**Data - Driven Analysis of EV Charging Infrastructure for a Smart City - Medium/Heavy Duty Vehicles**

A Project Report

Presented to

The Faculty of the Department of Applied Data Science San Jose State University In Partial Fulfillment

Of the Requirements for the Degree

Master of Science in Data Analytics

By

*Mahe Jabeen Abdul, Pranavi Sandrugu, Rohan Naga Venkata Mayukh Ungarala, Lohitha Vanteru*

*December 15, 2023*

*Copyright © 2023*

*Mahe Jabeen Abdul, Pranavi Sandrugu, Rohan Naga Venkata Mayukh Ungarala, Lohitha Vanteru*

*ALL RIGHTS RESERVED*

**APPROVED FOR DEPARTMENT OF APPLIED DATA SCIENCE**

Dr. Jerry Gao, Project Advisor

Dr. Lee C. Chang, Department Chair

**ABSTRACT**

The advent of electric vehicles (EVs) has ushered in a transformative era in urban transportation.In response to the rising prominence of electric vehicles (EVs) in urban transportation, this project conducts a meticulous data-driven analysis, concentrating on the unique challenges and opportunities presented by the charging infrastructure tailored for medium/heavy-duty vehicles within the dynamic context of a smart city. The overarching objectives of the study encompass forecasting the expanding demand for EVs, mitigating concerns related to driving range limitations, optimizing energy resource allocation, and strategically situating charging stations. Leveraging advanced forecasting models, such as the Prophet model for market growth and a Stacking Ensemble Regressor model with weighted fusion for range prediction, the study anticipates and addresses critical issues associated with EV adoption. Additionally, the Temporal Fusion Transformer model is applied to optimize energy resource allocation at EV charging stations, ensuring sustainability and cost-effectiveness. The placement of charging infrastructure is strategically addressed through the integration of a PuLP linear programming optimization model with K-means clustering, enabling the prediction of optimal locations for new charging stations that cater to diverse vehicle types and align with the evolving needs of urban mobility. This comprehensive project, employing cutting-edge machine learning methodologies, encapsulates a forward-thinking analysis of the evolving landscape of EV charging infrastructure. The findings aim to contribute actionable insights for the enhancement and strategic expansion of EV charging infrastructure, thereby propelling sustainable urban mobility into the future.

***Keywords***: *Machine Learning, Stacking Ensemble Regressor, Temporal Fusion Transformer*

**ACKNOWLEDGMENTS**

We would like to express our gratitude and special thanks to our project guide, Dr. Jerry Gao (Zeyu), Department of Applied Data Science, San José State University, and supervisor Dr. Lee C.Chang who is the Director of the Department of Applied Data Science at San José State University for providing us with a wonderful opportunity to study, research, and work on the project related to Natural Language Processing in the field of Artificial Intelligence. We would like to sincerely appreciate Dr. Jerry Gao (Zeyu), efforts in guiding the entire team in the right direction to carry out this project successfully. His motivation and guidance helped us to push ourselves in the most critical stages of our project. We would also like to appreciate all his efforts and the time he has given to our team to make this project a successful one by providing all the problem-solving guidelines.

Towards the end of the project, we were confident that we are technically skilled enough to take up any real-time project in a company and work on it. All the credits go to our project guide, Dr. Jerry Gao (Zeyu), Department of Data Science, San José State University. He taught us how to be open-minded and accept challenges and work on them step by step to achieve the end results successfully.

Furthermore, we would like to appreciate each other's efforts in the team for working in different timings coordinating well within the team, and performing the tasks that we assigned each other. The Data Collection part was the crucial stage for this project and our teammates together catered a lot of time and effort in forming the dataset without which getting good output results for our project would not be possible.

We would also like to thank our supervisor Dr. Lee C. Chang for designing this program and allowing us to understand our technical capabilities. Without this program, we would not have been able to upscale our knowledge and problem-solving skills. We are also grateful to our spouses for all their understanding and support throughout our project tenure. Finally, we would like to thank our family and friends for all their support and encouragement throughout the completion of this project work.

**TABLE OF CONTENTS**

**Chapter 1 Introduction**

1.1 Project Background and Execute Summary

1.2 Project Requirements

1.3 Project Deliverables

1.4 Technology and Solution Survey

1.5 Literature Survey of Existing Research

**Chapter 2 Data and Project Management Plan** 2.1 Data Management Plan

2.2 Project Development Methodology

2.3 Project Organization Plan

2.4 Project Resource Requirements and Plan

2.5 Project Schedule

**Chapter 3 Data Engineering**

3.1 Data Process

3.2 Data Collection

3.3 Data Pre-processing

3.4 Data Transformation

3.5 Data Preparation

3.6 Data Statistics and Data Analytics Results

**Chapter 4 Model Development**

4.1 Model Proposals

4.2 Model Supports

4.3 Model Comparison and Justification

4.4 Model Evaluation Methods

4.5 Model Validation and Evaluation Results

**Chapter 5 Data Analytics System**

5.1 System Requirements Analysis

5.2 System Design

5.3 System Development and Implementation

5.4 System Supporting Environment

**Chapter 6 System Evaluation and Visualization**

6.1 Analysis of Model Execution and Evaluation Results

6.2 Achievements and Constraints

6.3 Quality Evaluation of Model Functions and Performance

6.4 Evaluation of Models vs. Requirements

6.5 Project Information Visualization

**Chapter 7 Conclusion**

7.1 Summary

7.2 Benefits and Shortcoming

7.3 Potential System and Model Applications

7.4 Experience and Lessons Learned

7.5 Recommendations for Future Work

7.6 Contributions and Impacts on Society

**References**

**Appendices**

**Appendix A – System Testing**

**Appendix B – - Project Data Source and Management Store**

**Appendix C – Project Program Source Library, Presentation, and Demonstration**

1

**1. Introduction**

**1.1 Project Background and Executive Summary**

The transportation industry is undergoing a substantial shift towards environmentally conscious and sustainable methods, with a particular focus on electric vehicle technology. In this context, medium- and heavy-duty electric vehicles (EVs) are beginning to show promise as viable alternatives to their combustion engine counterparts. However, the successful integration of electric vehicles into the transportation system depends on the construction of a robust infrastructure for charging them. Authorities in the United States and around the world have recently been more and more skeptical about global warming, particularly in light of rising levels of greenhouse gas emissions. According to the U.S. Environmental Protection Agency (EPA), transportation is the main cause of greenhouse gas (GHG) emissions in the country, producing about 29% of all GHG emissions [1]. Promoting the use of electric cars (EVs), which are a well liked remedy for the pollution produced by automobiles running on fossil fuels, has grown more important as the hazard of climate change grows. Electric vehicles are essential for California to achieve its aspirational environmental and air quality targets. Without these cars, which have no tailpipe emissions, the state's goal of having 1.5 million zero-emission vehicles on Californian roads by 2025 would not be accomplished. Nevertheless, many important private sector automakers, such as Ford, BMW, Volvo, and others, have stated that they would switch to totally electric vehicles within the next ten years. If all of these changes in the public and private sectors were implemented, the demand for electric vehicles would skyrocket (EVs). To facilitate this shift, the state will need to invest a large amount of money in the infrastructure for electric vehicle charging. The California Energy Commission estimates that by 2030, the state may need up to 1.2 million EV chargers to support the eight million electric passenger cars that are

2

anticipated, as well as an additional 157,000 chargers to support non-passenger vehicles like trucks and buses. The charging needs of trucks and buses differ greatly from those of individual automobiles in terms of power demands, locations, and access. In addition to the $384 million the state would receive from the federal government over the following four years, the California Energy Commission (CEC) and the California Air Resources Board (CARB) recently announced that the state of California would contribute more than $5.5 billion to this effort.

In the current landscape of urban transportation [2], heavyweights in the form of medium/heavy-duty electric vehicles (EVs) find themselves in the spotlight due to their pivotal role in reducing emissions. However, a notable gap exists in the availability of tailored charging

infrastructure to support these vehicles. The existing charging infrastructure predominantly caters to light-duty vehicles, revealing a disparity that calls for a robust and dedicated solution. This underscores the urgency for a comprehensive analysis and enhancement of the charging ecosystem, ensuring that all segments of the EV market are adequately served.

Within the broader context of smart city initiatives, our project aligns seamlessly with the vision of a sustainable and energy-efficient urban environment. By contributing to the development of a green urban blueprint, we transcend the conventional notion of charging infrastructure as a mere means of powering vehicles. Instead, our focus extends to shaping a smarter and greener urban landscape. Beyond the act of charging, our project seeks to integrate seamlessly into the fabric of smart cities, fostering sustainability and contributing to the holistic development of urban spaces. In essence, it encapsulates a forward-thinking approach that not only addresses the charging needs of heavy-duty EVs but also envisions a future where urban mobility is intricately linked with ecological balance and energy efficiency. Medium and heavy duty electric vehicles, such as buses and trucks, play a crucial role in mitigating carbon

3

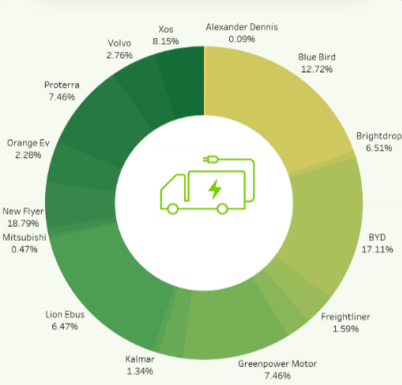
emissions [3] and improving overall energy efficiency within the transportation industry. In the dynamic urban environment of San Jose, continual construction endeavors and infrastructure enhancements are prevalent. Essential for the transportation of materials, execution of construction duties, and playing a pivotal role in the city's urban advancement, heavy-duty vehicles, including construction trucks and equipment, are indispensable. Medium and heavy duty vehicles play a vital role in San Jose's public transit network, serving as essential elements. They offer a crucial means of transportation for residents, aiding in the alleviation of traffic congestion and diminishing reliance on individual car usage.

In the dynamic landscape of sustainable transportation, pioneers in the medium and heavy-duty electric vehicle market, including trailblazers like Blue Bird, Lion Electric, and BYD, are at the forefront of driving innovation as shown in Figure 1. These companies are instrumental in the electrification of school buses, trucks, and diverse commercial vehicles, catalyzing a shift toward environmentally friendly practices. Their commitment to advancing

electric commercial vehicles underscores a collective endeavor to reduce carbon footprints and enhance the sustainability of transportation networks. The sector is witnessing notable technological trends, including advancements in fast-charging technologies, battery-swapping solutions, and smart grid integration, underscoring a commitment to efficiency and innovation in the EV charging ecosystem. Moreover, as environmental consciousness [4] intertwines with economic considerations, consumer adoption of electric medium and heavy-duty vehicles is on the rise. Fleet operators, logistics companies, and municipal services are increasingly drawn to electric vehicles, recognizing the dual benefits of addressing environmental concerns and realizing potential cost savings over time. This collective momentum among industry pioneers,

4

charging infrastructure leaders, and a growing consumer base reflects a transformative shift towards a more sustainable and electrified future in commercial transportation.

