Analyzing & Visualizing the EV Vehicles Population

Team Information

Team Name: SPIRIT
Section Number: 44517-04
Team Members:

- Venkateswara Rao Gude
- Vinay Marella
- Harsha Sai Teja Nukala
- Khaja Nayab Rasool Shaik

1 Project Idea

The project aims to analyze and visualize the Electric Vehicle (EV) population using data preprocessing, feature engineering, and machine learning. The project will utilize PySpark for distributed data processing and Tableau for visualizing insights.

1.1 Key Highlights

1. **Data Acquisition:** Download the EV population dataset in CSV format from Kaggle.
2. **Data Preprocessing:** Use Python libraries like Pandas to load the CSV file and perform essential data cleaning tasks, such as handling missing values, resolving data inconsistencies, and reformatting as necessary.
3. **Visualization with Tableau:** Use Tableau to create a range of visualizations that provide insights into the EV population data. Examples include:
   - Line charts to track the trend of EV adoption over time.
   - Bar charts to compare EV populations by region or manufacturer.
   - Histograms to visualize the distribution of EVs across different model years.

1.2 Tools and Technologies

- **PySpark:** For distributed data processing and manipulation.
- **Tableau:** A powerful data visualization tool for creating interactive dashboards and visualizations.

These tools and technologies will help efficiently access, visualize, and analyze the EV population data.

1.3 Dataset

The dataset used for this project is the Electric Vehicle Population Data available at:
`https://www.kaggle.com/datasets/ratikkakkar/electric-vehicle-population-data`.

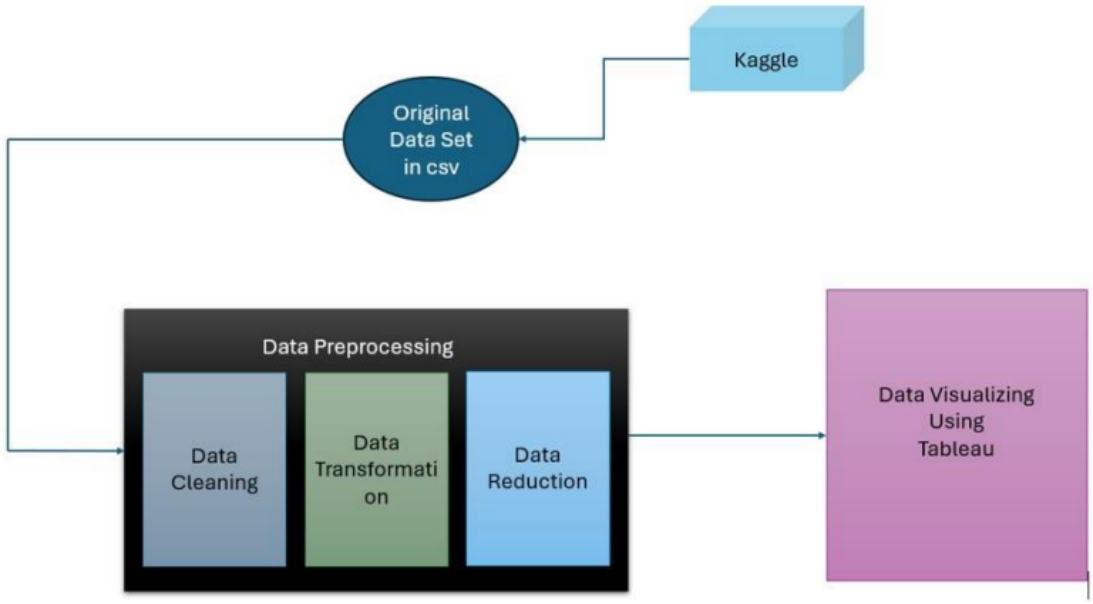

Figure 1: Data Flow Diagram

2 Data Flow Diagram

3 Block Diagram Explanation

1. The EV population data is obtained from external sources such as Kaggle in CSV format.
2. In the data preprocessing phase, relevant data is loaded into a DataFrame and unnecessary information is eliminated using SQL queries.
3. The preprocessing steps are customized using PySpark to align with project objectives.
4. After preprocessing, the data is visualized using various plots and graphs with comparisons displayed in Tableau.

4 Goals to Investigate

The following goals will be the focus of this project:

1. **Data Preprocessing and Feature Engineering:**
   - Perform data preprocessing tasks like handling missing values and scaling features using Apache Spark.
   - Engineer relevant features, such as subscription length, engagement frequency, and payment patterns.
2. **Develop a Scalable Machine Learning Model:**
   - Train a distributed machine learning model using Spark's MLlib to predict customer churn.
   - Explore algorithms such as logistic regression and decision trees.
3. **Model Evaluation:**
   - Evaluate the model using performance metrics like accuracy, precision, recall, and F1-score.

- Assess the model's ability to generalize across different customer segments.

4. **Identify Key Predictors of Churn:**
   - Investigate features (e.g., customer tenure, frequency of service usage, payment behavior) influencing churn.
   - Conduct feature importance analysis.
5. **Visualization and Risk Segmentation:**
   - Visualize churn predictions and customer risk segmentation using Tableau.
   - Create interactive dashboards to display churn risk levels for different customer groups.
6. **Actionable Insights for Retention Strategies:**
   - Provide recommendations for targeted retention campaigns based on churn predictions.
7. **Optimize Data Storage and Querying:**
   - Store processed data and churn prediction results in Hive for efficient querying.
8. **Scalability and Performance Optimization:**
   - Investigate how to scale the system to handle increasing data volumes while maintaining performance.

By addressing these goals, the team aims to build a robust, scalable system that predicts customer churn accurately and provides businesses with tools to improve retention.