**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 Methodology:**

To ensure predictions we adopt a comprehensive preprocessing approach. This approach entails exploring and understanding the relationships between these three datasets. We link demographic\_data and weather\_data to demand\_data based on attributes. This linking provides context to enhance our predictive model.

**Outlier Detection and Treatment:**

Outliers can have an impact on the performance of predictive models. Therefore we employ outlier detection techniques like Z score and interquartile range to identify and address any outliers in the datasets. We. Correct or remove these outliers depending on their impact.

**Standardization of Numerical Features:**

Standardizing numerical features is essential to ensure that all features are on a scale. This prevents any bias towards features during modeling particularly for models that rely on distance based calculations, like linear regression.

**Encoding Categorical Variables:**

Categorical variables in the dataset are transformed using a technique called label encoding. This conversion helps to represent values in a numerical format enabling machine learning algorithms to process them effectively.

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

1. **Linear Regression** - Linear regression serves as a baseline model establishing a relationship between input features and the target variable. It quantifies the impact of each feature on the prediction.

2. **Decision Trees**- Decision trees divide the feature space based on threshold values making them suitable for capturing interactions between variables.

3. **Random Forest**- Random Forest is a technique that constructs multiple decision trees and combines their predictions leading to improved accuracy and reduced overfitting.

4. **Neural Networks (using Keras)-** Neural networks are composed of interconnected nodes organized in layers. To design and train neural network architectures we utilize Keras, which's a user friendly neural network API. By exploring networks we can capture complex nonlinear relationships, within the data.

**Investigating Time Series Analysis Methods:**

Given the nature of transportation demand it is essential to use methods for analyzing time series;

**Autoregressive Integrated Moving Average (ARIMA)-** ARIMA models are effective in capturing the relationship between past values of the target variable and achieving stationarity by applying differencing. These models work well for identifying short term trends.

**Seasonal Decomposition of Time Series (STL)-** STL breaks down time series data into components like seasonality, trend and residuals. This breakdown helps us understand the underlying patterns and variations.

**Long Short Term Memory Networks (LSTM)-** LSTM networks are a type of recurrent neural network (RNN) designed to handle data sequences. LSTMs excel at capturing long term dependencies. Are useful, for modeling intricate time series patterns.

**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.