Forecasting Analysis of Automobile Sales

This project aimed to address one of the most critical challenges in today's automobile industry — **anticipating future demand** amid the ongoing transition from conventional vehicles to electric vehicles (EVs). Using monthly sales data from **2015–2023**, sourced from the Ministry of Road Transport and Highways (**VAHAN Dashboard**), the study forecasted **2024 EV and Non-EV sales** through advanced time-series models.

Business Context

India's automobile sector is experiencing a pivotal shift, driven by **government incentives, environmental awareness, and consumer adoption of EVs**. For manufacturers, accurate forecasting is not just about predicting numbers — it is about aligning **production capacity, supply chain management, and marketing strategies** with real market demand.

Inaccurate forecasts can lead to **overproduction, inventory pile-ups, or missed opportunities** in high-demand segments, making forecasting a strategic priority.

Approach

The project followed a structured analytics workflow:

- **Data Preprocessing**: Cleaning, handling missing values, detecting outliers, and normalizing sales data.
- **Exploratory Analysis**: Identifying long-term trends, seasonal cycles (festivals, subsidies, new launches), and anomalies.
- **Model Development**:
    - *Baseline models*: Moving Averages, Weighted Averages.
    - *Smoothing models*: Exponential Smoothing (SES, DES, Holt-Winters).
    - *Advanced models*: SARIMA for seasonal forecasting and LSTM for capturing non-linear dependencies.

- **Model Evaluation**: Accuracy assessed using **MAPE, RMSE, R^2, and MAD** for a comprehensive comparison.

Key Insights

- **EV Sales**: Strong upward trend with seasonal spikes during **Diwali and subsidy cycles**.
    - **SARIMA** achieved the **lowest forecast error (MAPE ~9.79%)**, making it the most reliable model for EV demand.
    - Holt-Winters also captured seasonality well.
    - LSTM highlighted complex non-linear patterns but required more granular data for higher accuracy.
- **Non-EV Sales**: Showed **steady but slower growth** with mild seasonality.
    - **Double Exponential Smoothing (DES)** proved the most reliable, with the **lowest MAPE ~7.22%**, thanks to its ability to capture stable trends efficiently.

Best Performing Models:

- **EVs → SARIMA** (lowest forecast errors, robust with seasonality).
- **Non-EVs → DES** (simpler, computationally efficient, strong trend fit).

Strategic Implications

- **Manufacturers**: Use **SARIMA-driven EV forecasts** to align production with **peak demand cycles** and reduce risks of under/over supply.
- **Supply Chain & Inventory**: Apply **DES forecasts for Non-EVs** to ensure smoother inventory planning in a predictable market.

- **Policy Makers & Investors**: Forecasts provide **data-backed evidence** of EV adoption momentum, supporting targeted policies and capital allocation.
- **Future Outlook**: With **larger, more granular datasets**, LSTM models could unlock **even more accurate long-term forecasts** as EV adoption accelerates.

Conclusion

By integrating **traditional statistical techniques (SARIMA, DES)** with **modern machine learning (LSTM)**, this project delivered not only **accurate sales forecasts** but also **actionable insights** for decision-makers.

The results clearly highlight:

- **SARIMA → Best for EVs** (MAPE: ~9.79%).
- **DES → Best for Non-EVs** (MAPE: ~7.22%).

These findings help bridge the gap between **data science and business strategy**, ensuring that the Indian automobile industry can **adapt effectively to the dynamic EV revolution**.

Starbucks Customer Segmentation

🛑 Problem

Starbucks serves a **diverse global customer base** with varying preferences, lifestyles, and behaviors. However, a **one-size-fits-all marketing strategy** often fails to maximize customer engagement, loyalty, and sales.

