214243P - Assignment Report

Project Title: Forecasting Monthly New Vehicle Registrations by Category in Sri Lanka

---

1. Problem Definition & Dataset Collection

1.1 Problem Description and Relevance

In the evolving transportation landscape of Sri Lanka, monthly vehicle registration volumes are subject to high volatility due to economic crises, import policy changes, and shifting consumer demand. While historical registration data is available, there is a lack of predictive tools that can anticipate future registration trends based on current market conditions. This project addresses the gap between historical record-keeping and proactive transportation planning by developing a model to forecast monthly new vehicle registrations across 9 vehicle categories.

1.2 Data Source and Collection

The dataset was compiled from official monthly vehicle registration records issued by the Department of Motor Traffic (DMT), Sri Lanka. This is a primary official source, ensuring high reliability and local relevance.

1.3 Features and Target Variable

Target Variable: New_Registration (Monthly count of newly registered vehicles per category).

- Original Features:
    - Standard_Category: Vehicle type (e.g., Motor Car, Motor Cycle, Bus)
    - Year: Registration year (2018–2025)
    - Month: Registration month
    - Transfer: Monthly vehicle ownership transfers
    - Yearly_Total_Stock: Cumulative registered vehicle fleet size
- Engineered Features:
    - Month_Num: Numeric encoding of month (1–12)

- Quarter: Derived fiscal quarter (1–4)
- Is_Peak_Season: Binary flag for high-demand months
- Is_Crisis_Period: Binary flag for economic downturn periods (2020–2022)
- Transfer_to_New_Ratio: Ratio of transfers to new registrations (secondary market activity)
- New_Registration_Market_Share: Category's share of total monthly registrations
- Prev_Month_New_Reg: Previous month's registration count (lagged variable)
- Monthly_Growth_Rate: Month-over-month percentage change in registrations

1.4 Dataset Size and Preprocessing

Size: 648 monthly records across 9 major vehicle categories (Motor Cycle, Motor Car, Three Wheeler, Bus, Lorry, Dual Purpose, Tractor, Prime Mover, Other) spanning 2018–2025.

Preprocessing:

- Data Integration: Consolidated monthly registration data from multiple annual reports into a single master dataset.
- Cleaning: Removed incomplete records and standardized category names across reporting periods.
- Imputation: Handled missing values using category-based mean imputation for numeric features.
- Feature Engineering: Created 8 new features based on temporal patterns, market dynamics, and economic conditions.
- Target Transformation: Applied log-transformation (log1p) to the target variable to reduce dominance of high-volume categories, enabling balanced cross-category accuracy.

1.5 Ethical Considerations

This project utilizes anonymized public government data from the Department of Motor Traffic. No personal, sensitive, or identifiable information (such as vehicle owner details) was used, ensuring compliance with ethical research standards.

2. Selection of a New Machine Learning Algorithm

2.1 Selected Algorithm: CatBoost (Categorical Boosting)

The selected algorithm is **CatBoost**, a high-performance gradient boosting framework.

2.2 Justification

- Not Covered in Lectures: This algorithm was not part of the standard curriculum, satisfying the "new algorithm" requirement.
- Handling Categoricals: Unlike standard models like Random Forest or k-NN, CatBoost uses a specialized ordered target encoding algorithm to handle categorical features (like "Standard_Category") without requiring manual One-Hot Encoding, making it ideal for distinguishing between 9 vehicle types.
- Robustness: It utilizes "Ordered Boosting" to overcome gradient bias, making it highly effective for smaller, high-quality tabular datasets like this one (648 records).
- Log-Transform Compatibility: CatBoost's RMSE loss function works effectively with log-transformed targets, enabling the model to optimize for relative errors rather than absolute errors — critical when vehicle registration volumes range from ~17 vehicles/month (Prime Mover) to ~17,000 vehicles/month (Motor Cycle).
- Superior Performance: A benchmark comparison against 6 other algorithms on the same dataset confirmed CatBoost as the best choice:

Algorithm	MAE (vehicles)	R^2 Score
CatBoost (Optimized + Log-Transform)	160.10	0.9842
Random Forest	407.22	0.9249
CatBoost (Baseline)	469.31	0.9415
Gradient Boosting (sklearn)	498.12	0.8666
k-NN (k=5)	1,008.40	0.4435
SVR (RBF)	1,635.12	0.2128
Linear Regression	2,238.79	0.3229

The optimized CatBoost model achieved a 60.7% lower MAE than the next best algorithm (Random Forest) and a 74.5% lower MAE than standard Gradient Boosting, demonstrating

the significant advantage of combining CatBoost with Bayesian hyperparameter optimization and log-target transformation.