 **Figure 1**. Pioneers of Electric Commercial Vehicles

This project seeks to perform an extensive data-driven examination of the charging infrastructure that supports medium and heavy-duty EVs. By exploring essential metrics, usage trends, and challenges linked to charging infrastructure, the project aims to offer valuable insights that will guide strategic decision-making, policy creation, and the development of infrastructure. The primary problem targeted for the project is to address the limited availability

5

of EV charging stations for commercial transportation by Placing optimal locations of New Charging Stations for each vehicle type such as Transit Buses, Delivery Trucks, and School Buses in the community, and to support the transition to be a more sustainable and environmentally friendly transportation system. There are three objectives of the project Firstly, to reduce carbon footprint and improve air quality. Secondly, to promote the widespread adoption of public electric vehicles by providing an affordable and accessible charging infrastructure. Furthermore, it will help to support the growth of the EV markets, creating new jobs and economic opportunities in the city. To achieve these goals, the project approach involves four tasks. Initially, the proposal entails forecasting the demand for Heavy Duty and Medium Duty Electric Vehicles (EVs) along with the required charging stations. The plan involves developing a predictive model using the Prophet model to anticipate EV demand, encompassing vehicle counts and the corresponding number of needed charging stations, up until the year 2035. We will assemble an extensive dataset containing information on vehicle quantities, zip codes, fuel types, vehicle brands, duty classifications, and production years. Secondly, Predicting Vehicle Range of Heavy Duty and Medium Duty EVs. Forecasting the range of a vehicle through charging data involves grasping the variations in the vehicle's state of charge (SOC) over time and comprehending its consequential impact on driving distance. Enhancing the operational efficiency of electric vehicles, including transit buses, school buses, and delivery trucks, entails making predictions about their range using charging data. Our research is centered on leveraging analytics of charging data to estimate the range of these three vehicle categories, aiming to enhance decision-making in fleet management. The third goal is Predicting Short and Long-term Energy Demand for Heavy Duty and Medium Duty EVs. The effective utilization of energy resources, the streamlined operation of diverse vehicle fleets—

6

ranging from transit buses to school buses and delivery trucks—and the optimization of charging infrastructure all hinge on the capability to anticipate the energy consumption of electric vehicles (EVs). This use case incorporates both long-term and short-term projections to meet various planning and operational goals. Finally, strategically locating new charging stations to align with the charging needs of diverse vehicle types to promote the uptake of electric vehicles (EVs) and advance sustainable transportation practices. After finishing these tasks, we will assess the performance of our models using metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2. Additionally, precision, recall, and F1 score will be considered for predictions on a station-wise, site-wise, and zipcode-wise basis. Subsequently, our attention will shift towards the development and implementation of the system, incorporating predictions for new charging station locations tailored to each vehicle category, namely school buses, transit buses, and delivery trucks, within a specified zip code in San Jose city of California.

Our project prioritizes the advancement of electric vehicles (EVs) for use in public transportation systems within the community. Through these initiatives, we anticipate fostering innovation in clean energy technology and the infrastructure of EV charging stations within the community.

The primary contributions of our project are outlined below:

1. Strategic Infrastructure Planning: Through a thorough examination of charging infrastructure, the project plays a pivotal role in strategic planning for the establishment of charging stations catering to medium and heavy-duty electric vehicles. This encompasses discerning insights into ideal locations, charging capacities, and the specific types of vehicles accommodated.

7

2. Anticipating Future EV Demand: A key contribution lies in our capacity to predict the demand for both heavy-duty and medium-duty electric vehicles. This foresight empowers stakeholders, including policymakers, manufacturers, and infrastructure developers, to anticipate the expanding market and adapt their strategies accordingly.

3. Forecasting Energy Consumption: The project's ability to predict the energy consumption of heavy and medium-duty vehicles is a significant feature. By understanding the fluctuations in the state of charge (SOC) over time and how it impacts driving distance, the project aids in forecasting the energy needs of these vehicles, which is pivotal for ensuring efficient fleet management.

4. Long-Term and Short-Term Planning: Forecasting the energy demand for medium and heavy-duty vehicles in both the long and short term has multifaceted benefits. It aids in strategic planning, enhances operational efficiency, facilitates optimal resource allocation, promotes cost optimization, fosters environmental sustainability, encourages technology innovation, and allows adaptability to market changes. The inclusion of both long-term and short-term projections within the project is a significant asset, ensuring that the predictions address immediate operational requirements while also facilitating strategic planning for the future.

The findings from our project, which include forecasting EV count and charging station demand, predicting vehicle range, and anticipating both long and short-term energy demand, offer numerous benefits. Initially, the application of the Prophet model to forecast EV count and charging station demand creates a data-driven guide for future EV infrastructure development. This holds significant value for legislators, utility corporations, and infrastructure planners, facilitating more efficient resource allocation and investment decisions through precise estimates

8

of car counts and charging station needs. The second advantage lies in predicting vehicle range through the application of a stacked regressor model. This proves advantageous in comprehending the fluctuations in the vehicle's state of charge (SOC) over time and how it influences driving distance. Such understanding is essential for enhancing the operational efficiency of electric vehicles such as transit buses, school buses, and delivery trucks. Our research concentrates on leveraging charging data analytics to gauge the range of these vehicle types, thereby enhancing decision-making in fleet management. Finally, the prediction of both short and long-term energy demand using a temporal fusion transformer model offers significant advantages by addressing both temporal scopes. Long-term energy demand forecasting provides vital insights for strategic planning and resource allocation, encompassing weekly, monthly, and yearly perspectives. This equips fleet managers and infrastructure planners with the ability to proactively manage variations in energy demand and ensure the availability of ample charging resources. On the other hand, short-term energy demand forecasting is focused on predicting energy needs for the upcoming day or within a shorter timeframe, typically within the next 6 hours. This form of forecasting plays a critical role in energy load balancing and making real time operational decisions, contributing to heightened operational efficiency, minimized energy waste, and the promotion of environmentally friendly transportation practices. The targeted problems along with the project deliverables are outlined in the following Table 1. **Table 1.** Targeted Problems and Project Deliverables

**Targeted Problems Project Deliverables**

Anticipating EV Market Growth

Count Forecast of Heavy/Medium Duty EV Vehicles and Charging stations using Prophet.

9

**Table 1.** *Cont.*

**Targeted Problems Project Deliverables**

Addressing Range Anxiety Range Prediction using Stacking Ensembler Regressor Model.

Efficient Energy Resource Allocation

Strategic Charging

Infrastructure

**1.2 Project Requirements**

EV Charging Station Energy Demand Prediction using Temporal Fusion Transformer Model.

Optimal Location Prediction for New Charging Station using PuLP linear programming optimization model with K-means clustering model.

The aim of this research focused on the data-driven analysis of EV charging station infrastructure for medium and heavy-duty vehicles, is to utilize data and analytical methods to comprehend the current status and anticipate future requirements of EV charging infrastructure. To identify locations in need of additional charging stations, it is imperative to analyze data on the current charging station locations, usage patterns, and capacities, as well as demographic and geographic information. The project involves the application of predictive modeling and simulation techniques to forecast future demand for charging infrastructure and optimize the strategic placement and design of additional charging stations. Ultimately, the project seeks to offer insights to policymakers, infrastructure planners, and other stakeholders, facilitating the creation of a more efficient and effective charging infrastructure to support the transition to electric vehicles.

Functional requirements refer to the attributes and capabilities that a system must possess or be able to achieve. Regarding data collection and management, the system should have the capacity to gather data from diverse sources, including information from charging stations, vehicle registrations, traffic statistics, California zip codes, transit bus stops, and vehicle fleet

10

transactions, consolidating it in a centralized database. To assess the system's capability to handle extensive data while upholding data integrity, it becomes crucial to scrutinize the completeness and quality of the acquired data. The system should possess the ability to analyze the gathered data, identifying patterns, trends, and insights about EV charging station usage, such as peak hours, popular locations, and charging durations. The accuracy of the analysis and the relevance of the generated insights can be used to evaluate this requirement. Additionally, the system should be equipped to forecast future requirements for EV charging infrastructure based on current usage patterns and other relevant factors like industry expansion, shifts in consumer behavior, and modifications in governmental regulations. This can be verified by comparing anticipated demand with actual consumption trends over time. Taking into account variables such as location, capacity, and user requirements, the system must possess the capability to optimize the placement and design of EV charging stations to meet both current and future demand. The effectiveness and efficiency of the optimized charging infrastructure can be assessed to verify this requirement. Additionally, the system should be adept at generating reports and visualizations that communicate outcomes and recommendations to stakeholders. The evaluation of the usefulness and relevance of these reports and visualizations to stakeholders serves as a test for this requirement. Furthermore, the initiative should outline specifications for seamlessly integrating data and analysis outputs with existing systems and platforms. Testing these specifications involves assessing the ease of integration with current systems and the accessibility and utility of shared data and analytical outputs for end-users. The functional requirements of our project are comprehensively outlined in Table 2.

Utilizing AI features enhances the system's sophistication and autonomy. The forecasting capabilities of the Prophet model and Neural network architecture particularly in forecasting

11

**Table 2.** Functional requirements of the project

**Task Functional Requirement**

Data collection and management

Data analysis and modeling

• Ability to collect data from multiple sources.

• Data cleaning and preparation to ensure accuracy and consistency. • Robust data management system to store and organize data. • Ability to update data regularly.

• Ability to perform statistical analysis on the data.

• Ability to develop predictive models to forecast future demand for charging infrastructure.

• Ability to simulate different scenarios to optimize the placement and design of charging stations.

• Ability to generate visualizations and reports to communicate insights and recommendations.

Efficiency Evaluation • Ability to assess the efficiency and efficacy of the optimized charging infrastructure to ensure optimal performance.

• Ability to assess how well the optimized infrastructure contributes to

reducing carbon emissions and promoting sustainable transportation

practices.

• Ability to use Continuous monitoring and feedback mechanisms to

contribute to ongoing improvements and the sustained efficiency of the

EV charging system.

Data Visualization • Ability to display charging station locations on an interactive map. • Ability to display real-time data on charging station availability for

each vehicle type, occupancy, and charging speed.

• Ability to customize the visualization based on user preferences.

Integration with existing systems

Ease of Integration Assessment

• Ability to integrate with geographic information systems (GIS) to visualize data on maps.

• Ability to integrate with transportation planning and decision-making tools.

• Ability to share data and analysis results with stakeholders.

• Ability to assess the ease of integration with the existing systems and the availability of shared data and analytical outputs for customers. • Ability to assess the robustness and flexibility of the system's Application Programming Interface (API).

12

charging station usage, energy demand, and availability, can be assessed by comparing predicted values with actual results and evaluating predictive accuracy. The efficacy of optimization algorithms in improving the positioning and design of new charging stations can be gauged by contrasting enhanced solutions with real-world outcomes. Pattern recognition algorithms driven by AI can be evaluated for precision and efficiency in identifying patterns and anomalies in charging station data. This includes recognizing stations frequently out of service or detecting consumption patterns inconsistent with typical usage. The effectiveness of AI-powered recommendation engines in suggesting optimal charging stations based on customer preferences is assessable by comparing recommendations with actual usage patterns. To test the system's robustness, performance, and reliability under different stress and load conditions, experiments can be conducted. This evaluation determines how well the AI-powered system can manage substantial data volumes and operate reliably and efficiently across various scenarios. Details regarding the AI requirements of the project are outlined in Table 3.

To assess the distribution and concentration of charging stations across diverse zip codes and counties in California, obtaining precise location data for charging stations is essential. A comprehensive understanding of electric vehicle counts and the demand for charging infrastructure, especially in areas with either underutilized or overburdened charging stations, necessitates access to vehicle registration data. This includes information on fuel types and vehicle manufacturers. Analyzing the energy demand of charging stations requires charging transaction data, encompassing details such as the starting state of charge (SOC) and ending SOC. For effective analysis and decision-making tailored to specific locations, geographic and environmental data play a pivotal role. This data may include population density, traffic patterns, and environmental considerations, aiding in pinpointing areas that require additional charging

13

**Table 3.** AI requirements of the Project

**AI Requirement Description Measurable Metric**

• The ability to use historical data to

forecast Electric Vehicle count and

future demand for Charging stations.

Predictive modeling

Sequence to

Sequence modelling

Linear Programming plus Clustering

analysis

Visualization

Optimization

algorithms

• The ability to use EV transaction data to predict vehicle range to optimize the operating efficiency of EV.

The ability to use EV transaction data and consider the temporal elements like day of the week, month, and year to determine EV charging station energy demand.

The ability to utilize the capabilities of linear programming and clustering method to suggest optimal locations for EV charging stations

The ability to present data in a clear, intuitive, and interactive way, such as maps, graphs, and heatmaps

The ability to optimize the placement and design of new charging stations based on various criteria, such as vehicle range, EV demand, and accessibility

Prediction accuracy, error rates, and correlation coefficients.

Prediction accuracy, temporal accuracy, seasonal pattern

recognition, peak demand

prediction, utilization efficiency.

Clustering analysis, precision values, recall values and F1 score.

User engagement metrics, such as time spent on the platform, clicks, zooms, and searches.

Charging station utilization, coverage and accessibility, EV range coverage, demand satisfaction and user satisfaction.

infrastructure based on requirements and constraints. To predict the demand for charging infrastructure and optimize the design and capacity of charging stations, detailed information on the usage of electric vehicles becomes paramount. This includes data on the quantity, types, and battery capacities of medium and heavy-duty electric vehicles. Understanding public transportation involves sourcing transit bus stop data for San Jose and providing insights into the locations of bus stations. Additionally, obtaining school bus terminal data for San Jose

14

contributes to a comprehensive overview of school bus terminal locations. For a deeper understanding of commercial fleet operations, heavy and medium-duty vehicle fleet transaction data has been acquired. This dataset encompasses real-world data for various weight classes of commercial fleet vehicles. In summary, the table labeled Table 4, provides a breakdown of the diverse data aspects essential for the successful execution of this project.

