Vehicle Sales Forecasting System Product Requirements Document

1. Introduction

This document outlines the design and architecture for a project aimed at developing statistical forecasting models to predict monthly vehicle sales at a regional dealership level. The models will incorporate both time-driven patterns and external economic indicators to improve accuracy and provide actionable insights for automotive original equipment manufacturers (OEMs). The primary goal is to provide sales executives and operations planners with accurate forecasts to support production planning, inventory management, and sales strategy.

2. Business Planning

The key business goals of this project are:

- **Improve Production Planning:** Provide accurate forecasts to align vehicle production with anticipated demand.
- **Optimize Inventory:** Reduce inventory holding costs and minimize imbalances by better understanding regional demand.
- **Enhance Sales Strategy:** Identify the causal impact of external factors to inform sales incentives and marketing campaigns.
- **Mitigate Risk:** Provide region-specific risk indicators to address potential market volatility.

3. Solution

a. HIgh Level

Data Ingestion and Preprocessing Layer: This layer is responsible for gathering raw data from various sources (e.g., simulated sales data, public economic datasets) and cleaning it.
Modeling and Analysis Layer: This is the core of the project, where statistical models are built, trained, and evaluated. It will focus on comparing baseline time-series models with more advanced causal models.
Visualization and Reporting Layer: This layer translates the model outputs into a format that is easy for business stakeholders to understand. It will include dashboards and summary reports to communicate key findings and recommendations.

b. Breakdown

- **Data Sources:**
  - **Simulated/Public Regional Sales Data:** Monthly vehicle sales data segmented by region and vehicle type. The dataset will span 3–5 years.

    Source: Government portal

  - **External Economic Data:** This will include time-series data for fuel/oil prices, interest rates, and a policy index.

    Source: Government international prices of crude oil (Indian basket)

  - **Calendar/Holiday Data:** Information on major holidays and festivals that may influence sales.

    Source: Python 'holidays' library

- **Data Cleaning and Transformation:**
  - **Handling Missing Values:** Use interpolation or other statistical methods to fill in gaps in the data.
  - **Outlier Detection:** Identify and handle anomalous sales figures that could skew model performance.
  - **Aggregation:** Ensure all data is aggregated to a monthly time step and aligned by date.
  - **Feature Engineering:** Create new variables from the raw data, such as a "holiday effect" flag or lagged variables for economic indicators.

- **Model Selection:**
  - **Baseline Models: ARIMA** (AutoRegressive Integrated Moving Average) and **Exponential Smoothing** will be used as baselines to capture time-series patterns.
  - **Causal Models: SARIMAX** (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) will be the primary model for incorporating external variables to measure their causal impact.

- **Model Development:**
  - **Train/Test Split:** The dataset will be split into a training set (e.g., the first 80%) and a holdout test set (the last 20%) to evaluate model performance on unseen data.

    - **Hyperparameter Tuning:** Use techniques like grid search or randomized search to find the optimal parameters for each model.
    - **Cross-Validation:** Employ time-series cross-validation to ensure model robustness.

- **Model Evaluation:**
    - **Primary Metric: Mean Absolute Percentage Error (MAPE)** will be the key metric to measure forecasting accuracy against the success criteria of < 10%.
    - **Secondary Metrics: Mean Absolute Error (MAE)** and **Root Mean Squared Error (RMSE)** will also be used to provide a comprehensive view of model performance.
    - **Residual Analysis:** Analyze the residuals (the difference between predicted and actual values) to check for patterns, which could indicate a need for model improvement.

- **Visualization Platform:** Tableau CRM (or a similar tool) or Jupyter-based visuals will be used to create interactive dashboards.
- **Dashboard Features:**
    - **Forecast Visualizations:** Line charts showing historical sales, forecasted sales, and confidence intervals (error bands) for different regions and vehicle types.
    - **Performance Metrics:** A summary of MAPE, MAE, and RMSE for each model and region.
    - **Causal Impact Insights:** Visuals demonstrating the relationship between external variables (e.g., fuel prices) and sales trends.
    - **Region-Specific Strategies:** A dedicated section to highlight model-driven recommendations for specific regions.

c. API Calls

i. API 1: `/api/forecast`

Generates a forecast for a specific region and vehicle type, returning the predicted sales values along with confidence intervals.

ii. API 2: `/api/performance-metric`

Calculates and returns key performance metrics (MAPE, MAE, RMSE) for a specified model and region on the holdout test set.