The challenge: **How can Starbucks segment its customers effectively to deliver personalized offers, improve loyalty, and drive long-term growth?**

🔍 Approach / Methodology

1. **Data Collection & Preparation**
    - Consolidated and pre-processed **three datasets (~300,000 entries)**.
    - Ensured data cleanliness, merged sources, handled missing values, and standardized formats.
2. **Exploratory Data Analysis (EDA)**
    - Identified customer spending patterns, demographics, and behavioral attributes.
    - Engineered features to capture **recency, frequency, and monetary behavior**.
3. **Dimensionality Reduction**
    - Applied **Principal Component Analysis (PCA)** to reduce data complexity while retaining key variance.
    - Used **t-SNE optimization** to visualize high-dimensional clusters.
4. **Clustering Algorithm**

- Implemented **K-Means clustering** on ~17,000 customer records (after feature selection).
- Determined **optimal number of clusters** using **Silhouette Score & Sum of Squared Errors (SSE)**.

💡 Solution / Results

- **K-Means Segmentation Outcome**: Customers segmented into **4 key clusters** (example interpretation):
  - **Busy Professionals** → High spenders, frequent mobile app users, prioritize convenience.
  - **College Students** → Price-sensitive, socially conscious, drawn to sustainability offers.
  - **Loyal Regulars** → Frequent store visitors, moderate spenders, respond well to loyalty rewards.
  - **Occasional Buyers** → Low-frequency customers, responsive to discounts and seasonal promotions.
- **Validation Metrics**:
  - **Silhouette Score** → Confirmed optimal separation between clusters.
  - **SSE (Elbow Method)** → Helped determine cluster count stability.

📑 Recommendations

- **Cluster 1 (Busy Professionals)** → Promote **mobile ordering & premium subscription offers**.
- **Cluster 2 (College Students)** → Discounts on reusable cups, emphasize **sustainability campaigns**.

- **Cluster 3 (Loyal Regulars)** → Strengthen **loyalty rewards & personalized offers**.
- **Cluster 4 (Occasional Buyers)** → Send **seasonal promotions & introductory discounts** to increase frequency.

🌍 Impact / Use

- **Personalized Marketing**: More relevant offers → higher engagement and conversion rates.
- **Customer Loyalty**: Tailored rewards strengthen emotional connection with the brand.
- **Sales Growth**: Targeted campaigns reduce marketing waste and maximize ROI.
- **Strategic Insight**: Helps Starbucks understand **demographic & behavioral drivers** for each segment.

Credit Card Fraud Detection

🛑 Problem

Credit card fraud is a **critical challenge for financial institutions** as the number of digital transactions continues to rise. The primary difficulty lies in **detecting fraudulent transactions in real time** while minimizing false positives. Even small detection errors can result in **financial loss, reputational damage, and reduced customer trust**.

🔍 Approach / Methodology

1. **Data Collection**
   - Used publicly available imbalanced datasets (fraudulent vs. non-fraudulent transactions).
   - Ensured data was representative of real-world imbalance (fraud cases <1%).
2. **Data Preprocessing**
   - Handled **imbalanced classes** using techniques such as **SMOTE (Synthetic Minority Oversampling)**.
   - Performed **feature engineering** to identify patterns linked to fraudulent behavior.
   - Standardized/normalized transaction attributes for model readiness.
3. **Exploratory Data Analysis (EDA)**
   - Visualized fraud vs. non-fraud distributions.
   - Analyzed transaction patterns (amount, time, frequency).
   - Identified anomalies and correlations critical for fraud detection.
4. **Model Development**

- Trained and optimized multiple algorithms:
    - Logistic Regression
    - Decision Tree
    - Support Vector Machine (SVM)
    - K-Nearest Neighbors (KNN)
    - XGBoost
    - Neural Networks
- Applied **Grid Search CV** for hyperparameter tuning.

5. **Model Evaluation**
    - Evaluated using: **Recall (to minimize false negatives), Precision, F1-score, ROC-AUC, Confusion Matrix**.
    - Chose recall as the **primary metric**, since failing to detect fraud (false negative) is more costly than false positives.

💡 Solution / Results

- Achieved **93% Recall** → High ability to correctly identify fraudulent transactions.
- **K-Nearest Neighbors (KNN)** emerged as the **top-performing model** with:
    - **Accuracy: ~95%**
    - Strong performance across F1-score & ROC-AUC.
- XGBoost and Neural Networks also performed well, but KNN provided the **best balance of interpretability, accuracy, and recall**.