3. Model Training and Evaluation

3.1 Methodology

- Data Split: The data was split into 80% training and 20% testing sets (518 training, 130 testing).
- Target Transformation: A log-transformation (log1p) was applied to the target variable to prevent high-volume categories (Motor Cycle, Three Wheeler) from dominating the loss function, enabling balanced accuracy across all 9 vehicle categories.
- Hyperparameter Optimization: Bayesian Optimization was performed over 80 trials using Optuna, yielding optimal parameters: learning rate of 0.0306, 1,876 iterations, tree depth of 4, L2 regularization of 0.989, subsample ratio of 0.734, column subsample of 0.667, random strength of 5.756, and minimum data in leaf of 3.
- Loss Function: RMSE was used on the log-transformed target, which effectively optimizes for percentage errors rather than absolute errors.
- Metrics: The model was evaluated using Mean Absolute Error (MAE) to represent the average vehicle count deviation, R-squared (R^2) to measure variance explained, and Median Absolute Percentage Error (MedAPE) to provide a scale-independent accuracy measure across categories.

3.2 Results Obtained

The model achieved an R^2 score of 0.9842 on the test set, indicating that the engineered features (previous month registrations, market share, fleet size, economic conditions) are highly predictive of monthly registration volumes. The test set MAE is 160.10 vehicles/month, with a Median Absolute Percentage Error of just 8.9%, meaning predictions are typically within 9% of the actual registration count. Per-category performance demonstrates balanced accuracy: Bus (MAE=10), Dual Purpose (MAE=21), Lorry (MAE=29), Motor Car (MAE=182), Motor Cycle (MAE=666), Prime Mover (MAE=7), and Tractor (MAE=61). The log-transform approach reduced MAE by 47.8% compared to the baseline model (306.68 → 160.10), proving its effectiveness in handling the extreme scale differences across vehicle categories. This low error margin is reliable for real-world transportation planning applications in Sri Lanka.

4. Explainability & Interpretation

4.1 Explainability Method: SHAP (Shapley Additive explanations)

- To ensure the model is not a "black box," the SHAP method was applied using TreeExplainer to interpret individual predictions in real time.

4.2 Interpretation

- Feature Influence: SHAP analysis reveals that Prev_Month_New_Reg (previous month's registrations) is the strongest predictor, followed by New_Registration_Market_Share (category's market share) and Yearly_Total_Stock (total fleet size). This confirms that recent registration momentum and relative market position are the primary drivers of future registrations.
- Domain Alignment: The model correctly learned that the Is_Crisis_Period feature has a strong negative impact during 2020–2022, aligning with Sri Lanka's known economic crisis when vehicle imports were severely restricted (Motor Car registrations dropped from 6,000+/month in 2018 to just 83/month in 2022). The recovery pattern in 2025 (Motor Car rising from 111 in January to 12,710 in December) is also well-captured through the Monthly_Growth_Rate and Prev_Month_New_Reg features.
- Real-Time Explainability: Each prediction on the web application displays a SHAP-based horizontal bar chart showing the top 5 influencing factors with human-readable labels, enabling users to understand why the model forecasts a particular registration volume for their selected category and time period.

5. Critical Discussion

5.1 Limitations and Data Quality

A primary limitation is the dataset size (648 records across 9 vehicle categories from 2018–2025), which may not fully capture extreme black-swan economic events beyond what occurred during Sri Lanka's 2020–2022 crisis period. Furthermore, data quality is dependent on the accuracy of the Department of Motor Traffic's monthly reporting and category classification standards. The model's strong performance on test data (R^2 = 0.9842, MAE = 160.10 vehicles) suggests that the available historical data captures stable market patterns well, but unprecedented external shocks such as sudden import policy changes, new taxation regimes, or global supply chain disruptions could affect prediction accuracy.

Additionally, the dataset has a notable gap (no 2023–2024 data), which means the model bridges from crisis-period patterns directly to the 2025 recovery, potentially missing transitional market dynamics. The extreme scale variance across categories (Motor Cycle ~17,000/month vs. Prime Mover ~17/month) was mitigated through log-transformation, but low-volume categories inherently have higher relative prediction uncertainty despite low absolute MAE values.

5.2 Ethical and Real-World Impact

While the model provides transparency through real-time SHAP analysis on every prediction, there is a risk of misuse if used by vehicle importers or dealers to artificially manipulate inventory levels or pricing strategies based on anticipated registration volumes. Ethically, this tool should remain a public asset for transportation planning, policy-making, and infrastructure development. The high accuracy (98.42% variance explained) necessitates responsible deployment to prevent market manipulation by sophisticated actors for example, a dealer anticipating a surge in Motor Car registrations could stockpile inventory to inflate prices. The model also carries the risk of self-fulfilling prophecy: if policymakers use predicted low registration months to relax import duties, the resulting increase in registrations could invalidate the original forecast. Furthermore, the crisis period data (2020–2022) reflects extreme economic conditions unique to Sri Lanka, and care must be taken not to over-generalize these patterns to future economic scenarios. Responsible use guidelines are provided on the application's Explainability page to promote transparency and appropriate interpretation of forecasts.