**Table 4.** Data requirements of the project

**Data Requirement Description Source** Number of vehicles registered each

CA Vehicle registration data CA EV charging stations data

year for every zip code and county in CA

Real-time data of EV charging stations for each zip code in CA

Data of public bus station locations for

California Department of Motor Vehicles

U.S. Alternate Fuel Data Center

Transit bus stops data

transit bus in San Jose CA open data portal

School bus terminals data

Heavy/Medium vehicle fleet transaction data

Data of school bus terminal station locations in San Jose

Real-world data of commercial fleet vehicle operating data for each weight class

CA Department of

Education

National Renewable Energy Laboratory

The team comprises members dedicated to specific use cases outlined in the project, including: 1) Forecasting electric vehicle count and charging station demand for Medium and Heavy-duty EVs using time series modeling, 2) Predicting the vehicle range of Medium-duty and Heavy-duty EVs through ensemble machine learning techniques, 3) Projecting short and long term energy demand of medium and heavy-duty EVs using a sequence-to-sequence modeling approach, and 4) Optimizing the placement of new charging stations for each vehicle type using

15

linear programming and clustering methods. Responsibilities for these tasks have been distributed among team members to ensure the timely achievement of objectives. Each member's roles and responsibilities are detailed as follows:

**Table 5.** Team members and their allocated roles

**Task Assigned**

Data Extraction Lohitha, Mahe,

Pranavi

Data Cleaning Pranavi, Rohan

Data Transformation Pranavi, Lohitha

Feature Selection Mahe, Lohitha,

Rohan

Exploratory Data Analysis All

Forecasting EV count and charging station demand of Medium and Heavy duty EVs based on electric vehicle registrations

Predicting vehicle range of Medium and Heavy duty EVs based on Medium/Heavy duty transactions

Predicting short and long-term energy demand of Medium and Heavy duty EVs based on Medium/Heavy duty transactions

Optimal placement of new charging stations for each vehicle type such as school bus, transit bus and delivery truck

Pranavi, Mahe Lohitha, Mahe Pranavi, Rohan Lohitha, Rohan

Final Report and Presentation All

**1.3 Project Deliverables**

As part of the project, deliverables include a project proposal that defines the research problem statement, related references, and a background survey. Using the Cross-Industries Standard Process for Data Mining (CRISP-DM), these deliverables are divided into various

16

stages. A literature review is conducted to identify the research gap, followed by an analysis of possible technology solutions to the problem statement. The project's next phase begins with an introductory chapter that describes the project's requirements and defines the scope of the project. Among the deliverables of the third phase is the Data and Management Plan for the project, which discusses how data is collected, as well as the insights gained by studying the given data, as well as how the tools and software will be utilized with cost justifications.

As part of this phase, a Work Breakdown Structure (WBS) will be provided, displaying each task's timeline. WBS can be created with project management tools such as JIRA or ClickUp. According to CRISP-DM, this matrix shows the hierarchical representation of work packages, tasks, and deliverables of a project. A research problem requires a certain amount of effort to solve. This is the effort estimation. Based on the project's complexity and the availability of individual team members, various processes are estimated. In agile methodologies, work is estimated using story points. You can calculate the time required to complete a user story based on the complexity of the task. There are two types of project scheduling charts: Gantt charts and PERT charts. There is significance to each of them. A Gantt chart is a graphical representation of timelines, tasks, sub-tasks, dates, milestones, and resources allocated to each task/sub-task. By using a Gantt chart, it shows how the subtasks are related to one another. The PERT chart is used to organize and schedule projects by breaking down, mapping, and visualizing tasks. An estimate of the time needed to complete a project can be made using a PERT chart.

As soon as the data and project management plan are complete, the next phase will be data engineering. In this phase, data is preprocessed, transformed including dimensionality reduction, and provides the final data set for training and testing the modeling models. In this

17

phase, algorithms' performance parameters are measured and compared. Each team member will submit a report explaining the model's features, framework, efficiency, and limitations. The final deliverables include California, will be able to manage its electric vehicle recharging infrastructure comprehensively. The proposed solution comprises an interactive website interface that is designed to facilitate user interactions by allowing queries for analytical results and providing a user-friendly mechanism for suggesting new charging station locations. Fleet managers can use this information to ensure the smooth and effective running of their EV fleets by learning the best charging locations. Users can benefit from the system's accuracy in predicting optimal charging station locations, enabling them to plan routes and optimize charging stops. The results of the various experiments on the datasets will be compared in a table comparing the results of these models. At the end of the project, the group submits a report and presentation. The project deliverables are listed in Table 6, along with their corresponding descriptions and timelines.

**Table 6.** Project Deliverables with their respective description and due dates. **Deliverable Description Due Date** Project Abstract Contains a literature survey and a sample data file. In

addition, it contains the background, examples of data

sources, motivation, and the problem statement for the

study.

Data Preparation Plan The study of data generation models and the understanding of the public transportation management

system helps fleet operators better manage charging

schedules.

Project Introduction The document includes the project background and executes summary, project requirements, project

17 February 2023 10 March 2023

deliverables, technology, and literature surveys. 15 March 2023

18

**Table 6.** *Cont.*

**Deliverable Description Due Date**

Work Breakdown Structure

In a hierarchical structure, the work must be done by the project team to produce the required deliverables. This breakdown includes the start and end dates of each task.

22 March 2023

Clean Data Set An exploratory data set using dimensionality reduction and exploratory analysis, ready for modeling. 10 April 2023

Data Engineering This stage provides a complete explanation of the data

process, data collection, data pre-processing, data

preparation, data transformation, and data statistics

steps.

Research Report The report contains all the necessary details about the research, such as how the data was transformed, project

requirements, and methodologies.

14 April 2023 12 May 2023

System Design and Implementation

Model Evaluation and System Visualization

Web system design and

This document contains the system development process, where the detailed structure and functionality of the system are defined.

This stage provides model improvements, system components, and processes, facilitating a clearer understanding of the system’s architecture and functionality.

Provides a user interface for a web-based application,

02 October 2023 13 November 2023

development

Research Report Presentation

creating a functional and interactive web system. 2 27 November 2023

ThThe final research report will be presented in a

PowerPoint presentation on this day. 11 December 2023

Final Project The project integrates with the public transportation management system to empower users to easily locate

optimal charging stations. By using the website, users

15 December 2023

can locate charging stations nearby, making it easier for them to adopt electric vehicles.

19

**1.4 Technology and Solution Survey**

In this section, configurations of multiple types of charging stations were examined to select the relevant technology for our use case. Later, the suitable models were scrutinized and compared on a higher level and then an in-depth comparison analysis among all the existing technological surveys was done which will give a clear overview of the model requirements of this research.

Table 7 provides a summary of different EV charging technologies and their respective specifications. It can be used to compare the effectiveness and efficiency of different charging technologies in a data-driven analysis of EV charging infrastructure. Here, various technologies for electric vehicle (EV) charging infrastructure are briefly summarized. These technologies include Level 1 and Level 2 charging, DC fast charging, wireless charging, smart charging, solar-powered charging, and battery swapping. Power output, connector type, network accessibility, location, utilization rate, and availability are a few of the criteria used to evaluate each technology. In both residential and commercial settings, level 1 and level 2 charging are typical. They provide medium power outputs and have affordable installation and maintenance expenses. DC fast charging is better suited for high-traffic locations due to its higher power outputs but higher installation and maintenance expenses. Physical connectors are not required for wireless charging; however, installation costs are higher and usage rates are lower. Smart charging features high utilization rates, low installation, and maintenance costs, and optimizes charging based on demand and grid circumstances. Although solar-powered charging has lower power outputs and moderate installation costs, it decreases pollution and reliance on the grid. While battery swapping offers quick charging times, it also necessitates greater room and more expensive infrastructure. Finding the best options for various locations and use cases requires

20

comparing and assessing various EV charging systems. To support the rising demand for electric vehicles, this information can assist in guiding decisions regarding where and how to invest in EV charging infrastructure.

**Table 7.** Technology Comparison Table for EV Charging

**Technology Level 1 Level 2 DC Fast Charging Wireless Charging** Power Output 1.4 kW 7.2 kW 50-350 kW 3.7-7.7 kW

Charge Time (0- 100%)

8-12 hrs 4-8 hrs 20-60 min 3-4 hrs

Connector Types NEMA 5- 15

J1772 CCS1, CCS2, CHAdeMO, Tesla Supercharger

Qi, SAE J2954

Compatibility Light Duty

All EVs Medium and Heavy Duty Limited

Cost Low Medium High Very High Convenience Low High Medium High Availability High High Low-Medium Low

Table 8 offers a thorough overview of the models utilized to analyze the infrastructure of electric vehicle (EV) charging stations. It describes the different models, their advantages and disadvantages, and particular use cases. Traditional regression models are used to determine the number of stations required, while more sophisticated methods such as Temporal Fusion Transformer (TFT) are used to forecast the demand for energy consumption in the smart grid. Each model is designed to a different set of requirements, providing advantages as well as limitations. Some of the models anticipate the occupancy of infrastructure, others identify clusters with similar usage patterns, and yet others optimize the placement of charging stations inside road networks. A model is chosen based on the specific use case and the intended result,

21

considering variables like interpretability, computational complexity, and handling of various data kinds.

**Table 8.** Technology Survey on Models Used for EV Charging Station Infrastructure Analysis

**Ref**

**Model Description Pros Cons Use case ID**

[5] Regression models

[6] Time

series

forecasting

models

[7] Neural networks

A statistical

model that

analyzes the

relationship

between

dependent and independent

variables.

Models that

analyze historical data to make

predictions about future trends.

A model that simulates the

behavior of the human brain to process complex data.

Easy to

interpret, can be used with different data types.

Can capture seasonal

patterns and make

predictions based on

historical data.

Can capture complex and non-linear

relationships, can handle large datasets.

May not capture complex

relationships or nonlinear patterns in data.

May not account for sudden

changes or

external factors that could impact trends.

Requires a large amount of data to train, can be

computationally intensive, and can be difficult to interpret.

Predicting the number of EV charging stations needed in a

specific location.

Forecasting the demand for EV charging stations in a specific

location.

Predicting the optimal placement of EV charging stations.

[8] Clustering A technique that groups similar

data points based

on their

characteristics.

Can identify patterns and relationships in data that may not be

immediately apparent.

Can be difficult to interpret and may require subjective

determination of the number of clusters.

Identifying clusters of EV charging stations that have similar usage

patterns.

22

**Table 8.** *Cont.*

**Ref**

**Model Description Pros Cons Use case ID**

[9] Prophet The prophet model is used to

identify and

simulate the

time-dependent

patterns and

structures that

are present in the

charging load

data for EVs.

Can manage unavailable

data efficiently, distribution of uncertainty

intervals.

Over

simplification of intricate patterns. Limited attention to outside

influences.

Load Forecasting of Battery Electric Vehicle Charging Stations.

[10] Gradient Boosting

Classifier

[11] Linear

Programming

[12] Temporal Fusion

Transformer

Model

A model that

predicts the

occupation status of charging

infrastructure, utilizing binary training data.

An optimization technique based on integer linear programming is used to place

electric vehicle charging stations strategically

throughout a

network of roads.

To address the increasing energy consumption in the residential sector

Manages

complex

interactions

well and

provides

excellent

predicted

accuracy.

By providing mathematical rigor and

efficiency,

linear

programming ensures the best possible

options for the location of EV charging

stations.

TFT provides flexibility in predicting day, weekly, and monthly energy use prediction horizons.

Can be

computationally demanding and prone to

overfitting,

particularly when dealing with

noisy data.

Practical

applications may face issues related to sensitivity to input data

accuracy and

significant

computing

complexity.

The complex

internal workings of the model

present

interpretability issues.

Predict charging infrastructure

occupancy.

Efficient

Placement of

Electric Vehicles Charging Stations.

