cost |
|---|---|---|---|
| Python libraries | Scikit-learn, pandas, NumPy, Matplotlib, Seaborn, Plotly, TensorFlow/keras, PyTorch, XGBoost, LightGBM, Flask, OpenPyXL | -Scikit-learn: for preprocessing and machine learning models.<br>-Pandas: for data manipulation and analysis.<br>-NumPy: For numerical operations.<br>-<br>Matplotlib/Seaborn/Plotly: | No external cost (Open sources) |

| | | | |
|---|---|---|---|
| | | for visualizations (static and interactive).<br>-TensorFlow/Keras, PyTorch:  for advanced deep learning models (if required).<br>-XGBoost/LightGBM: For gradient boosting techniques in classification.<br>-Flask: for model deployment as a web application.<br>-OpenPyXL: For working with Excel files (if needed for data export) | |
| Development environment | Google Colab | For coding, testing, and running machine learning workflows using cloud-based computational resources. | Free (pro-option: $10/month) |
| Cloud computing services | Google Colab Pro (optional) | Provide enhanced resources like faster GPUs/TPUs and longer session durations | $100/month (optional) |
| Visualization tools | Power BI | For creating interactive dashboards to present insights from EDA, feature engineering, and modeling results. | No Cost (Educational License) |
| Data Storage | Google Drive (integrated with Colab) | To securely store large datasets and ensure easy access for collaboration workflows. | Free (15 GB) or $10monthly for extra storage |
| Documentation tools | Microsoft Word, Google Docs, Microsoft presentations | To create detailed project charter, reports, and final documentation, presentations. | No Cost (Educational tools) |
| Deployment Frameworks | Flask or Django | To deploy the trained machine learning model as a web-based or API application. | No external cost (open source) |
| Monitoring Tools | Google Colab Logs, Cloud Monitoring | To monitor resources utilization, model | Included in Google Colab Pro (optional) |

| | | | |
|---|---|---|---|
| | | performance, and deployment metrics | |

## 8. Risk Management Plan:

| Risk | Likelihood | impact | description | Mitigation strategy |
|---|---|---|---|---|
| Delayed Data access | Medium | High | Potential delays in acquiring the dataset from the data provider, impacting project timelines. | -initiate early communication with the data provider.<br>-Establish clear deadlines for data delivery.<br>-Prepare a backup plan, such as using synthetic or public dataset for initial testing. |
| Data Quality | High | Medium | Challenges with missing, inconsistent, or noisy data that may require extensive cleaning efforts. | -Allocate Extra time in the project schedule for data preprocessing.<br>-Develop a structured approach for handling missing or inconsistent data.<br>-use automated data quality checks and validation scripts. |
| Computational Resources constraints | Medium | High | Insufficient computing power for model training | -utilize Google Colab Pro for enhanced |

| | | | | |
|---|---|---|---|---|
| | | | and testing, especially for large datasets or complex models. | computational resources.<br>-stagger model training workloads to optimize resource usage.<br>-Monitor resource utilization and adapt workflow accordingly. |
| Skill gaps in advanced machine learning | Low | Medium | Limited expertise in advanced ML techniques (e.g., hyperparameter tuning, ensemble methods). | -conduct internal training sessions and workshops.<br>-Assign tasks based on team members' strengths.<br>-encourage the use of online resources, such as documentation and tutorials, to bridge knowledge gaps. |
| Model Deployment challenges | Medium | High | Difficulties in deploying the trained model due to compatibility issues or lack of deployment experience. | -Choose a lightweight and flexible deployment framework (e.g., Flask or Django).<br>-Conduct early deployment tests to identify issues.<br>-Include deployment-specific resources in team training sessions. |

| | | | | |
|---|---|---|---|---|
| Data Security Breaches | Low | High | Risk of unauthorized access, data leaks, or breaches during processing, storage, or transmission. | -Use secure platforms like Google Drive with access controls.<br>-Encrypt sensitive data during storage and transmission.<br>-Regularly audit and update security measures. |
| Data Loss | Medium | Severe | Risk of accidental deletion or corruption of critical data files during project workflow | -Use automated backups for datasets and project files.<br>-Maintain version control for scripts and data files using platforms like Git,<br>-Implement recovery protocols for lost or corrupted files. |
| Timeline Overruns | Medium | High | Risk of missing deadlines due to unforeseen challenges or bottlenecks in the workflow. | -Use agile practices to monitor progress and adjust priorities.<br>-schedule regular team check-ins by scrum master to identify and address delays early.<br>- maintain buffer period for critical tasks. |

| | | | | |
|---|---|---|---|---|
| Inaccurate model prediction | Medium | Medium | Risk of the model not meeting performance expectations due to insufficient features engineering or training. | -Conduct thorough features engineering and hyperparameter tuning.<br>-use cross-validation to ensure robustness.<br>-iterate on model improvements based on performance metrics. |

## 9. Budget:

While the project primarily utilizes free and open-source software, minor costs are associated with optional cloud storage and computing resources:

- **Google Drive Storage:** $15/month for 3 months = $45
- **Cloud computing:** $100/month for 2 months = $ 200
- **Total Estimated cost:** $ 350