📑 Recommendations

- Deploy **KNN as the primary fraud detection model** with continuous monitoring.
- Use **XGBoost as a secondary ensemble model** for large-scale, high-volume transaction systems.
- Implement a **hybrid system**:
    - High-recall models (KNN/XGBoost) for initial screening.
    - Manual review or rule-based filters for flagged high-value transactions.

🌍 Impact / Use

- **Financial Institutions**: Prevent substantial losses by detecting fraud in real-time.
- **Customer Trust**: Enhanced protection builds loyalty and reduces customer churn.
- **Operational Efficiency**: Reduces false positives → fewer unnecessary investigations.
- **Scalability**: Models can adapt to evolving fraud patterns with retraining.

✅ Conclusion

This project demonstrates the effectiveness of **machine learning for fraud detection** on highly imbalanced datasets.

- **KNN** achieved the best performance with **95% accuracy** and **93% recall**, making it the most suitable model for deployment.
- Fraud detection accuracy directly translates into **reduced financial losses, improved customer trust, and better compliance for financial institutions**.
- The study highlights the importance of **precision-recall tradeoff** in fraud analytics, emphasizing recall as the key metric in high-risk financial applications.

Interactive Business Dashboards

This project showcases a collection of interactive dashboards built to **analyze, monitor, and present business insights** across HR, Sales, and E-commerce domains. The dashboards were created using **Power BI / Tableau / Python visualization libraries** (depending on the use case), focusing on clarity, storytelling, and actionable decision support.

1️⃣ HR Analytics Dashboard

Problem

Employee attrition is a critical issue in HR management, impacting workforce stability and organizational costs.

Approach / Methodology

- Processed HR datasets (~1,470 employees).
- Performed **EDA** to uncover attrition drivers (age, role, job satisfaction, education field).
- Designed **interactive visuals**: attrition by demographics, job roles, tenure, and salary distribution.

Solution / Result

- Dashboard highlighted **237 attritions (16%)** with highest churn among **sales executives, lab technicians, and younger employees (26–35 years)**.
- Attrition drivers such as **low job satisfaction** and **shorter tenure** were visually mapped.

Impact / Use

- HR teams can identify **high-risk employee groups** and design **targeted retention strategies**.
- Supports proactive workforce planning and cost reduction in rehiring.

2️⃣ E-commerce Sales Dashboard

Problem

Businesses need real-time visibility into sales performance, profit contribution, and customer purchasing patterns to optimize strategy.

Approach / Methodology

- Processed e-commerce transactional data (5,615 orders).
- Tracked KPIs: **Profit, Sales Amount, Quantity, Payment Mode**.
- Built **drill-down dashboards** to analyze trends by category, sub-category, geography, and customer segments.

Solution / Result

- **Total Profit:** ₹37K | **Total Sales Amount:** ₹438K.
- **Top-performing categories:** Clothing (62.6%) & Electronics (20.5%).
- **Profitability drivers:** Printers & Bookcases, while some categories (Tables, Accessories) showed losses.
- **Payment preference:** COD (43.7%) dominated, followed by UPI (20.6%).

Impact / Use

- Helps business teams **understand product profitability** and **regional performance**.
- Supports **inventory optimization** and **marketing campaigns** for specific customer groups.

3️⃣ Additional Dashboards

- Financial dashboards

- Customer behavior analytics
- KPI monitoring panels

📌 Recommendations

- Use HR insights to **reduce attrition through role-specific engagement**.
- Optimize e-commerce strategy by **focusing on profitable sub-categories** and **encouraging digital payments**.
- Continuously update dashboards with **real-time data pipelines** for maximum business value.

🛠️ Tools & Tech Stack

- **Power BI, Tableau** → Dashboard creation, interactive visuals
- **Python (Pandas, Matplotlib, Seaborn, Plotly)** → Data cleaning, EDA, visualizations
- **Excel** → Data preprocessing and validation

🚀 Business Impact

These dashboards transformed raw data into **actionable insights**, enabling decision-makers to:

- Enhance **employee retention strategies**.
- Increase **sales profitability and customer targeting**.
- Streamline **reporting and performance monitoring**.

👉 This project reflects my expertise in **data storytelling, visualization, and business analytics** — bridging the gap between raw data and strategic decision-making.