Forecasting energy consumption

demand of

customers in the smart grid

23

Table 9 provides a thorough summary of various research initiatives targeted at tackling significant components of integrating electric vehicles (EVs) into transportation networks. Every topic describes the objective, methods, and important findings of the corresponding studies and covers a variety of subjects, including energy-efficient routing, charging optimization, infrastructure planning, and charging station placement techniques. The research emphasizes the significance of evaluating different charging systems, investigating the expansion of electric vehicles in commercial sectors, and utilizing cutting-edge techniques such as SHapley Additive exPlanations (SHAP) to improve energy efficiency. It also recognizes the difficulties and constraints these studies face, from operational adaptability issues and computational complexity to the requirement for more thorough analyses, like comparing the evaluations of urban and rural routes or validating machine learning models against external datasets. In the ever-changing field of electric vehicle integration and sustainable transportation planning, the survey provides a useful resource by highlighting the developments, gaps, and possible directions for further research.

**Table 9.** Technology survey methods of different Electric Vehicles (EVs) **Objective Methods Description Negatives**

Planning Electric vehicle infrastructure for various types of vehicle

Heavy- and medium-duty vehicles

[13]

Heavy-duty trucks only [14]

Uses the Return-to-base model and the on-route charging model, and

summarizes the challenges of charging commercial electric vehicles (CEVs) at public locations.

Joint routing and charging (JRC) scheduling approach for electric trucks, to

minimize costs and delivery delays.

Need for a balanced

exploration of alternative charging strategies beyond the return-to-base model and V2G technology.

Computational complexity of the optimization problem, operational adaptability in environments.

24

**Table 9.** *Cont.*

**Methods Description Negatives**

Utilizing the Energy Demand of Electric vehicles (EVs)

Methods used for locating charging stations

E-Trucks & Buses

[15]

Shapley

Additive

exPlanations (SHAP)

[16]

Energy

efficient route planning

[17]

Hybrid

simulation model

[18]

Activity

based

approach

[19]

Route-based analysis

[20]

GIS-based approach

[21]

Growth of electric trucks and buses by industry

engagement, facilitating the electrification of vehicles.

Examines (BEVs) to identify factors influencing BEV energy consumption and provide insights for

enhancing energy efficiency and informing transport policies.

Estimating energy

consumption in electric vehicles compared to

conventional methods.

Examine EVs through strategies addressing

greenhouse emissions from electricity generation and EV adoption in passenger transportation.

Locate the (sub)optimal locations for charging stations using multiday travel activity constraints.

Analyze the infrastructure development and EV

efficiency in minimizing travel time and maximizing user comfort.

Identify the optimal

locations for EVSE in terms of improved efficiency and lower emissions.

There is a need for

adjustments in utility rates and a lack of adequate charging infrastructure.

More detailed comparative analysis of urban and rural routes, and validation of ML models against external datasets.

Can improve by providing more detailed results of statistical tests.

Oversimplified alternative options, lack of a detailed transportation demand forecast, and unclear model validation.

Lack of information on grid stability and electricity demand charging

infrastructure utilization.

Uncertainty in optimal charging stop selection and waiting times was not included in the study.

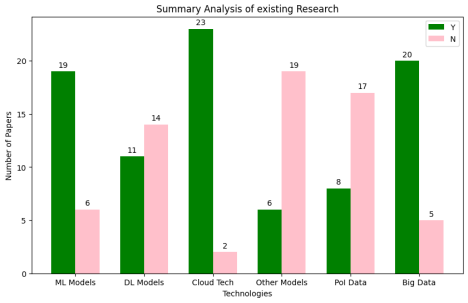
Lack of available specialists in electric mobility and incomplete vehicle

registration data.

25

**1.5 Literature Survey of Existing Research**

The objective of this section is to organize the literature survey on sustainable and efficient deployment of Heavy Duty and Medium Duty EVs: forecasting demand for EV and charging stations, predicting vehicle range, estimating short and long-term energy demand, and optimizing the placement of new charging stations to identify gaps in the field and, more importantly, to establish a standard for measuring our model execution performance. Several pertinent studies have already been cited in earlier sections. To categorize the overall strengths and shortcomings of those methodologies, we assess our findings holistically and suggest prospective paths for research into the field of heavy-duty and medium-duty electric vehicles (HEVs) and their associated charging infrastructure.

 **Figure 2**. Analysis of existing research

26

***1.5.1 Forecast Heavy Duty and Medium Duty EV and Charging Stations Demand*** The increasing use of Heavy Duty and Medium Duty Electric Vehicles (EVs) makes a detailed analysis of their demand necessary in the future. The current literature on predicting demand for these EVs and the related charging stations is examined in this review of the literature. A significant portion of research employing the ANN, RNN, LSTM, and Transformer models is accessible. To have a better understanding of implementation, a comparison of research publications based on various models is done in Table 10.

**Table 10.** Comparison of existing research on EV and Charging Stations Demand

**Ref ID**

**Objective Region Dataset ML Models**

**DL**

**Models**

**Other models**

**Eval**

**Metrics**

[22] Forecasting medium term public EV

charging demand

Scotland, UK

Eleven charging stations

LinReg, RF,

SVM,

KNN

ANN ARIMA MAE, RMSE,

SMAPE

[23] Forecasting demand for EV charging

post-session start

Morocco Two Public charging

stations

N ANN, RNN,

LSTM,

GRU

N MAPE, RMSE

[24] EV charging demand predictions for short

term and long-term

forecasting

Boulder, Colorado

Twenty-five Public

charging

stations

N ANN, LSTM,

Transfor

mer

ARIMA, SARIMA

RMSE, MAE

[25] Addresses the challenges of mass

electric vehicle (EV)

adoption on power

systems

Shenzhen China

real-world EV charging station

datasets

N DNN, RNN,

LSTM,

GRU

N NRMS, normalize

d mean

absolute

error

(NMAE)

[26] Smart EVC

solution—intelligent

charging station

management

platform based on AI

United Kingdom

real-time data on the energy consumption and charging demand.

N N Custom Reservation

Algorithm

Reservation

Charging time

27

**Table 10.** *Cont.*

**Ref ID**

**Objective Region Dataset ML Models**

**DL**

**Models**

**Other models**

**Eval**

**Metrics**

[27] Forecast electric vehicle charging

demand

Atlanta, USA

EV charging dataset of

Georgia Tech

N Deep Learning

-

based

LSTM,

DLSTM

arithmetic optimizatio n algorithm (AOA),

empirical mode

decomposit ion (EMD).

MAE,

RMSE, Accuracy

***1.5.2 Predicting Vehicle Range of Heavy-Duty and Medium-Duty EV***

Precisely forecasting the range of Heavy Duty and Medium Duty Electric Vehicles (EVs) is essential to performance optimization in the ever-ch anging world of electric transportation. Through an exploration of techniques and innovations, this study seeks to provide information that is essential for improving operational effectiveness and resolving customer issues. A comparison of research publications based on different models is done in Table 11 to provide a detailed comprehension of implementation.

**Table 11.** Comparison of existing research on Vehicle Range

**Ref**

**ID**

**ID**

**Objective Region Dataset ML Models Deep Learning**

**Other models**

**Eval Metrics**

[28] Analyze the effects of

different

environmental

parameters on

energy

consumption

and driving

range

Various regions of China

Real-world BEV

N N Micro simulation

model

Accuracy, Relative error,

Absolute error.

28

**Table 11.** *Cont.*

**Ref ID**

**Objective Region Dataset ML Models Deep Learning**

**Other models**

**Eval Metrics**

[29] Improve the driving range

prediction

accuracy

Beijing Real-world data from

Baic New

Energy

Automobile

Classification and

Regression tree, MLR, GBDT

N N MAE, Accuracy

Maximum,

minimum

error

[30] Reduce

driver’s range

anxiety by

estimating the

real-time

energy

consumption

of EVs

United States

Nissan Leaf 2013,

Argonne

National

Laboratory (ANL)

N Deep

Convolutional

Neural

Networks

N RMSE, MAE, K-fold

correlation

[31] Predict the remaining

driving range

of EVs

Major cities in China

National Big Data

Alliance of New Energy Vehicles

XGBT,

Boosting Regression Tree,

XGBoost

N Blended Model

MAE, RMSE, MAPE

[32] Accurately estimating the

driving range

of electric

vehicles

China National Monitoring

and

Management

Platform for

New Energy

Vehicles

Gradient

boosting

decision tree (LGBM),

SVM

N N MAE, MSE, RMSE, R2

(R-Square)

[33] Impact of driving

electric

vehicles

(EVs) at

highway

speeds, using

auxiliary

loads

Australia Mitsubishi i MiEV and

Nissan Leaf

test car data

N N Vehicle mathematical

modeling

Mean

unsigned

error, Energy consumption

29

***1.5.3 Predicting Short and Long-term Energy Demand of Heavy-Duty and Medium-Duty EV*** Predicting the energy consumption of Heavy Duty and Medium Duty Electric Vehicles (EVs) both in the short and long term is critical as we move toward a sustainable transportation paradigm. In addition to addressing the changing environment of charging behavior and grid interactions, this research examines research methodology and insights into factors driving energy usage patterns. Table 12 presents an extensive overview of implementation through a comparison of research findings based on various models.

**Table 12.** Comparison of existing research on Energy Demand

**Ref ID**

**Objective Region Dataset ML Models**

**Other models Eval Metrics**

[34] Predicting the energy usage during charging sessions for plug-in

electric vehicles

(PEVs)

Nebraska, USA

public

charging stations

RF,

SVM,

Xgboost, LinReg

N RMSE, MAE, R2

[35] Predicting the energy demand for EVs

Dundee city, UK

Charging stations data(CS)

RF, DT, KNR,

SVR,

SGDR

Federated

Energy Demand Learning

(FEDL),

Clustering based EDL,

RMSE

[36] Predicting the energy consumption in electric vehicles (EVs)

Brussels Controller Area Network

(CAN) bus

signals of a

2012 Nissan

Leaf

Multiple linear

regressio n

Macro and hybrid model

correlation coefficient (R2),

Accuracy

[37] China's road transport sector's energy

consumption and

greenhouse gas

emissions at the

provincial level up to

2050

Mainland China

National

Bureau of

Statistics of China, Annual statistic of

population

N CPREG model, GHG Emissions

Analysis model

BAU

scenario, LC

scenario, GDP

growth rate

30

**Table 12.** *Cont.*

**Ref ID**

**Objective Region Dataset ML Models**

**Other models Eval Metrics**

[38] Analyze future trends in China's road

transport sector,

evaluating direct and

life cycle energy

demand and

greenhouse gas

emissions

China China

Automotive

Technology

and Research

Center

(CATARC)

data

N Gompertz curve model, bottom

up model

BAU

scenario, HEV

application rate

[39] Reliability and sustainability of smart

city transportation

systems

United States

Electric Vehicle Charging Dataset Kaggle

KNN,

DT, RF, SVM

Deep Neural Networks

(DNN), LSTM

MSE,

Gradient loss,

Accuracy, MAPE

***1.5.4 Optimal Placement of New Charging Stations for Heavy Duty and Medium Duty EV*** The positioning of charging stations is crucial for the successful integration of Heavy Duty and Medium Duty Electric Vehicles (EVs) in the context of electric mobility. The research on the best locations for new charging stations considers grid efficiency, user convenience, and accessibility. The study's analysis of current approaches aims to put an insight into important considerations for the construction of strategic infrastructure. Similar to previous modules, a brief comparison of studies on the optimal placement of new charging stations given in Table 13. **Table 13.** Comparison of existing research on Optimal Placement of New Charging Stations

**Ref ID**

**Objective Region Dataset ML Models**

**Deep**

**Learning**

**Other models**

**Eval Metrics**

[40] Optimizing the selection

of EV

charging

station

SingaporeUSA, UK

Road

network data,

charging station data

N N Proximal Policy

Optimization

algorithm,

RL

profit increase, travel time, charging time

31

**Table 13.** *Cont.*

**Ref ID**

**Objective Region Dataset ML Models**

**Deep**

**Learning**

**Other models**

**Eval Metrics**

[41] Resolving EV charging

infrastructure

planning

Sydney, Australia

Traffic

flow

Distribution network

data

N CNN (GCN)

Cournot

competition game model

Travel time, wait time, charging time

[42] Determining the

placement of

charging

stations

Germany Charging Station data

LinReg N N MAE, MAPE, RMSE

[43] Optimization model using

GIS data and

grid

partitioning

to locate and

size public

charging

stations for

EVs.

