Project Summary

# Overview

Automatidata is consulting with the New York City Taxi and Limousine Commission (TLC) to create a predictive regression model. The goal is to estimate taxi fares before rides, utilizing a large dataset of taxi trip data provided by TLC.

## Key Problem

The main problem is to accurately estimate taxi fares based on historical data, allowing drivers to have a better understanding of expected costs prior to starting a trip.

**Stakeholders**

* **Internal**:
  + Udo Bankole (Director of Data Analysis)
  + Deshawn Washington (Data Analysis Manager)
  + Luana Rodriguez (Senior Data Analyst)
  + Uli King (Senior Project Manager)
* **External**:
  + Juliana Soto (Finance and Administration Department Head, TLC)
  + Titus Nelson (Operations Manager, TLC)

## Objectives

1. Provide visuals about the descriptive analysis of the dataset.
2. **Build a regression model** capable of predicting taxi fares based on data features such as distance, duration, time of day, and location.
3. **Ensure model consistency** through rigorous testing and performance evaluation.
4. Present findings and recommendations to all stakeholders.

# Data Preparation and Exploration

## Data Overview

The dataset includes over a million trips per day from more than 200,000 licensed taxi and limousine drivers in NYC. Data points include ride start and end times, trip distance, pickup/dropoff locations, fare amounts, and potentially customer demographics or vehicle types.

**Tasks**

1. **Data Scrubbing**:
   * Cleaning the dataset by removing or addressing missing, null, and outlier values.
   * Ensuring all variables, especially those affecting fare (e.g., distance, time of day, traffic conditions), are properly formatted for analysis.
2. **Exploratory Data Analysis (EDA)**:
   * Explore the dataset to identify trends, correlations, and outliers.
   * Generate summary statistics for key features (distance, fare, ride duration).
   * Visualize data distributions to inform modeling decisions.
3. **Feature Engineering**:
   * Create new features that might improve model performance, such as “mean duration” or “rush hour” indicators.
   * Test relationships between fare and distance, traffic, and location.

# Model Building

**Tasks**

1. **Regression Model Development**:
   * Choose an appropriate regression algorithm (e.g., linear regression, random forest, or gradient boosting) based on the complexity of the relationships between features.
   * Train the model on historical taxi data and split the dataset into training and testing sets.
   * Ensure that performance metrics like Mean Absolute Error (MAE) or Root Mean Squared Error (RMSE) are in acceptable ranges.
2. **A/B Testing**:
   * Conduct A/B testing on the relationship between key variables (e.g., distance vs. fare) to confirm that predictions align with expected fare structures.
3. **Testing for Consistency**:
   * Evaluate the model's performance across different time periods, locations, and ride types to ensure robustness.
   * Perform cross-validation to ensure that the model does not overfit or underfit.

# Delivering and Presenting Results

**Tasks**

1. **Model Presentation**:
   * Create a final report and presentation summarizing the key findings, model performance, and recommendations.
   * Visualize the relationship between key variables (e.g., fare vs. distance) in a way that is digestible for non-technical stakeholders.
   * Provide predictions for sample trips and explain how the model handles various scenarios (e.g., traffic congestion, rush hour).
2. **Deliverables for TLC**:
   * Provide a set of visual aids that Titus Nelson can share with TLC executives, explaining the business value of fare predictions.
   * Highlight the cost and time-saving potential of the model in TLC’s operations.
3. **Future Steps**:
   * Recommendations for deploying the model in TLC’s systems for real-time fare estimates.
   * Propose a monitoring plan to continuously evaluate the model's performance and adjust for new data trends.

# Project Milestones

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **PACE Stage** | **Responsible Team** | **Due Date** |
| Project Kickoff & Stakeholder Alignment | Plan | Automatidata, TLC | Week 1 |
| Data Cleaning & Preprocessing | Analyze | Automatidata Data Team | Week 2 |
| Exploratory Data Analysis (EDA) | Analyze | Automatidata Data Team | Week 2 |
| Model Selection and Training | Construct | Automatidata Data Team | Week 3 |
| A/B Testing and Consistency Check | Construct | Automatidata Data Team | Week 3 |
| Final Model and Performance Review | Execute | Automatidata, TLC | Week 4 |
| Presentation and Report Delivery | Execute | Automatidata, TLC | Week 5 |