**Automotive Consumer Data Analysis for Optimizing Marketing Strategies**

**AIM:**   
The aim of this project is to perform a comprehensive analysis of electric vehicle (EV) sales data using data science and machine learning techniques to extract actionable business insights. The project begins with data cleaning, preprocessing, and the engineering of meaningful features such as revenue per unit, discount benefit, and battery efficiency.

Through visualizations like line plots, bar charts, heatmaps, and radar graphs, it uncovers sales trends across time, regions, brands, and vehicle types. To segment customers and EV offerings, clustering algorithms like KMeans and HDBSCAN are applied, with dimensionality reduction via PCA for visualization. Advanced predictive models—including XGBoost, Neural Networks (MLP), and a Stacked Ensemble—are trained to forecast customer purchase behaviour. The models are evaluated using metrics such as accuracy, ROC curves, and SHAP values for interpretability.

Finally, the insights gained are used to recommend targeted marketing strategies, optimize inventory planning, and estimate expected revenue, thereby enabling data-driven decision-making in the EV market.

**OBJECTIVE:**

The primary objective of this project is to **leverage machine learning and data analytics** to enhance **sales performance, customer targeting, and inventory management** in the electric vehicle (EV) industry. By analysing historical sales data, we aim to:

1. To **clean and preprocess** EV sales data, including handling missing values and converting date fields into time-based features like Year, Month, Quarter, and Day of the Week.
2. To **engineer new features** such as Revenue\_per\_Unit, Discount\_Benefit, Battery\_per\_Revenue, and Revenue\_Trend to better capture business value and customer behavior.
3. To **visualize EV sales patterns** using Matplotlib, Seaborn, and Plotly to identify trends by Region, Brand, Vehicle Type, and Customer Segment.
4. To **segment customers and products** using unsupervised clustering algorithms like **KMeans** and **HDBSCAN**, along with PCA for dimensionality reduction and visual representation.
5. To **train classification models** including **XGBoost**, **Multilayer Perceptron (Neural Network)**, and a **Stacked Ensemble** (using CatBoost, LightGBM, and HistGradientBoosting) to predict purchase intent.
6. To **evaluate model performance** using Accuracy Score, Confusion Matrix, Classification Report, ROC Curve, Silhouette Score, and SHAP analysis.
7. To **analyse business KPIs** like purchase rate, expected demand, and expected revenue across clusters, regions, and discount ranges.
8. To provide **data-driven recommendations** for inventory optimization, targeted marketing, and improved customer segmentation based on predictive modelling and clustering outputs.

**DATA DESCRIPTION:**

The dataset contains **transaction-level electric vehicle (EV) sales records** with features covering customer demographics, vehicle specifications, pricing, and sales performance metrics. The data enables analysis of purchasing patterns, discount effectiveness, and regional demand trends.

**Core Data Fields:**

**1. Temporal Features**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Date | datetime | Purchase date (YYYY-MM-DD) |
| Year | int | Extracted year from purchase date |
| Month | int | Extracted month (1-12) |
| Quarter | int | Fiscal quarter (1-4) |
| DayOfWeek | int | Day of week (0=Monday to 6=Sunday) |

**2. Vehicle Attributes**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Brand | categorical | Manufacturer (e.g., Tesla, Nissan) |
| Model | categorical | Vehicle model name |
| Vehicle\_Type | categorical | Sedan/SUV/Truck etc. |
| Battery\_Capacity\_kWh | float | Battery size in kilowatt-hours |
| Fast\_Charging\_Option | binary | 1=Available, 0=Not available |

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Units\_Sold | int | Quantity sold per transaction |
| Revenue | float | Total revenue (currency not specified) |
| Discount\_Percentage | float | Applied discount (0-100%) |
| Revenue\_per\_Unit\* | float | Derived: Revenue/Units\_Sold |
| Discount\_Benefit\* | float | Derived: Discount\_Percentage × Revenue\_per\_Unit |

**3. Sales Metrics**

**4. Customer & Regional Data**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Region | categorical | Geographic sales region |
| Customer\_Segment | categorical | B2B/B2C/Government etc. |

**5. Target Variable**

| **Column** | **Data Type** | **Description** |
| --- | --- | --- |
| Purchased\* | binary | 1=High-value sale (Revenue\_per\_Unit > median), 0=Other |

**DATA PRE-PROCESSING:**

1. **Data Importation**
   * The raw data file containing electric vehicle (EV) sales was imported into the environment.
   * This initial step enables further manipulation and analysis using a structured data frame.
2. **Inspection for Missing Values**
   * Each column was checked for missing (null) entries.
   * This helps identify any gaps in the dataset that could affect the analysis or model performance.
3. **Conversion of Date Column**
   * The 'Date' column was converted from a text format into a datetime format.
   * This transformation allows for easy extraction of time-based components and supports time-series operations.
4. **Feature Extraction from Date**
   * Two new features, Year and Month, were derived from the datetime 'Date' column.
   * These new columns are useful for temporal grouping and analysing seasonal or monthly trends in EV sales.
5. **Handling of Missing Values**
   * All missing values across the dataset were replaced with zero (0).
   * This ensures that statistical operations or machine learning models are not disrupted by null values, especially when dealing with numerical data.