Thailand GIS data, Vehicle

Population

N N Optimization and spatial

analysis

models

EV–CP ratio, Service time

[44] Allocating charging

stations in

large-scale

transportation

networks for

electric

vehicles

(EVs)

[45] Address the multi-stage

placement of

electric

vehicle (EV)

charging

stations

Southern

Sweden

San Pedro

District of Los Angeles.

The

Swedish National Road

Database (NVDB)

zip code tabulation area

(ZCTA), Road

network data

N N Probabilistic rule, Integer

programming

for solving

the

optimization

problem

N N Nested logit model,

Bayesian

game

analysis

Maximal route cost

Charging

Demand, Total Profit, Overall Utility.

32

The literature review on the analysis of electric vehicle (EV) charging infrastructure offers a thorough look at previous studies and research that have been done on different facets of EV charging networks is given in Table 14. The research was conducted on forecasting demand for EVs, including heavy- and medium-duty models, as well as the vehicle's range and short- and long-term energy requirements. Also, explored the optimal placements for new charging stations, considering factors such as geographic locations, traffic patterns, and user behavior. An increasing amount of information about the intricacies of EV charging infrastructure, including user behaviors, technological developments, and the effect on electricity grids, is available in the literature survey. Geospatial analyses have been used to investigate the best locations for new charging stations, highlighting the significance of key areas to fulfill the increasing demand for electric mobility. By synthesizing these diverse strands of research, the literature review serves as a valuable resource for shaping the future of electric vehicle charging infrastructure and fostering sustainable transportation solutions.

**Table 14.** Literature Review in Analyzing Electric Vehicle Charging Infrastructure

**Ref**

**Region Purpose Models Evaluation Metrics ID**

[46] Adaptive Charging Network

[47] Santa Monica, California

Finding clusters of EV charging behavior and classifying future sessions to predict which cluster they belong.

Predicting charging session duration and energy

consumption

K- means.

KNN

Hybrid

estimator

using GKDE and DKDE

Accuracy: 97.9 AUC: 0.99

MED: 0.74 hr MED: 1.68 Kw/hr

[48] Shenzhen China Predicting hourly charging load of public stations

RNN Based Models

GRU NRMSE: 2.89 %

33

**Table 14.** *Cont.*

**Ref**

**Region Purpose Models Evaluation Metrics ID**

[49] England, U.K Estimating Traffic flow and based on that predicting EV

arrival rates

[50] Austin, Texas Classifying Charging profiles of EV users

CNN MAPE: 3.21 %

DGM Accuracy: 0.98 F1- Score:0.8

[51] UCLA Campus, Santa Monica,

California

Predicting charging behavior based on the labels obtained from clustering

ANN Accuracy

[52] Beijing, China Classifying whether the user will use fast charging or not

[53] Netherland Predict the time to the next plug for residential charging.

Binary Log regression

SVR with radial basis K-means

Accuracy: 0.894

MAE: 0.124 RMSE:0.158

[54] Dutch

metropolitan area

Creating EV Profiles that capture charging behavior

GMM ARI:0.6

[55] UCLA Campus, California

Predicting session duration and energy consumption in both residential and non-residential areas

Ensemble Model with SVR, RF, DKDE

SMAPE: 10.4

Reduced peak load by 27%, charging cost by 4%

**2. Data and Project Management Plan**

**2.1 Data Management Plan**

The data management plan is designed to articulate the strategies for data collection, methods of management, storage procedures, and utilization mechanisms for the project. The

34

data will be collected from a variety of sources. Initially, vehicle registration data will be obtained, providing annual counts of registered vehicles for each zip code and county in California. Subsequently, real-time information on EV charging stations across each zip code in California will be collected. Additionally, data on transit bus stop locations in San Jose, school bus terminal locations in San Jose, and real-world operational data for heavy/medium-duty commercial fleet vehicles across different weight classes will be gathered. This comprehensive dataset aids in understanding the various factors influencing charging station usage.

To improve the quality of data and maintain consistency in information, the following data management process can be applied: The first step is data cleaning, which involves identifying and removing inaccuracies, inconsistencies, and errors in the data. It is an essential step to improve data quality and ensure that the data is reliable and accurate. The second step is data transformation, it helps convert the data from one format to another to make sure the data is in a standardized format. For example, converting all other data types to a structured format CSV. So that can be easily analyzed and interpreted. The last step is to combine data from multiple sources into a single, comprehensive view. In addition to direct data collection, the project explores data partnerships and collaborations with relevant stakeholders. Collaborative efforts with school districts, commercial fleet operators, and local authorities contribute to a more holistic understanding of the EV ecosystem. These collaborations may involve data requests, sharing agreements, or even joint initiatives for data collection. By fostering partnerships, the project aims to enrich its datasets and ensure a comprehensive representation of the diverse facets of the medium/heavy-duty EV landscape. The combination of these data collection approaches ensures that the project starts with a robust foundation of accurate, diverse, and timely datasets, setting the stage for insightful analyses and informed decision-making in the

35

realm of EV charging infrastructure. The entire Data management plan is summarized in below Table 15.

**Table 15.** Summary of Data Management Plan

**Aspect Description**

Data Collection Sources • Vehicle registration data for annual counts by zip code and county in California.

• Real-time EV charging station data for each zip code in California. -

Transit bus stop locations data in San Jose.

• School bus terminal locations data in San Jose.

• Real-world operational data for heavy/medium-duty commercial fleet

vehicles across different weight classes.

Data Management Process

• Data cleaning: Identify and remove inaccuracies, inconsistencies, and errors.

• Data transformation: Convert data to standardized formats (CSV). • Data integration: Combine data from multiple sources into a single, comprehensive view.

Data Storage Methods • Cloud-based storage on Amazon Web Services (AWS S3) for scalability and accessibility.

• Structured database for efficient organization.

• Version control (GitHub) for traceability.

Data Utilization Mechanisms

• Analytical models for forecasting vehicle count, range predictions, and energy demand.

• Optimization models for strategic placement of new charging stations. • Visualization tools(Tableau) for effective communication of insights.

***2.1.1 Data Collection Approaches***

For primary data sources, collaboration with essential stakeholders is paramount. Engaging with charging station operators, city authorities, and heavy/medium-duty vehicle owners will provide real-time insights into charging station usage, energy demand patterns, and vehicle counts. Secondary data sources will complement primary data, drawing from publicly

36

available datasets, government reports, and industry publications. Additionally, sensor networks will be deployed at existing charging stations to capture live data on usage behaviors and charging dynamics.In the pursuit of a comprehensive Data-Driven Analysis of EV Charging Infrastructure for a Smart City for Medium/Heavy Duty Vehicles, our data collection strategy employs a multi-faceted approach. Primary data collection involves establishing robust partnerships with key stakeholders in the EV ecosystem. Through engagements with charging station operators, city authorities, and owners of medium/heavy-duty electric vehicles, we aim to extract real-time insights into the dynamic landscape of charging station usage, energy demand fluctuations, and the daily vehicular counts. This primary data, rich in its immediacy and specificity, forms the backbone of our analysis.

In addition to primary sources, secondary data collection is integral to our strategy. This involves tapping into publicly available datasets, government reports, and industry publications to augment and validate our primary data. By leveraging these authoritative secondary sources, we ensure a holistic and well-rounded understanding of the broader context in which our project

operates. Furthermore, the deployment of sensor networks at existing charging stations adds granularity to our dataset, capturing live data and providing detailed insights into usage behaviors and charging dynamics. This combination of primary, secondary, and sensor-derived data enhances the depth and accuracy of our dataset, forming the basis for informed decision making.

***2.1.2 Data Technology Selection***

The information is securely housed in an Amazon Web Services (AWS) cloud-based environment, ensuring both scalability and adaptability in data storage. Accessibility to the data is facilitated from any location with an internet connection. In terms of utilizing the data on

37

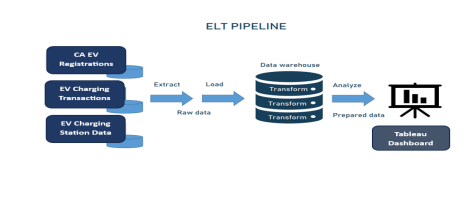
Amazon Web Services (AWS), there are various services available. Dataflow serves the purpose of collecting data from diverse sources, while Cloud Data Prep aids in data preparation. Additionally, AWS SageMaker is employed for real-time data processing, enabling the continuous monitoring of usage patterns and immediate issue detection. These functionalities make use of machine learning and deep learning models to extract insights into usage patterns and enhance capacity planning.

To examine the EV charging infrastructure, Python will serve as the primary programming language, leveraging several data science libraries like Pandas, Numpy, Scikit learn, and TensorFlow. TensorFlow, specifically, will be applied for the implementation of deep learning models, including temporal fusion transformers. Additionally, time series modeling techniques will be employed to scrutinize the seasonality and trends in electric vehicles, aiding in the prediction of demand.

***2.1.3 Data Engineering Process***

The data engineering process for the project encompasses multiple stages. Initially, we identified pertinent data sources crucial for our analysis, including EV charging station data, EV transaction data, transit bus data, school bus terminal data, and fleet transaction data. Subsequently, we procured and extracted the necessary data from diverse sources such as the CA open data portal, the CA Department of Education, the National Renewable Energy Laboratory, and others. The next steps involved data cleaning and preprocessing to ensure the data's quality and suitability for analysis. After collecting the data, we performed exploratory data analysis to gain insights and identify any issues or anomalies in the data. We also performed data transformation, data aggregation, and feature engineering to create new variables that would be useful for our analysis. Once the data was cleaned and transformed, we stored it in a suitable data

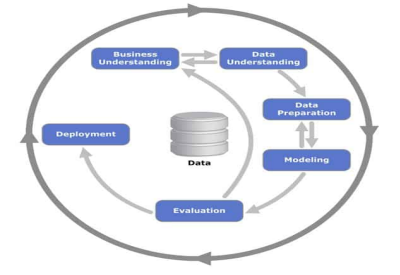
38

storage system over the cloud. Finally, we developed data pipelines and automated processes to update the data and perform regular analysis, ensuring that the insights and recommendations from the project are based on up-to-date and reliable data. Overall, this data engineering process ensures that the project is based on a solid foundation of relevant, accurate, and well-organized data. The entire data engineering process which is the ELT pipeline is shown in Figure 3. **Figure 3.** Data Engineering Process - ELT Pipeline

**2.2 Project Development Methodology**

A CRISP-DM methodology is used for this project, expanding into a CRISP approach Industry Standard Data Mining Process. This approach provides explanations of the duties involved in each project stage, descriptions of typical stages, and the connections between these activities. The CRISP-DM process model provides an overview of the data mining life cycle. The life cycle model contains six phases, and arrows indicate their most important and frequent dependencies. The phases can occur in any order. Projects frequently move back and forth between stages as needed. The model is highly flexible and easily adaptable. To put it another way, CRISP-DM aids in developing a data mining model that satisfies particular needs.

39

The CRISP-DM methodology includes the following phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. The following subsections go into great detail about each of these phases as they relate to this project. The different phases of CRISP-DM Methodology are shown in Figure 4 below. **Figure 4.** CRISP-DM Methodology Life Cycle

***2.2.1 Business Understanding***

The initial and most significant phase in strategic planning is to have a thorough grasp of the business environment. Here, the emphasis is on defining precise requirements and work objectives for the data-driven examination of the infrastructure supporting electric vehicle (EV) charging, especially for medium- and heavy-duty automobiles. The development of a reliable and widely accessible charging infrastructure is imperative given the growing demand for electric vehicles (EVs), particularly those in the medium- and heavy-duty segment. The problem is where to put charging stations that can effectively provide the electricity needed to recharge

40

EVs. Public spaces, shopping centers, and tourist spots are prime places to put them. More EV charging stations that are specifically designed to meet the charging requirements of medium and heavy-duty vehicles must be planned, designed, and built to promote the wider adoption of electric vehicles.

