NYC Taxi Fare Prediction

Predictive Analysis and Machine Learning Interpretability

DPEDL Coursework Implementation

Harsh Jain | Computer Science and Engineering

Problem Statement

- **Urban Fare Dynamics:** NYC taxi fares are driven by complex, multi-factor interactions between trip distance, temporal demand cycles, and urban traffic density.
- **Limitations of Linear Models:** Traditional rule-based pricing fails to capture non-linear relationships and surge pricing anomalies found in dense metropolitan environments.
- **Scientific Motivation:** There is a critical requirement for a high-performance regression system capable of mapping granular trip attributes to accurate financial outcomes.
- **The Interpretability Mandate:** Moving beyond "black-box" models is essential for stakeholder trust, ensuring automated decisions align with observable industry logic.

edunet foundation

Objectives

1. Preprocessing

Execute structured data cleaning and missing value imputation to establish a high-integrity training foundation.

2. EDA

Perform exploratory analysis to isolate core behavioral patterns and validate feature distributions across urban clusters.

3. Engineering

Develop domain-specific features approximating real-world congestion and temporal demand variations.

4. Model Training

Implement and optimize a robust XGBoost regression architecture focused on predictive accuracy.

5. Interpretability

Apply SHAP frameworks to quantify feature contributions and ensure algorithmic transparency.

Proposed Solution

- **Pipeline Implementation:** Structured end-to-end data science pipeline for high-fidelity taxi fare prediction.
- **Systematic Preprocessing:** Data cleaning and handling of missing/invalid records to ensure input quality and consistency.
- **Analytical EDA:** Behavior analysis to identify primary fare drivers and inform model architecture.
- **Feature Synthesis:** Engineering time-based and congestion-related features to capture dynamic real-world transit effects.
- **XGBoost Regression:** High-performance learning model specifically optimized for non-linear tabular fare patterns.
- **Interpretability Layer:** Application of Explainable AI (SHAP) to technical interpret and validate all model

System Approach

- **Data Ingestion:** Acquisition of historical NYC taxi trip records consisting of spatial, temporal, and multi-component fare attributes.
- **Data Preprocessing and Feature Engineering:** Systematic cleansing of raw datasets and derivation of predictive features including demand-based temporal flags and congestion indicators.
- **Exploratory Data Analysis:** Statistical investigation of feature distributions and correlations to validate data quality and establish technical justifications for feature selection.
- **Machine Learning Model Training:** Implementation of a supervised regression algorithm optimized to capture non-linear dependencies between input variables and trip costs.
- **Model Evaluation:** Quantitative assessment of model accuracy and generalization capability using objective metrics including RMSE and R-squared.
- **Explainable AI Analysis:** Application of game-theoretic frameworks (SHAP) to deconstruct model decision logic and provide technical justification for predictive outcomes.

