**Design and Methodology for Prediction of Demand for Public Transportation Services Using Data Analytics.**

**Abstract:**

In this paper we present a design and methodology for conducting experiments on different data analytics techniques to predict the demand for public transportation services. The experiment utilizes three datasets; demand\_data, demographic\_data and weather\_data. Our proposed methodology includes steps such as data preparation, preprocessing, outlier detection feature standardization, encoding of categorical variables and exploration of multiple machine learning (ML) and time series analysis methods. Our ultimate goal is to identify the accurate predictive model that can improve the efficiency and reliability of public transportation services.

**Introduction:**

Efficient and functioning public transportation is essential for urban planning and managing the movement of people as it addresses issues like traffic congestion, environmental concerns and overall livability in cities. Accurately predicting transportation demand plays a role in optimizing resource allocation and ensuring passenger satisfaction. This prediction is invaluable because it allows transit authorities to allocate resources effectively providing services during peak hours while implementing cost saving measures during off peak periods. Moreover precise demand forecasts result in waiting times, reduced overcrowding, improved service quality and ultimately higher passenger satisfaction levels. Consequently this encourages people to utilize public transportation services. To develop models for public transportation demand cities can leverage data analytics techniques such as machine learning and statistical modeling. These models rely on data encompassing information, about passenger counts routes taken, weather conditions and more. By incorporating these models into real world transportation management systems cities can make decisions based on data analysis that contribute to the long term sustainability of their transportation networks while consistently improving services to meet evolving needs.

**Preparing the Data:**

Data Collection-In this step we gather the datasets. Demand\_data, demographic\_data and weather\_data. These datasets contain information that's relevant to the problem or task at hand.  
Data Cleaning; It is important to clean the data to ensure its quality and reliability. This involves;  
Handling Missing Values; Identifying and addressing any data points. This can involve replacing missing values with estimated ones or removing rows/columns with many missing values.  
Removing Duplicates; Identifying and eliminating records to avoid repetition and ensure consistency in the data.

**Data Cleaning:**

It is important to clean the data to ensure its quality and reliability. This involves;

Handling Missing Values; Identifying and addressing any data points. This can involve replacing missing values with estimated ones or removing rows/columns with many missing values.

Removing Duplicates; Identifying and eliminating records to avoid repetition and ensure consistency in the data.

Ensuring Data Consistency; Making sure that the data is consistently formatted, such as maintaining uniform date formats, units of measurement and categorical values.

**Inclusive Preprocessing Approach;**

Exploring Relationships; Understanding the connections between the three datasets (demand\_data, demographic\_data and weather\_data) is crucial. Exploratory data analysis techniques can be used to identify patterns, correlations and dependencies, between variables.

Linking Datasets; To improve modeling accuracy we link the demographic\_data and weather\_data to the demand\_data based on common attributes or keys. This linking provides context and features that enhance our predictive model.

**Outlier. Treatment;**

Detecting Outliers;Outliers are data points that deviate significantly from the behavior of the dataset. To identify these outliers you can use techniques like Z score or interquartile range (IQR). The Z score measures how far a data point is from the mean in terms of deviations while IQR defines the range where most data points fall.

Dealing with Outliers; Once you detect outliers you have options for handling them depending on their impact. You can correct values remove outliers altogether or transform them to fit within a specific range. The choice will depend on your datas nature and the analysis or modeling you're conducting.

**Standardizing Numerical Features;**

Why Standardize; Standardizing numerical features involves scaling them so that they have a value of 0 and a standard deviation of 1. This is crucial to ensure that all features are on a scale. It becomes particularly important for models that rely on distance based calculations like linear regression and k means clustering because it prevents features with scales from dominating the models behavior.

**Encoding Categorical Variables;**

Label Encoding; Categorical variables are typically non numeric. Need to be converted into a numerical format, for machine learning algorithms to effectively process them.

Label encoding is a technique that assigns a numerical value to each category within a categorical variable. As an illustration consider the "Color" variable with options, like "Red," "Blue," and "Green." In label encoding these categories could be represented by the integers 0 1 and 2 respectively.

**We explore machine learning methods for predicting transportation demand**

Machine learning encompasses a variety of techniques and there are four main methods that stand out. Linear Regression is a model that establishes a linear connection between input features and the target variable. It quantifies the impact of each feature on predictions. On the hand Decision Trees are versatile tools that can be used for both classification and regression tasks. They divide the feature space based on thresholds effectively capturing complex interactions between variables. Random Forest, which is a method combines multiple decision trees to improve accuracy and overcome overfitting issues. Lastly Neural Networks, with the help of the user Keras API utilize interconnected nodes in layers to uncover intricate nonlinear relationships within data. They excel in tasks such as image recognition and natural language processing. These techniques offer a range of approaches to address machine learning challenges. The choice, among them depends on problem requirements and dataset characteristics. In applications it's common to use a combination of these methods or more advanced models to optimize predictive performance.

**Investigating Time Series Analysis Methods:**

Understanding and predicting transportation demand heavily relies on time series analysis, which involves commonly used methods. One important method is the Autoregressive Integrated Moving Average (ARIMA) model, which utilizes differencing to achieve stationarity and captures relationships between values of a time series. ARIMA models are particularly effective in identifying short term trends and dependencies making them valuable for short term forecasting. Another approach is the Seasonal Decomposition of Time Series (STL) which breaks down time series data into components such as seasonality, trend and residuals. This decomposition provides an understanding of underlying patterns and variations aiding in the detection of recurring patterns and anomalies. Additionally Long Short Term Memory Networks (LSTMs) a type of neural network offer a flexible solution for modeling complex time series patterns. LSTMs excel at capturing both term and long term dependencies within the data making them especially suitable for analyzing intricate time series data characterized by extended temporal relationships like stock prices or natural language data. The choice of method ultimately depends on the nature of the data and analysis objectives since each approach offers unique advantages, for different aspects of time series analysis.

**Performance Evaluation and Model Selection:**

The performance of each model is assessed by using metrics like mean squared error (MSE) root mean squared error (RMSE) and R squared. The model that exhibits the amount of error and the highest R squared value is chosen as the most reliable predictor, for transportation demand.

**Summary:**

This paper presents an approach to predict the demand for public transportation services using data analytics techniques. The study involves steps such as data preparation, preprocessing, outlier detection, standardization, categorical encoding exploring different machine learning methods and investigating time series analysis techniques. By combining these steps our goal is to create a predictive model that can improve the efficiency and reliability of public transportation services. Ultimately this will contribute to urban mobility and passenger satisfaction.

In summary the methodology described in this paper offers a strategy to tackle the challenges of predicting transportation demand using data analytics. By incorporating techniques and exploring both machine learning and time series analysis approaches we can effectively use available data to develop accurate and dependable predictive models for public transportation services. Through refinement and optimization of these models they can play a crucial role in urban planning and resource allocation – ultimately resulting in enhanced transportation experiences, for the general public.