***2.2.2 Data Understanding***

Understanding the nature and characteristics of the data pertaining to the infrastructure for charging electric vehicles (EVs) in medium- and heavy-duty vehicles is the primary objective at the data understanding stage. Acquiring knowledge about the distribution, structure, and essential characteristics of the dataset is necessary for this. It is crucial to comprehend the underlying trends, patterns, and possible difficulties in the data. Significant factors are analyzed, including utilization trends, charging capacities, geographic distribution of charging stations, and any external influences on the data. To gain useful insights for optimizing EV charging infrastructure, appropriate issues, hypotheses, and modeling methodologies must be formulated. This fundamental understanding sets the basis for succeeding steps in data-driven research. ***2.2.3 Data Preparation***

The main objective of the Data Preparation stage is to prepare the data for further modeling; the Google Cloud Platform is used to store the raw data. In this step, careful data wrangling for EV charging station data is done, and any missing, erroneous, or duplicate values are checked. Moreover, any unnecessary or redundant data that is not relevant to the research is removed. The process of data transformation involves altering the data types, standardizing the formats, and sometimes adding computed columns once the data wrangling is finished. Subsequently, the data is partitioned into training and testing sets. The preprocessed data is

41

saved on Google Cloud, serving as the input for the subsequent models in the analysis of EV charging infrastructure for Medium/Heavy-duty vehicles.

***2.2.4 Modeling***

The modeling method in this analysis employs a diverse set of models to uncover patterns and insights within the data. More specifically, the temporal fusion transformer model, PuLP Linear Programming Optimization with K-means Clustering, and Stacking Ensemble Regressor (Weighted Fusion) ensemble methods, used alongside the time-series forecasting model Prophet, have been chosen meticulously, are used to predict, and forecast the future count of Heavy Duty and Medium Duty Vehicles and the required EV charging stations to cater the demand. Applying the Temporal Fusion Transformer Model's capabilities, enables adaptable forecasts of energy consumption on a daily, weekly, and monthly basis, considering the fluctuating demands for EV charging. The Stacking Ensemble Regressor integrates multiple algorithms and gives each one a weight to improve prediction accuracy and forecast the range of heavy and medium-duty vehicles of each weight class using characteristics like charging duration, and state of charge to the battery. Furthermore, the strategic placement of charging stations within the road network is optimized by the integration of PuLP Linear Programming Optimization and K-means Clustering. Each model utilizes historical charging data, geographical information, and other pertinent factors to provide a comprehensive and accurate assessment of the charging infrastructure requirements.

***2.2.5 Evaluation***

Through this phase, we will assess our built models by passing them through a test data set. First, model performance checks with unknown data. Then, to analyze the model performance of the algorithms, we have chosen these cluster performance indicators, such as

42

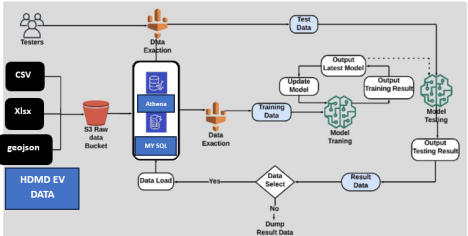
Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score. We also evaluate how well the optimization strategy works to deploy new charging stations in optimal locations for all kinds of vehicles. The Objective Function Value (Cost), Coverage Metric, Utilization Metric, and Equity Metric are the four metrics applied to our optimization problems. We analyze these measures in a variety of data models to determine which model has the best-predicted accuracy and best fit. The most reliable data-driven model is chosen for deployment based on this careful evaluation, which ensures that it is prepared for practical significance in the context of EV charging infrastructure for medium- and heavy-duty vehicles.

***2.2.6 Deployment***

Finally, in the Deployment phase, we continuously monitor and maintain models in order in order to ensure predictable and consistent performance after deployment. This includes meticulously tracking the models' behavior, quickly identifying any possible problems, and, if necessary, retraining the models to improve accuracy. A website is also created to make it easier to install EV charging stations in areas where they are most required. This forward-thinking strategy makes sure that the website contributes to a sustainable and effective transportation ecosystem by anticipating and adapting to the changing environment of electric mobility in addition to meeting current infrastructure needs. The project's dedication to accuracy, flexibility, and creativity in influencing the direction of EV infrastructure planning is demonstrated by the website. The culmination of this phase involves the compilation of all requisite data for the project report and presentation, summarizing the comprehensive data-driven analysis of EV charging infrastructure for Medium/Heavy-duty Vehicles.

43

***2.2.7 Intelligent System Engineering Process***

To summarize and create a well-defined intelligent system for our project, the first step in the process was to define the problem statement and gather the necessary data from various sources. The data was then cleaned and preprocessed using various techniques such as text normalization and feature extraction. Next, we employed various machine learning algorithms such as Prophet model, Stacking Regressor model, Temporal Fusion Transformer and Linear Programming and Clustering model to perform different tasks such as demand forecasting, range prediction, energy demand forecasting and finding optimal locations of new charging stations. We also utilized various time series modeling techniques to analyze the seasonality and trends of electric vehicle demand. The entire flow of the project is shown as below Figure 5. **Figure 5.** Project Data Flow Methodology

The performance of these algorithms was evaluated using various metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), and R2 Score. Based on the performance, we selected the best-performing algorithms for each task.

44

Finally, we deployed the selected algorithms in an integrated system that could take input data from different sources and provide output predictions in real-time. The system was tested for various use cases and the results were validated with real-world data. Throughout the process, we followed best practices such as using open-source tools, documenting the code and processes, and collaborating with domain experts to ensure the accuracy and usefulness of the insights generated by the system.

**2.3 Project Organization Plan**

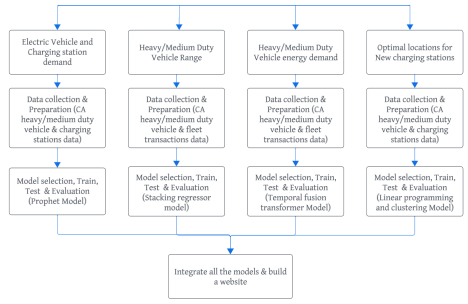
We are using the CRISP-DM technique to define the work breakdown structure for our project. A WBS tool is used to manage the objectives of a project by breaking it down into multiple tasks, and team members can be assigned different tasks to finish the tasks on time. As depicted in Figure 6, our project has been divided into several phases.

The Business Understanding phase is the first and most important step in the process. During this phase, we have gathered background information for our data-driven research on the deployment of EV infrastructure in urban areas. Our research into ML/DL efforts for this issue led us to establish the project's objectives. Here, our project aims to provide a comprehensive overview of how models such as Prophet, Stacking Ensemble Regressor (Weighted Fusion), the temporal fusion transformer model, PuLP Linear Programming Optimization with K-means Clustering can be used to analyze and forecast charging behavior and demand to enable the necessary infrastructure even in remote areas and communities that are too frequently ignored.

In the second phase, known as Data Understanding, we looked into the pertinent EV charging station data sources that were needed for the project. In this stage, we have selected California’s EV charging station usage data in accordance with the requirements for determining whether it is adequate for training and testing. We have also performed a quality analysis on the

45

dataset and developed a data management strategy to store our image data in the Google Cloud Platform (GCP).

**Figure 6.** Flow chart of project phases

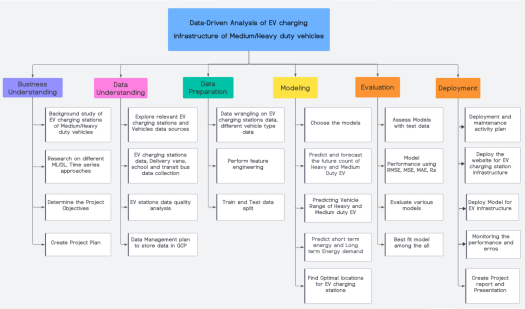
The third stage is Data Preparation, which involves processing and cleaning the data. Performing feature engineering enhances the performance of the ensemble learning model by selecting the appropriate features for the model and preparing the attributes in a way that is suitable for the model and helpful for the following phases. The data is divided for the models' training and testing and is then used in the following stages.

The Modeling stage comprises selecting an ensemble model and training that model with our preprocessed EV charging stations data set. specifically utilized and enhanced by the time series forecasting model Prophet, the Temporal Fusion Transformer model, PuLP Linear Programming Optimization with K-means Clustering, and Stacking Ensemble Regressor

46

(Weighted Fusion) ensemble approaches are utilized. These models are used to forecast and predict the number of Heavy Duty and Medium Duty Vehicles in the future, as well as the number of EV charging stations needed to meet demand. By utilizing the Temporal Fusion Transformer Model's features, it is possible to generate flexible energy consumption estimates on a daily, weekly, and monthly basis, which can be adjusted to meet the changing needs for EV charging. And using PuLP Linear Programming Optimization with K-means Clustering, can find the optimal locations for the charging stations.

In the last stage of deployment, we will deploy the ensemble model and create a website that can be useful for installing EV charging stations wherever they are required and doing the necessary maintenance and monitoring tasks. The last stage will be to compile a project report and presentation with all the required information.

**Figure 7.** WBS Chart

47

**2.4 Project Resource Requirements and Plan**

In this project, it is essential to find a suitable dataset of electric charging stations and to select appropriate software and hardware resources. The project's budget must be planned with an understanding of the costs and resources involved as given in Table 16. This project uses Amazon Web Services (AWS). It is a free public dataset that we collected from the source of data that we used in our study. AWS environments support platform-as-a-service, infrastructure as-a-service, and other cloud computing execution models. Affordability, adaptability, and traceability are all advantages of these technologies. A consistent approach is required to manage the different services offered by AWS, which include a variety of hardware and software resources. The tool known as the "AWS CLI" is the Amazon Web Services command-line interface. This unified tool offers a command-line interface for users to interact with a range of AWS services. With the AWS CLI, users can effectively manage their AWS resources, configure services, and automate tasks through scripting.

The first thing we need in this project is a virtual machine. We require two or more GPU worker nodes, 64 GB of RAM, and 16 virtual CPUs to run machine learning models uninterrupted. A bucket and folder will organize all project-related data on Aamazon S3. The AWS is a centralized repository for storing files so that you can access them quickly. Because we store data at each stage, such as raw, processed, trained, tested, and archived, we need 30 GB of storage.

Various data cleaning and wrangling procedures are used to prepare the gathered data for analysis. The free Google Colab software will be used for data preparation and modeling. For data wrangling and data transformation, we will use Python programming language. We can code and develop the model quickly using all Python 3.9 packages and modules. In order to

48

analyze and visualize data effectively, Pandas, NumPy, and Matplotlib libraries will be used. PIL (Python Imaging Library) and Torchvision libraries are given in Table 17. The platform facilitates the development, testing, and deployment of machine-learning models. **Table 16.** Hardware Requirements

**Hardware**

GCS- Standard Dual Region (us-central1 (Iowa) and us west1 (Oregon))

GCS-Standard (us-east4 (Northern Virginia))

**Configuration Purpose**

20 GB To store the collected data in each stage like raw,

processed, train and test data.

10 GB To archive the data in a separate region for disaster

recovery.

GC Virtual machine OS: Red Hat Enterprise Linux n1-standard-16

(vCPUs: 16, RAM: 60GB),

NVIDIA TESLA T4 GPU:2

**Table 17.** Software Requirements

To run machine learning models uninterrupted for the high-resolution image data.

**Software Version Purpose**

Python Programming Language

3.9.0 Performing exploratory data analysis

Google Colab Python Libraries: Pandas version 1.5.0

Matplotlib version 3.60

Touchvision 0.13.1

PIL 8.32

AWS Sage Maker NumPy version 1.21.0 Scikit-learn version 1.1.2

SciPy version 1.9.2

Performing data wrangling and data transformations

Developing, testing, and deploying unsupervised clustering machine learning models.

49

In accordance with CRISP-DM methodology, we prepared the Workbench structure, Pert, and Gantt charts using the Click Up tool given in Table 18. To collaborate online, we used Zoom, a cloud-based, free video conference application. To redistribute our preprocessed data and metadata documentation, we use GitHub, an open-source platform. As part of our project, we also use the MS Office 365 suite under a student license to prepare reports and presentations. **Table 18.** Tools and Licenses

**Tool License Purpose**

ClickUp Unlimited (Paid)

For our project management, we need to create a WBS, a Gantt chart, and a Pert chart using CRISP-DM.

Zoom Free Project planning and discussion can be done online.

Github Free Data acquisition, usage, and distribution documentation and preprocessed data will be redistributed.