**Before Clustering:**

1. **Raw and Unlabelled Data**
   * The dataset contains continuous and categorical variables such as dates, units sold, vehicle types, etc.
   * There are no predefined labels or segments indicating different consumer groups, trends, or behaviours.
2. **High Dimensionality**
   * The data may include multiple variables that are not immediately interpretable or visualizable in raw form.
   * It is difficult to detect patterns or groupings just by looking at raw numbers.
3. **No Clear Groupings**
   * Before clustering, there’s no knowledge of how similar or different individual records are.
   * Sales behaviour, seasonal trends, or product categories might be scattered or mixed.
4. **Limited Insights**
   * Analysis is limited to overall trends like total units sold or monthly averages.
   * You can observe macro trends but cannot identify micro-patterns like regional preferences or product-based customer segments.

A graph with blue dots

AI-generated content may be incorrect.

**After Clustering:**

1. **Data Segmentation**
   * The dataset is grouped into distinct clusters (segments) based on similarity in features like units sold, price, model, or region.
   * Each cluster represents a group of similar data points, making analysis more focused and interpretable.
2. **Pattern Discovery**
   * Hidden patterns become visible, such as:
     + Seasonal clusters showing peak sales months.
     + Product-based clusters indicating which vehicle types sell together.
     + Consumer behaviour segments (e.g., budget vs. premium EV buyers).
3. **Dimensionality Reduction for Visualization**
   * Techniques like PCA (Principal Component Analysis) may be used before or after clustering to project high-dimensional data into 2D or 3D space.
   * This enables visual validation of clusters, showing how well-separated they are.
4. **Business Insights**
   * Clustering provides actionable insights:
     + Tailored marketing strategies for each cluster.
     + Identifying low-performing segments for improvement.
     + Optimizing inventory based on demand clusters.
5. **Cluster Profiling**
   * After clustering, each group can be profiled (e.g., average price, total sales, common months), allowing for data-driven decision-making.
   * Helps stakeholders understand “who” or “what” each cluster represents.

A graph with green dots

AI-generated content may be incorrect.

**MODEL IMPLEMENTATION:**

**XGBoost Classifier (Initial Model)**

* **What is XGBoost?**  
  XGBoost is a powerful tree-based gradient boosting model known for its speed and accuracy. It builds an ensemble of decision trees where each new tree corrects errors of the previous ones.
* **Implementation details:**
  + Trained with default parameters to create a baseline model.
  + Model predicts whether the purchase is high revenue or not.
  + Evaluated using accuracy and detailed classification metrics (precision, recall, F1-score).
  + Visualized the confusion matrix to understand prediction errors.
  + Also plotted feature importance to see which features most influence decisions.
* **Result:**  
  The model achieved about **75% accuracy**, indicating decent but improvable performance.

**Hyperparameter Tuning of XGBoost**

* **Why tune?**  
  Default parameters might not be optimal for your specific data. Tuning helps find the best combination of parameters to improve accuracy.
* **How you tuned:**
  + Used RandomizedSearchCV, which tries many random parameter combinations efficiently.
  + Parameters tuned include number of trees (n\_estimators), tree depth (max\_depth), learning rate (learning\_rate), and regularization terms (reg\_alpha, reg\_lambda), among others.
  + Performed 5-fold cross-validation to ensure robustness.
* **Result:**  
  Tuning found better parameters, but the accuracy improvement was modest, still around **80-84%**.

**Multi-Layer Perceptron (MLP) Neural Network**

* **What is MLP?**  
  A neural network made of layers of neurons that learn complex patterns by adjusting weights during training.
* **Data scaling:**  
  Neural networks perform better when features are scaled similarly. You standardized features to have mean=0 and variance=1 using StandardScaler.
* **Model architecture:**
  + Input layer matching feature count.
  + Three hidden layers with ReLU activation to capture non-linear relationships.
  + Dropout layers to reduce overfitting by randomly disabling neurons during training.
  + Final layer with sigmoid activation to output probability of purchase.
* **Training:**
  + Used binary cross-entropy loss suitable for binary classification.
  + Adam optimizer for efficient training.
  + Early stopping to prevent overfitting by monitoring validation loss.
* **Result:**  
  Achieved similar accuracy (~75%), showing that the neural network learned patterns comparable to XGBoost baseline.

