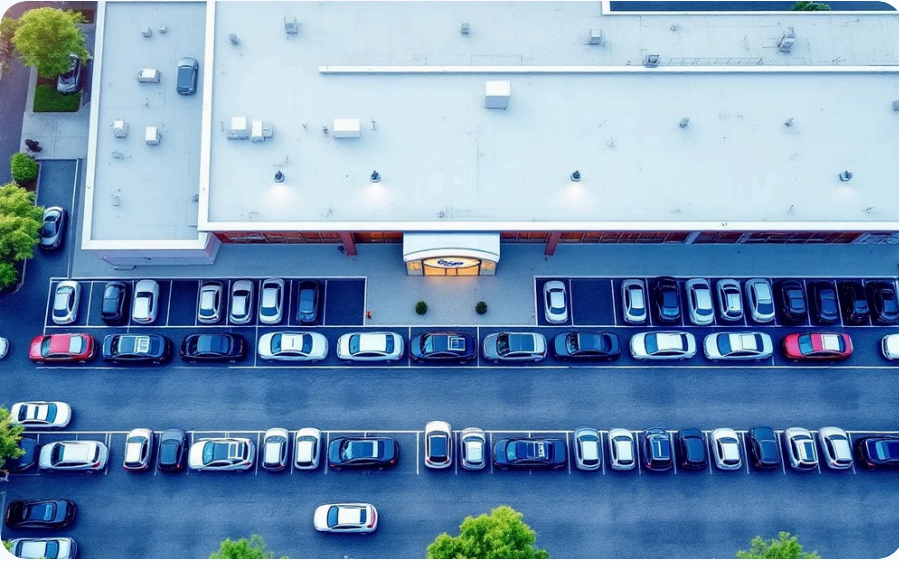# Car Sales Forecasting

Leveraging Machine Learning to Optimize Inventory and Maximize Profitability

Date: July 26, 2025

Organization: Data Science Institute – University of Toronto

Presented by: ML Team 3

Made with GAMMA

Business Problem & Objective

The Challenge

A large dealership network struggled with accurately forecasting monthly sales in highly competitive markets.

Our Solution

Our machine learning model predicts the **top 3 best-selling** vehicles each month using historical sales data and car attributes. This solution helps dealers to:

- Stock the right vehicles
- Reduce overstock inventory
- Increase overall profitability

Dataset Overview

Data Source

Car Sales Report dataset from Kaggle + 3 synthetic features

~28,000 records over 24 months (Jan 2022 – Dec 2023)

Key Features

- Car specifications
- Buyer information
- Dealer region
- Family Size (synthetic)
- Gas Mileage (synthetic)
- Crash Test Score (synthetic)

Limitations

- **Limited data volume**: ~28K samples limit deep modeling.
- **Severe imbalance**: Only 7% labeled "best sellers".
- **Compute constraints**: Limited to models with reasonable training times.