MS Office 365 suite

Student As part of our project, we will create reports and presentations

Based on the configurations and the duration of the various tools and resources we use. Table 19 provides an overview of the budget projections for these tools and resources. In total, it is estimated that the calculation services will cost approximately $2875 in total. A Google Cloud Pricing Calculator is a tool that estimates costs for all of the available GCP services. **Table 19.** Project Cost Estimation and Justification

**Purpose Resource Type Resource Time Duration Cost in USD**

Cloud Service Management Tool

Software GCloud CommandLine

02/17/2023 – 11/30/2023 ~ 12 months

Free

50

**Table 19.** *Cont.*

**Purpose Resource Type Resource Time Duration Cost in USD**

Data Storage Hardware AWS S3 02/17/2023 – 12/17/2023

~ 10.5 months

$15.20

($ 3.80 per 1

month for 50 GiB)

Virtual Machine Hardware GCVM n1- standard- 16

(vCPUs: 16,

RAM: 60GB),

NVIDIA

TESLA T4

GPU:2

Data Preprocessing Software Google Colab Python 3.7

Version

02/20/2023 – 12/15/2023

~ 10 months

02/19/2023 – 12/15/2023

~ 10 months

$2400($2.36 per

GPU/hour, we will shut down the virtual

machine as much as possible. There may be a need for 2 GPUs)

Free

Model

Development

Project

Management

Software AWS Sage Maker

Tool ClickUp version 3.0

02/22/2023 – 12/15/2023

~ 10 months

03/18/2023 – 12/03/2023

~ 9 months

$70.75

$36

($7.2 per each member)

Data Redistribution Tool Github version 3.6.3

03/29/2023 – 11/29/2023

~ 9 months

Free

Project Work Collaboration

Project

Documentation

Tool Zoom version 5.1.2

Tool Ms Office 365 Suite version

2209

01/29/2023 – 12/15/2023

~ 2.5 months

02/15/2023 – 12/15/2023

~ 10 months

Free

Free (Student License)

51

**2.5 Project Schedule**

The most well-known tool for organizing project goals or tasks is the Gantt chart which helps to organize the project by establishing completion dates for the activities and allocating resources who will be in charge of those tasks. We made a chart with the task name, start date, deadline date, estimated completion time, and team member responsible for the assignment. The start and end dates, as well as their durations, are important since they show how long the activity will last. We have not included weekends, Thanksgiving break, or any other holidays that fell on our timetable when allocating these deadlines. We have coordinated the deadlines with the given deadlines for the deliverables. Also, there are still many tasks that need to be completed. Figure 8 is shown in the Gantt chart.

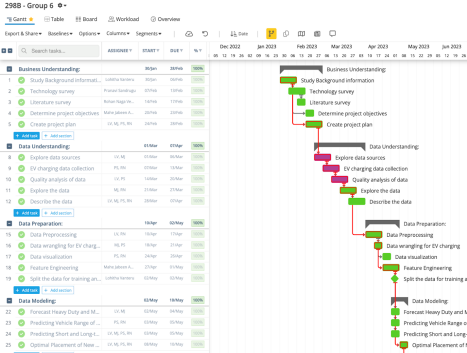
One of the most often used project management tools is a PERT chart shown in Figure 9, which displays the tasks, duration, and assignees inside rectangular shapes. PERT chart is used to organize and determine the approximate duration of a project's tasks. Tasks on the critical path in project management are essential to finishing a project. For the PERT chart, we used a task oriented methodology. Here, the dependencies between the tasks are defined, and assigned to the team members, and the time required to complete each task has been computed. The red arrows indicate the critical path; each task has a task number.

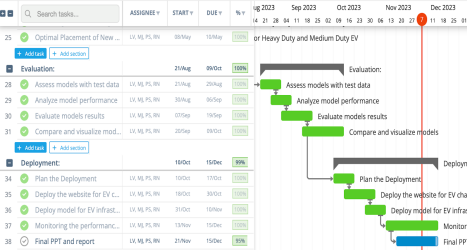
**3 Data Engineering**

**3.1 Data Process**

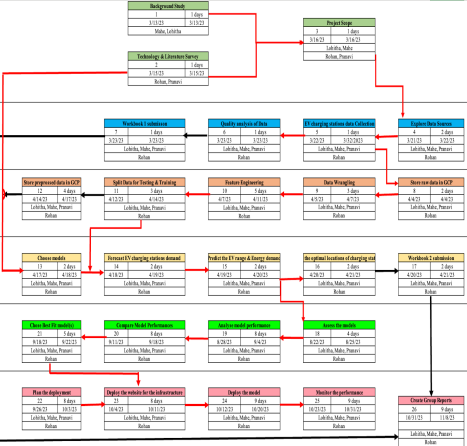
The raw form of data lacks utility, and data processing is the essential procedure of taking this raw data, refining it, and converting it into usable information. Commencing with the collection of raw data, the process involves enhancements, organization, processing, evaluation, archival, and ultimately presenting the data in a more structured manner.

52



**Figure 8.** Gantt Chart

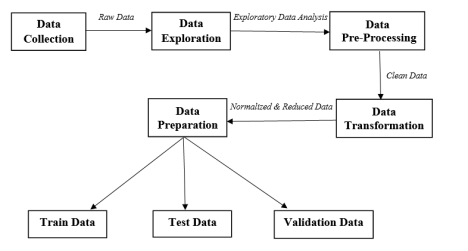
53

**Figure 9.** Pert Chart

This transformation provides the necessary shape and context for machines to analyze and derive meaningful insights, uncovering patterns or trends that might be challenging to discern otherwise. It holds significance as a crucial stage that facilitates informed decision making and underscores the informational value available in the process. In broad terms, data processing for this project encompasses data gathering, data exploration, data cleaning, data

54

transformation, and data preparation. The entire data process is summarized as a flow chart in the following Figure 10.

**Figure 10.** Data Process Flow Chart

The initial step in the data processing cycle is the collection of raw data, and the accuracy and relevance of the final results are heavily dependent on the quality of this unprocessed information. It is crucial to source preliminary data from trustworthy and dependable sources to ensure the validity and significance of the outcomes. In our project, the assessment of potential locations for new charging stations relies solely on real-world data within the Electric Vehicle Charging Infrastructure framework. Although acquiring data in this domain can pose challenges, real-world data provides insights into the actual behaviors of EV owners and potential EV users, making it more realistic than simulation-based results. The process of addressing machine learning problems commences with data acquisition, followed by exploration to assess its quality and accuracy, and comprehension through the interpretation of various visualizations. Post cleaning, the data may transform to ensure suitability for analysis, involving tasks such as

55

converting categorical variables to numerical ones, scaling or normalizing data, and employing feature engineering techniques to create new variables. After transformation, the data is segmented into training, validation, and test sets. The training set is employed for model training, the validation set aids in tuning hyperparameters, and the test set is instrumental in evaluating model performance. The adequacy of training data is crucial for model consistency, while an insufficient amount of testing or validation data can lead to increased variability in performance metrics. Adhering to these principles, the modified data is prepared for modeling, with an 80% allocation to the Train set and the remaining 20% divided into 10% for Test and 10% for Validation sets, used for model evaluation. The final dataset is then employed for both descriptive and predictive analyses after completing the data pre-processing stage. The established methodology can propose optimal locations for new charging stations by leveraging real-world data, particularly focusing on minimizing the average drop in charging usage within a charging zone. This essentially translates to maximizing the overall aggregate usage within a charging zone once the new charger is operational. Beyond pinpointing suitable locations for new charging stations, this methodology sheds light on the usage patterns of existing charging stations and the charging behaviors of EV owners. Consequently, providers of EV charging infrastructure can employ this proposed technique as a decision-support tool for expanding their charging network.

The California Open Data Portal is an online platform dedicated to fostering transparency and open government by providing public access to a diverse array of datasets related to the state of California. It covers various topics such as demographics, education, health, and transportation. The datasets encompass information on electric vehicle (EV) charging infrastructure, adoption rates of electric vehicles, emissions reductions, and environmental data.

56

The portal serves a broad audience, including researchers, policymakers, developers, and the general public, who can freely access, download, and analyze datasets in machine-readable formats, facilitating seamless integration and analysis for gaining insights into different facets of California's public life. The California Department of Motor Vehicles (DMV) is a government agency responsible for vehicle registration, driver's license issuance, and maintaining driving records to ensure safety and proper documentation in the state. It also contributes data on electric vehicles (EVs) through registration and licensing records, providing statistics on the quantity and types of electric vehicles registered. This information is valuable for tracking the expansion of the electric vehicle market in California, evaluating the popularity of various EV models, and understanding demographic trends in electric vehicle ownership across the state. The National Renewable Energy Laboratory (NREL) operates as a research institution under the U.S. Department of Energy with a primary focus on advancing renewable energy and energy efficiency technologies. It serves as a crucial source of research and data on various renewable energy technologies, including electric vehicles (EVs). NREL conducts comprehensive studies on EV technology, charging infrastructure, battery advancements, and the overall integration of electric vehicles into the broader energy framework. The laboratory offers valuable insights into the performance, efficiency, and environmental advantages of electric vehicles. Researchers and professionals in the industry frequently depend on NREL's data to comprehend the current state of EV technology and guide future advancements. The U.S. Alternative Fuel Data Center (AFDC) is a comprehensive platform managed by the U.S. Department of Energy (DOE), offering a wealth of information, datasets, and tools related to alternative fuels and advanced vehicles. Serving as a central repository, it covers various data sets related to alternative fuels, with a specific focus on electric vehicles (EVs). The datasets include valuable information such

57

as charging station data, fueling station data, and vehicle data. Researchers and policymakers frequently leverage the AFDC's data to analyze patterns, strategize infrastructure development, and make informed decisions concerning the promotion and adoption of alternative fuel vehicles, particularly electric vehicles.

The data processing for forecasting the demand for electric vehicle charging stations and predicting the count of Electric Vehicles (EVs) for Medium and Heavy-duty vehicles involves obtaining data from the California Department of Motor Vehicles for vehicle details and the U.S. Alternate Fuel Data Center for charging station information across CA zip codes. This dataset includes details such as year, model, make, vehicle types, and zip codes within California, spanning the years 2010 to 2022, with a count of 35,742. The aim is to predict future EV counts, forecast charging station demand, and visualize seasonal patterns. Additionally, we gathered medium and heavy-duty vehicle fleet transaction data from the National Renewable Energy Laboratory, encompassing charging start and end times, starting, and ending State of Charge (SOC), and total energy delivered. This dataset covers the period from 2018 to 2022, with a count of 428,963. The objective here is to predict the range of both heavy and medium-duty vehicles and visualize daily, weekly, and monthly energy demand forecasts. To enhance our understanding of public transportation, we also collected transit bus stop data from the CA Open Data portal and school bus terminal data from the CA Department of Education. The summary of open-source datasets that are collected for this project is outlined in the following Table 20.

These diverse data sources collectively serve as the foundation for our data-driven analysis, enabling us to extract valuable insights into electric vehicle adoption and the deployment of charging stations. Following the data collection phase, we performed data cleaning, essential aggregation, and transformation. Subsequently, we prepared the datasets,

58

divided them into an 80:10:10 ratio, and proceeded to the modeling stage. We applied the model fitting process to generate accurate predictions, and in addition, we conducted fine-tuning procedures to enhance the efficiency of the models.

**Table 20.** Open-Source Datasets

**Dataset Description**

CA Vehicle Registration Data Number of vehicles registered each year for every zip code and county in CA.

CA EV charging Stations Data Real-time data of EV Charging stations for each zip code in CA.

Transit Bus Stops Data Data of public bus stations locations for transit bus in San Jose. School Bus Terminal Data Data of school bus terminal stations locations in San Jose.

Heavy/Medium Fleet Vehicle Transactions

**3.2 Data Collection**

Real-world data of commercial fleet vehicle operating data for each weight class.

The process of collecting data follows a methodical approach to obtain pertinent information from various sources. The fundamental steps in this data collection process are elucidated below:

**1. Identifying Data Sources:** The initial phase involved identifying the primary data sources crucial for the analysis. These sources included the California Department of Motor Vehicles, supplying details on vehicle specifics; the U.S. Alternate Fuel Data Center, providing data on charging stations; the CA Open Data portal, serving as a source for transit bus stops; and the CA Department of Education, acting as a source for school bus terminal data. Additionally, the National Renewable Energy Laboratory offered data on Heavy/Medium fleet transactions.

59

**2. Retrieving Data from the California Department of Motor Vehicles:** Information about electric vehicles, encompassing details like the year, model, make, vehicle types, and zip code information within the state of California, was sourced from the California Department of Motor Vehicles. This dataset covered the period from 2010 to 2022. This dataset has 35, 742 records.