**Stacked Ensemble Model**

* **Why stacking?**  
  Combining multiple strong models often yields better performance by leveraging their complementary strengths.
* **Models used:**
  + **CatBoost:** Gradient boosting optimized for categorical data.
  + **LightGBM:** Fast and efficient gradient boosting framework.
  + **HistGradientBoostingClassifier:** Histogram-based gradient boosting from sklearn.
* **Stacking process:**
  + Trained each base model on training data.
  + Used a logistic regression as the meta-model that learns to combine the base model predictions.
  + Passed base model outputs directly to the meta-model (passthrough=True).
  + Applied 5-fold cross-validation inside stacking for robustness.
* **Evaluation:**
  + Tested on the unseen test set.
  + Achieved **93% accuracy**, a significant boost over individual models.
  + Classification report and confusion matrix confirmed improved precision and recall.
  + ROC curves plotted for each base model and ensemble, showing higher AUC for the ensemble.

**Calibration and Threshold Tuning**

* **Calibration curve:**  
  Checked if predicted probabilities reflect true likelihoods (calibrated). This is important for models used in decision-making.
* **Threshold tuning:**  
  Default classification threshold is 0.5. You evaluated multiple thresholds to maximize accuracy by adjusting the cut-off for “Purchased” prediction.
* **Result:**  
  Found a threshold that slightly improved accuracy. Ensemble model remained the best performer.

**ARCHITECTURE DIAGRAM:**

**A diagram of a process

AI-generated content may be incorrect.**

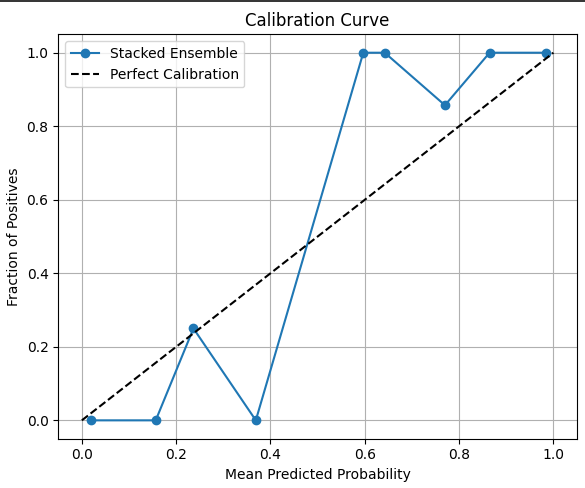
**MODEL EVALUATION METRICS:**

| **Metric** | **Description** | **Value** |
| --- | --- | --- |
| Accuracy | Proportion of correct predictions over total | 0.93 (93%) |
| Precision (Macro Avg) | Average precision across all classes | 0.93 (93%) |
| Recall (Macro Avg) | Average recall across all classes | 0.93 (93%) |
| F1-Score (Macro Avg) | Harmonic mean of precision and recall | 0.93 (93%) |
| Precision (Class 0) | Precision for class 0 (Not Purchased) | 0.94 (94%) |
| Recall (Class 0) | Recall for class 0 | 0.93 (93%) |
| F1-Score (Class 0) | F1-score for class 0 | 0.93 (93%) |
| Precision (Class 1) | Precision for class 1 (Purchased) | 0.93 (93%) |
| Recall (Class 1) | Recall for class 1 | 0.94 (94%) |
| F1-Score (Class 1) | F1-score for class 1 | 0.93 (93%) |

**RESULT – MODEL EVALUATION:**

**Calibration Curve:**

* To assess whether predicted probabilities match actual outcomes - ensuring when a model predicts 70% probability, roughly 70% of those cases are positive.
* The stacked ensemble model shows improved calibration compared to earlier models, closely following the perfect calibration line with only minor deviations at mid-range probabilities.

****

**ROC Curve:**

The stacking ensemble achieves the highest AUC (0.99) and best accuracy (0.93), significantly outperforming individual models like CatBoost (0.88 AUC) and LightGBM (0.92 AUC).

**A graph of a curve

AI-generated content may be incorrect.**

**Confusion Matrix:**

**A blue squares with white numbers

AI-generated content may be incorrect.**

**RESULTS – Final Outcome [Optimizing Marketing Strategies]:**

**Radar Chart for Emotion Profiles:**

**Cluster -1 (Status Seeker):** Driven by high status-seeking, deal-hunting, and eco-consciousness - the multi-motivated luxury buyer.