Modeling Approach

Data Preparation

Separated numeric and categorical features

Addressed class imbalance via weighting

Model Selection

- Logistic Regression

- Random Forest

- XGBoost

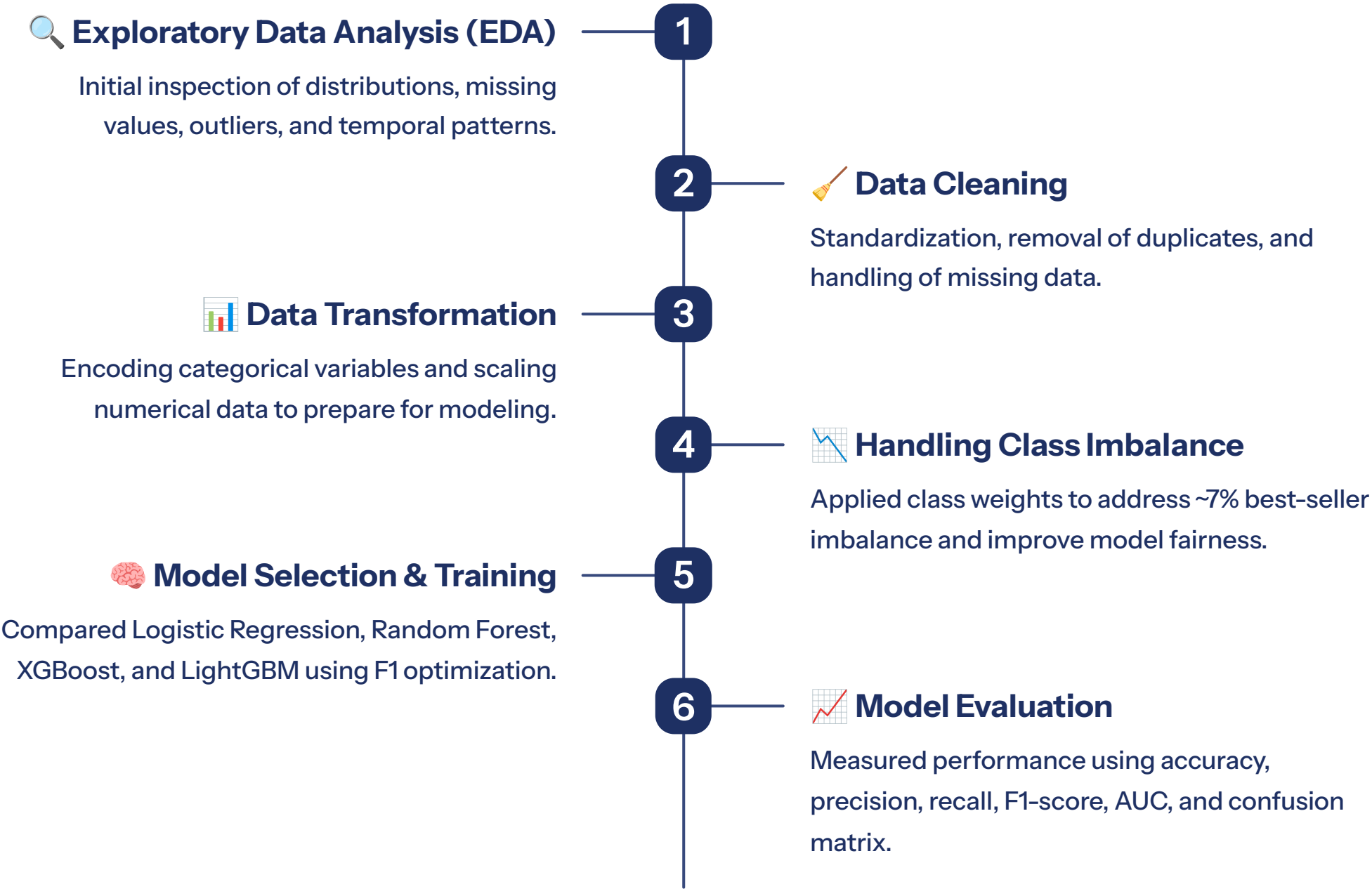- LightGBM

Optimization

Grid search with F1-score as optimization metric

Test data = last 3 months to simulate real-world forecasting

Analysis Methods and Tools

📌 Methodology Overview

🔍 **Exploratory Data Analysis (EDA)** — **1**

Initial inspection of distributions, missing values, outliers, and temporal patterns.

2 — 🧹 **Data Cleaning**

Standardization, removal of duplicates, and handling of missing data.

📊 **Data Transformation** — **3**

Encoding categorical variables and scaling numerical data to prepare for modeling.

4 — 📈 **Handling Class Imbalance**

Applied class weights to address ~7% best-seller imbalance and improve model fairness.

🧠 **Model Selection & Training** — **5**

Compared Logistic Regression, Random Forest, XGBoost, and LightGBM using F1 optimization.

6 — 📈 **Model Evaluation**

Measured performance using accuracy, precision, recall, F1-score, AUC, and confusion matrix.

🛠️ Tools Used

Pandas

Utilized for powerful data manipulation and analysis, especially with structured data.

NumPy

Leveraged for numerical computing, providing support for arrays and mathematical functions.

Matplotlib & Seaborn

Employed for creating static, interactive, and animated visualizations to explore data patterns.