**3. Utilizing the U.S. Alternate Fuel Data Center:** Real-time data concerning charging stations in every zip code in California was acquired from the U.S. Alternate Fuel Data Center. This dataset encompassed essential information, including the year, cities, zip codes, latitude, and longitude, in addition to charger types such as EV level 1, level 2, and DC fast count throughout the state. This dataset spans from 2010-2022 and has 14,862 records.

**4. Exploring the CA Open Data Portal:** To gain a deeper understanding of public transportation, information on transit bus stops was sourced from the CA Open Data portal. This dataset played a crucial role in comprehensively grasping the dynamics of public transit. Both datasets cover the period of 2022. Transit bus stops data has 1289 records.

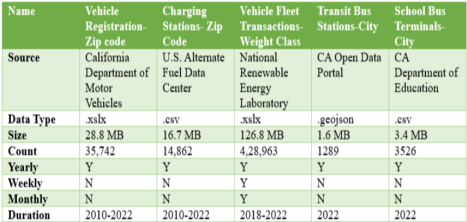
**5. Retrieving Data from the CA Department of Education:** For a more profound insight into public transportation, data on school bus terminals was obtained from the CA Department of Education. This dataset proved essential in gaining a comprehensive understanding of public transit dynamics. The dataset encompasses the timeframe of 2022 and has 3526 records.

**6. Deepening Commercial Fleet Analysis:** Lastly, to enhance our understanding of commercial fleet operations, we procured 'Heavy/Medium Vehicle Fleet Transactions

60

Data' from the National Renewable Energy Laboratory. This dataset furnishes real-world data about different weight classes of commercial fleet vehicles and was acquired from the National Renewable Energy Laboratory. The dataset spans from 2018-2022 and has 4,28,963 records.

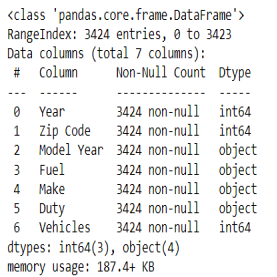
In conclusion, the process of collecting data entailed a careful identification of sources, obtaining access to pertinent datasets, and establishing a thorough groundwork for the ensuing data-driven analysis. The detailed information on datasets is summarized in below Figure 11.

**Figure 11.** Summarized information of the datasets

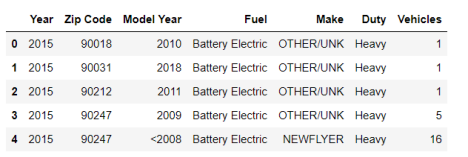
Raw dataset samples offer an initial look at the unprocessed and varied data that underpins the Data-Driven Analysis of EV Charging Infrastructure for Medium/Heavy Duty Vehicles project. These samples cover electric vehicle details, charging station information, transit bus stops, and fleet transactions, encompassing essential variables like a year, make, model, location, and charging specifics. These samples act as the foundational material for subsequent processes such as data processing, analysis, and modeling, providing a starting point

61

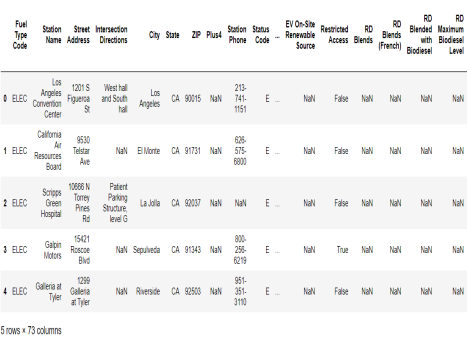
to derive meaningful insights into the intricate dynamics of electric vehicle adoption and charging infrastructure deployment. The figures below illustrate raw samples of the data collected.



**Figure 12.** Showing the raw sample data from vehicle registration data

**Figure 13.** Summary of the vehicle registration data

62

**Figure 14.** Showing the raw sample data from EV charging stations

**Figure 15.** Showing the raw sample data of vehicle fleet transactions

63

**Figure 16.** Summary of EV charging stations data

**Figure 17.** Showing sample raw data samples from transit bus stop and school bus terminals data

64

**Figure 18.** Summary of vehicle fleet transaction data

**Figure 19.** Summary of the transit bus stop and school bus terminals data

65

**3.3 Data Pre-processing**

In the data preprocessing stage, for all the four models namely, the time-series forecasting model Prophet, Temporal Fusion Transformer (TFT) model, PuLP Linear Programming Optimization with K-means Clustering, and Stacking Ensemble Regressor (Weighted Fusion), a meticulous and customized approach is used in order to guarantee the precision and dependability of the ensuing analyses. The initial step involves the storage of raw data in a robust platform, providing a secure and scalable environment for data handling. The selected approach offers a scalable framework that is suitable for managing large datasets typical of electric vehicle charging infrastructure assessments, in addition to guaranteeing the confidentiality and dependability of the data.

In the next step, in-depth data wrangling techniques are applied to the data to handle a variety of problems, such as resolving missing, erroneous, or duplicate values. Each model has specific preprocessing processes that are meticulously built to meet its specific needs and characteristics. The customized preprocessing processes created for every model are one of the distinguishing features of this technique. Considering the distinct needs and attributes of every model, particular modifications and improvements are implemented.

***3.3.1 Forecast Heavy Duty and Medium Duty EV and Charging Stations Demand*** In this step, for the heavy and medium-duty EV charging dataset, the data preprocessing is conducted to enhance its consistency and effectiveness. This preparatory step is undertaken to ensure that the dataset is optimized for subsequent analyses, fostering greater usability and coherency.

• **Define Column Data Types:**

66

To ensure correct data interpretation, the initial step is to define specific data types for columns using a ‘column\_data\_types’ dictionary. This is useful for ensuring that the columns are interpreted correctly.

• **Clean Column Names:**

The function, ‘clean\_column\_names’ is defined to is to replace underscores for spaces in column names. By doing this, a standardized naming system is established, which will facilitate data manipulation. This is done to create a consistent naming convention, making it easier to work with the data.

• **Convert Year to Datetime:**

The 'Year' column in the data frame is transformed using ‘pd.to\_datetime’ to datetime format. This transformation is likely applied to facilitate temporal analysis or plotting. • **Removing irrelevant columns:**

The 'Model\_Year' column is removed from the data frame, this is due to the non relevance of this intended analysis. This sequence of preprocessing steps aims to refine the dataset, making it more amenable to subsequent analyses and modeling tasks in the realm of electric vehicle data exploration.

***3.3.2 Predicting Vehicle Range of Heavy-Duty and Medium-Duty EV***

Below are the data preprocessing steps involved in the predicting Vehicle Range of heavy-duty and medium-duty electric vehicles.

• **Datetime conversion:**

Date and time stamped columns such as Local Disconnect Time, Local Connect Time, Local Charge Start Time, Local Charge End Time, Date, should be converted to datetime format for more straightforward manipulation and examination.

67

• Some of the categorical columns like ‘Vehicle ID’, ‘Charger ID’ are converted to numerical columns using the label encoder.

The sample data is shown in Figure 20 after completing all the data preprocessing steps. **Figure 20.** Sample data after the data preprocessing

***3.3.3 Predicting Short and Long-term Energy Demand of Heavy-Duty and Medium-Duty EV*** The below are the data preprocessing steps involved in the predicting short- and long term energy demand of heavy-duty and medium-duty electric vehicles.

• **Clean Column Names:**

68

clean\_column\_names' function is created to in order to replace column names with underscores instead of spaces. The 'charging\_data' dataframe is needed to be used with the function to ensure uniform column naming.

• **Remove the irrelevant columns:**

Local\_Connect\_Time" and "Local\_Disconnect\_Time" columns are removed from the dataframe.

• **Renaming the column names:**

Some of the column names are changed for the better understanding of the data such as, 'Local\_Charge\_E/Time' is changed to 'Local\_Charge\_End\_Time'.

• **Convert to Datetime:**

The columns 'Local\_Charge\_Start\_Time' and 'Local\_Charge\_End\_Time' are converted to datetime format.

The sample data snippet is shown in Figure 21 after completing all the data preprocessing steps. **Figure 21.** Sample data after the data preprocessing

69

***3.3.4 Optimal Placement of New Charging Stations for Heavy Duty and Medium Duty EV*** The below are the data preprocessing steps for the Optimal Placement of New Charging Stations for Heavy Duty and Medium Duty EV.

• **Change column names:**

The column names having spaces are replaced with underscores using replace function in Python, is done for the better understanding of the data.

• Filtering based on the city column where we have filtered the dataframe to include only rows where the 'City' column is equal to 'San Jose’.

The sample data snippet is shown in Figure 22 after completing all the data preprocessing steps. **Figure 22.** Sample data after the data preprocessing

**3.4 Data Transformation**

One of the most important stages of data analysis is data transformation, involving the systematic conversion of raw data into a more structured and suitable format for analysis and modeling. This multifaceted process encompasses a range of operations aimed at enhancing data

70

compatibility with specific analytical tools or models. Common data transformation tasks include Standardizing data formats, correcting outliers, and handling missing, erroneous, or duplicate values. In addition, transformations could entail disaggregating or aggregating data, feature engineering new variables, or changing data types in accordance to analytical specifications. Getting the data ready for further analysis in a way that best facilitates insightful conclusions, precise forecasts, and robust decision-making in subsequent stages of analysis. ***3.4.1 Forecast Heavy Duty and Medium Duty EV and Charging Stations Demand***

In time series analysis, grouping and aggregating temporal data is a common transformation as seen in the forecasting Heavy Duty and Medium Duty EV and Charging Stations Demand.

• **Grouping and Aggregation:**

The 'Year' column in the original dataframe (df) is used to group the data, and groupby and sum are then used to determine the total number of 'Vehicles' for each year. • Once the data has been converted, the function 'fit\_and\_forecast' is used to set the growth, seasonality mode, and changepoint scales, among other parameters for the modelling part.

• Extraction and printing of the anticipated EV car counts for each subsequent year until 2045 constitute the last phase, showcasing the effective use of data transformation. • **Merging DataFrames:**

Using the 'Zip\_Code' column, the 'df' and 'zc' dataframes are combined as part of the data transformation. Next, the 'zip' column is removed and the column order is changed in the resulting'vehicle\_data' dataframe. The desired column order is specified in the 'new\_order' list, and the'vehicle\_data' dataframe is reorganized in accordance with that

71

order. The modified dataframe now has columns labeled 'Year,' 'Fuel,' 'Make,' 'Duty,' 'Vehicles,' 'Zip\_Code,' 'primary\_city,''state,' and 'county' in the following sequence. The sample data snippet is shown in Figure 23 after merging the dataset.

**Figure 23.** Sample data after the data merge

***3.4.2 Predicting Vehicle Range of Heavy-Duty and Medium-Duty EV***

• **Dropping Null/Missing values:**

Dropping the null or missing data values present in the dataset.

• **Creating new columns:**

A new column named ‘Range’ created which later uses in predicting the vehicle range of heavy and medium duty EV. The column 'Range' is created by applying a formula involving 'Delta\_SOC' to estimate the range.

• **Feature Selection:**

Features that are relevant to the prediction task are selected. "Charging\_Duration," "Average\_Power," and "Delta\_SOC" are a few of these.

72

**Figure 24.** Sample data after performing the data transformation.

***3.4.3 Predicting Short and Long-term Energy Demand of Heavy-Duty and Medium-Duty EV*** • **Handling missing values:**

Removed rows from the Data Frame charging\_data that contain missing (null) values. SimpleImputer is used to handle missing values in the resampled data. Missing values are filled with the mean value of the respective column using the fit\_transform and transform methods for daily, weekly, and monthly forecasts, respectively.

• **Creating new columns:**

The new variables such as 'Charging\_Duration,' which is created by finding the time difference in hours between 'Local\_Charge\_Start\_Time' and 'Local\_Charge\_End\_Time'. And, Average\_Power,' and 'Delta\_SOC.' are created and added to the existing data to better represent and summarize aspects of the data.

• **Feature Engineering:**

Feature engineering done selecting the relevant columns which are created such as 'Delta\_SOC' and 'Charging\_Duration', feature selection, and defining the target variable ('Range’).

73

The sample data snippets are shown in Figure 25 before and after handling the null values. **Figure 25.** Sample data before and after handling the null values.

The sample data snippet is shown in Figure 26 after performing the data transformation techniques.

**Figure 26.** Sample data after performing the data transformation.