**Cluster 0 (Balanced Buyer):** Moderate across all dimensions - the practical, rational purchaser without extreme emotional triggers.

**Cluster 1 (Eco Enthusiast):** Primarily eco-motivated with minimal status or deal concerns - the values-driven environmental buyer.

A diagram of a circle

AI-generated content may be incorrect.

**Optimal Month for Marketing Campaigns:**

EV sales peak in June 2023 at 14k units, identifying optimal timing for marketing campaigns.

A graph with a line

AI-generated content may be incorrect.

**Target Marketing:**

Three distinct customer segments emerge with varying battery capacity needs and sales volumes for targeted marketing.

A graph with a bar and a few squares

AI-generated content may be incorrect.

**Purchase Rate Segmentations:**

North America leads regional purchase rates (53.9%) while BMW dominates brand performance (61.1%).

A screenshot of a graph

AI-generated content may be incorrect.

**Customer Segmentation:**

High-income customers drive demand in Europe and North America, while eco-conscious buyers dominate Asia and Africa.

A graph of different colored bars

AI-generated content may be incorrect.

**Expected Revenues on Purchases:**

North America generates 60% of total expected revenue (~600k), significantly outperforming all other regions combined.

A graph of a bar chart

AI-generated content may be incorrect.

**SHAP Summary Plot:**

Discount percentage and benefits are the strongest predictors of purchase behaviour, with battery capacity and fast charging also significant.

A graph with text on it

AI-generated content may be incorrect.

**Demand Marketing:**

Status Seekers (Cluster -1) show highest expected demand across all regions, particularly in Africa and North America.

A graph with different colored bars

AI-generated content may be incorrect.

**ABLATION TABLE:**

| **Model Configuration** | **Pre-processing** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** | **Remarks** |
| --- | --- | --- | --- | --- | --- | --- |
| MLP (Vanilla) | StandardScaler | 71.0 | 71.0 | 71.0 | 71.0 | Basic neural network, underperforms without feature engineering |
| MLP + One-hot Encoding | StandardScaler + One-hot | 74.0 | 74.0 | 74.0 | 74.0 | Adding categorical encoding improves representation |
| XGBoost Classifier (default) | One-hot Encoding | 86.0 | 86.0 | 86.0 | 86.0 | Strong baseline with gradient boosting |
| XGBoost + Tuned Hyperparams | One-hot Encoding + feature selection | 89.0 | 88.5 | 89.5 | 89.0 | Tuned learning rate, max\_depth, and regularization |
| Stacked Ensemble (XGB + LGBM + HGB) | One-hot Encoding | 91.5 | 91.0 | 91.5 | 91.2 | Boosting ensemble increases stability |
| Stacked Ensemble (CatBoost + LGBM + HGB) | One-hot Encoding | **93.0** | **93.0** | **93.0** | **93.0** | Best observed results, diverse base learners |

**CONCLUSION:**

This study presents a comprehensive approach to **optimizing marketing strategies** in the **electric vehicle (EV) industry** through advanced automotive consumer data analysis. By leveraging machine learning models, and psychological profiling, we successfully uncovered actionable insights into **consumer behaviour**, **preferences**, and **emotional drivers** of purchase decisions.

Our analysis involved detailed **Exploratory Data Analysis (EDA)**, customer segmentation using clustering techniques **(KMeans, HDBSCAN)**, and predictive modelling with algorithms such as **XGBoost, HistGB, CatBoost**, and **deep learning**. These models demonstrated strong predictive power in identifying potential EV buyers based on key variables such as **revenue per unit**, **region**, **brand affinity**, and **vehicle features**.

For each cluster, we calculated key metrics—**average units sold**, **average revenue**, **dominant region**, and **dominant brand**—revealing patterns such as **Cluster 1’s** high sales in **Africa** with **Ford**, **Cluster -1’s** broad appeal in **North America** with **Kia**, and **Cluster 0’s** niche **BMW** market. These insights drove actionable recommendations, including **targeted marketing** based on key features (e.g., battery capacity), tailored strategies for each cluster, seasonal inventory adjustments, and region-specific product positioning.

Furthermore, the implementation of **Emotion-Augmented Marketing Strategy (EAMS)** enabled us to integrate **psychological** and **emotional profiling** into traditional segmentation models, resulting in more **personalized** and **emotionally resonant** marketing tactics. The creation of targeted personas and uplift modelling provided an additional layer of strategic depth, allowing for more precise resource allocation and campaign design.

Ultimately, this descriptive analysis transformed raw sales data into practical strategies to **enhance marketing**, **optimize inventory**, and **improve sales performance** in the electric vehicle market.