Scikit-learn

The primary library for implementing various machine learning algorithms and predictive modeling.

Model Evaluation

Performance Metrics

Model: **XGBoost (final selected)**

- Accuracy: 0.9513

- Precision: 0.5534

- Recall: 0.8411

- F1 Score: 0.6676

> 🗋 Final model (XGBoost) selected based on balanced performance across precision and recall. High recall ensures that most best sellers are correctly identified, supporting inventory planning decisions.

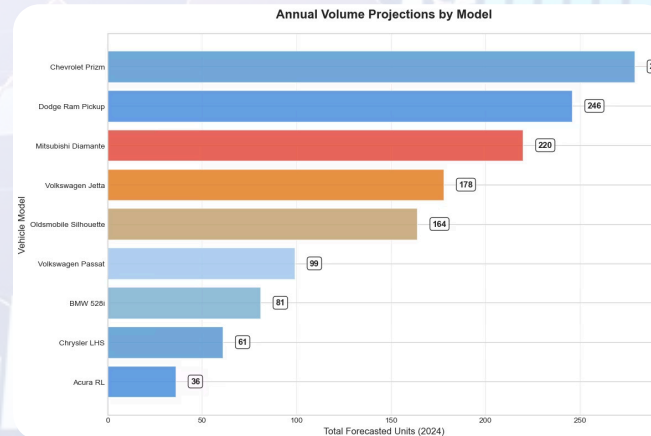Confusion Matrix: Actual vs Predicted
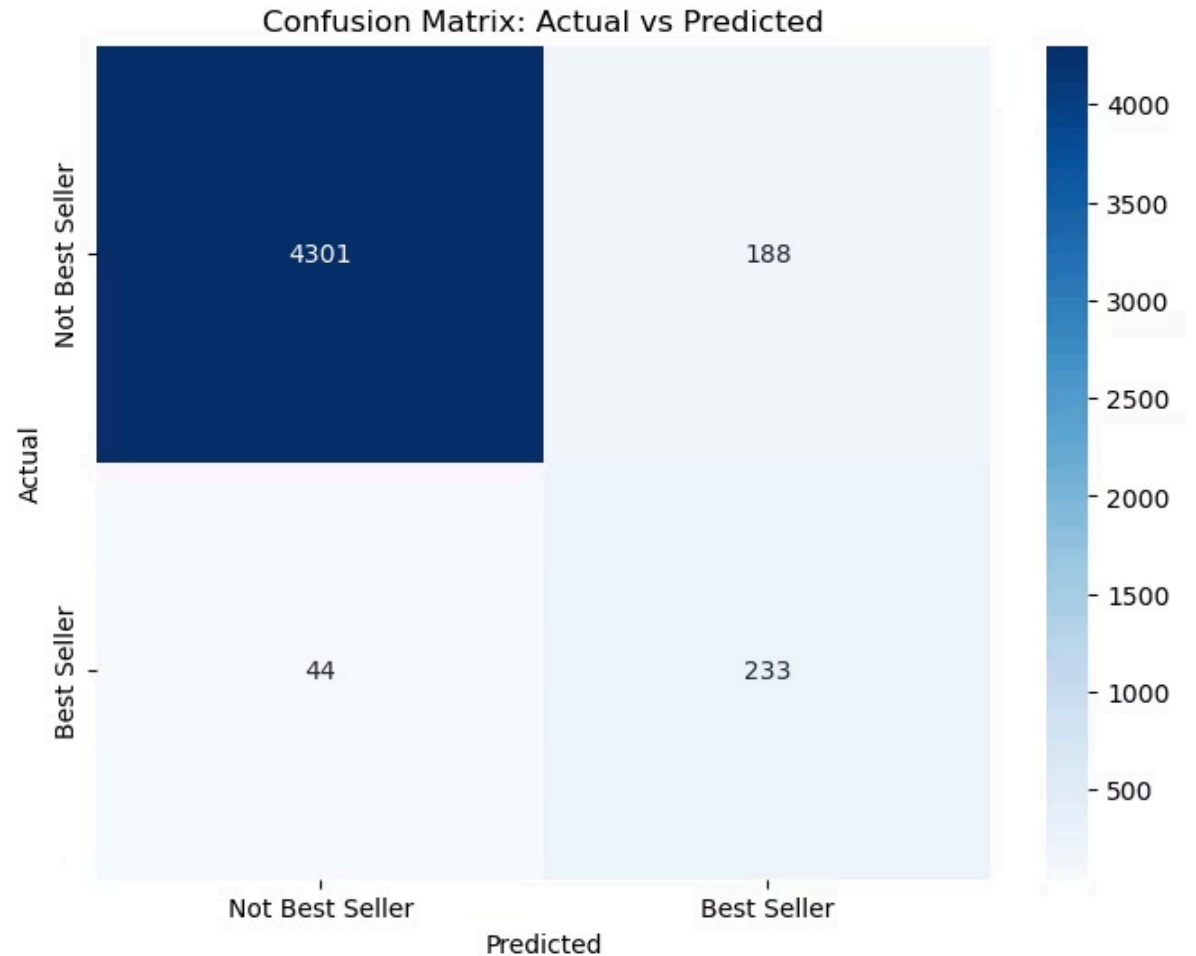
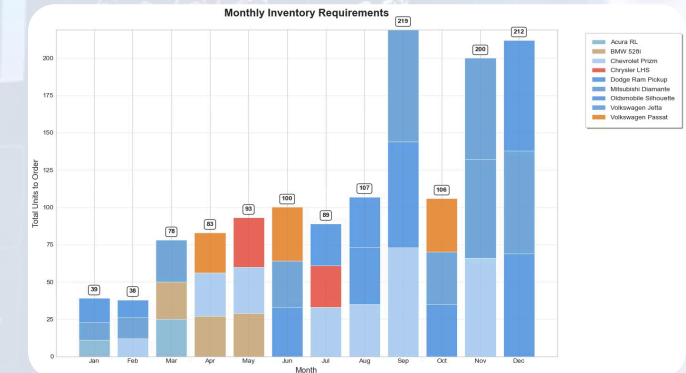# Sales Forecast Highlights – Top Models and Inventory Planning
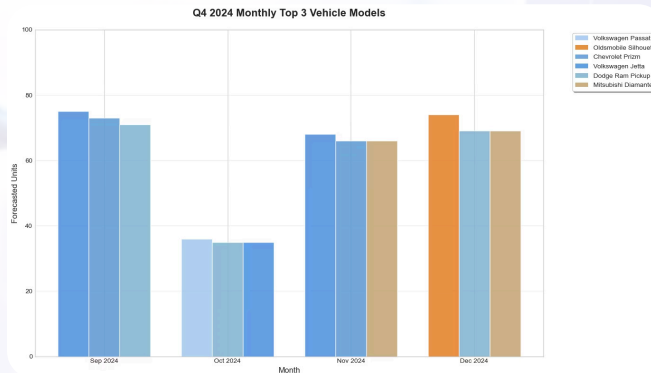
Q4 2024 Forecast

Reveals seasonal trends and the most in-demand models to prepare for year-end sales peaks.

Annual Projections

Identifies the highest-volume vehicles forecasted for 2024, enabling informed stocking decisions.

Monthly Inventory Requirements

Supports monthly order planning by anticipating vehicle demand and minimizing overstock or stockouts.

Made with GAMMA

Business Impact

Improved Inventory

Dealers can now anticipate best sellers monthly, optimizing stock levels.

Higher Sales

Right cars at the right time leads to increased sales volume.

Reduced Risk

Lower chance of overstock and associated carrying costs.

Customer Satisfaction

Improved availability of desired models enhances customer experience.

Next Steps

1 **Integration**

Integrate the model into dealer dashboards or API endpoints

2 **Retraining**

Implement periodic retraining with new sales data

3 **Model Enhancement**

Expand feature set and explore ensemble techniques to improve prediction accuracy.

Made with GAMMA

Thank You

Questions?

All code and documentation are available in our GitHub repository:

Scan to access the GitHub repo:

View GitHub Repository