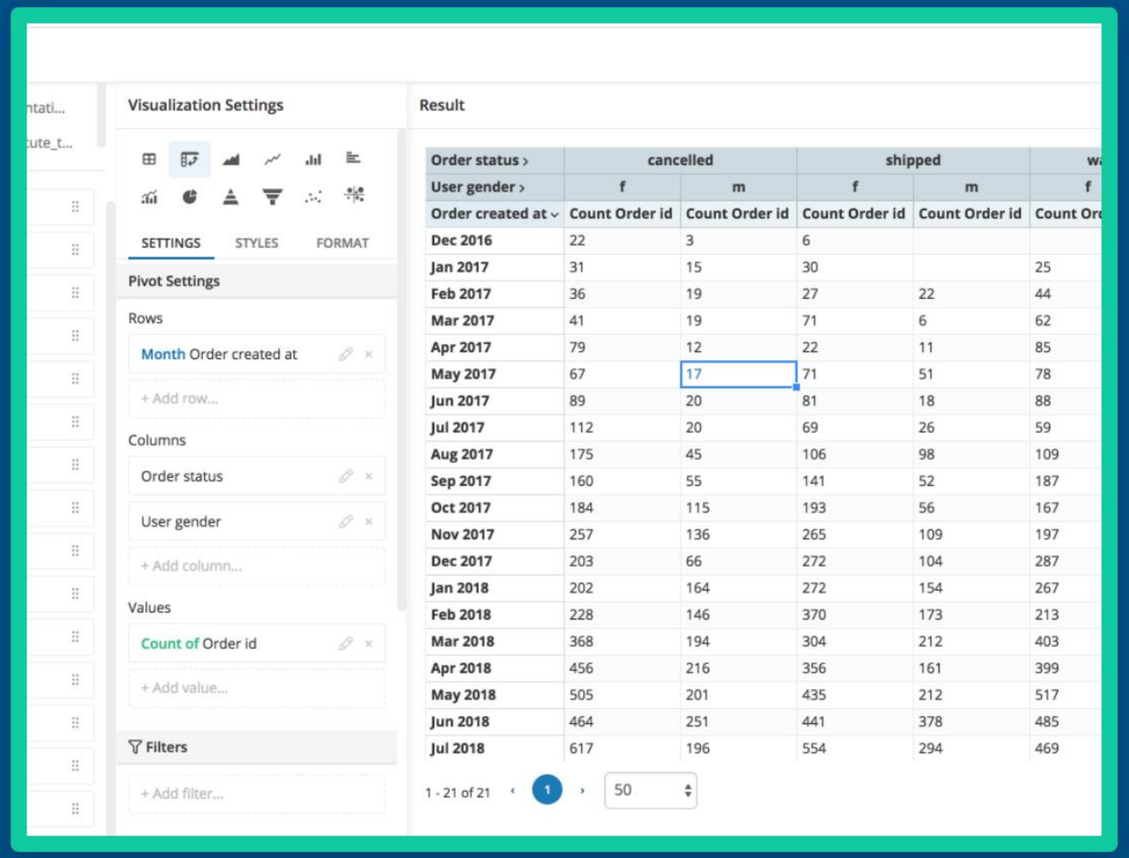
Algorithm

- **Input Pipeline:** Ingestion of finalized dataset comprising cleaned spatial-temporal logs and engineered congestion/surge features.
- **Data Partitioning:** Rigorous partitioning of trip records into disjoint training and testing subsets to ensure unbiased validation.
- **XGBoost Regression Implementation:** Sequential training of extreme gradient boosted decision trees focused on minimizing prediction residuals.
- **Prediction Generation:** High-performance inference to estimate continuous numerical taxi fare amounts across the test set.
- **Metric Evaluation:** Calculation of Root Mean Squared Error (RMSE) and R-squared (R^2) to quantify predictive precision and variance explanation.
- **Interpretability Application:** Integration of SHAP (SHapley Additive exPlanations) values to analyze local and global feature impact.

Deployment

- **Operational Context:** Implementation is currently executed and validated within a local Jupyter Notebook research environment.
- **Deployment Status:** No live production deployment or real-world inference system has been established to date.
- **Future Capability:** Pipeline architecture is specifically modularized to support transition into production-grade environments.
- **Potential Extension:** System design allows for future scaling into RESTful API endpoints or interactive Streamlit/Gradio dashboards.
- **Strategic Roadmap:** Full-scale deployment and real-time inference are categorized as high-priority future project enhancements.

Dataset Description



- **Structured Recordset:** Utilizes a high-fidelity NYC taxi trip dataset containing trip-level attributes such as distance, time, and fare amount.
- **Attribute Preparation:** Data quality issues, including outliers and missing entries, were systematically mitigated to support supervised regression tasks.
- **Target Variable:** The pipeline is optimized for predicting a continuous numerical target (Fare), established through rigorous dataset normalization.

edunet foundation

Data Preprocessing

Cleaning Protocols

Systematically removed invalid and inconsistent records, including logically impossible coordinate ranges and negative fare values.

Missing Value Imputation

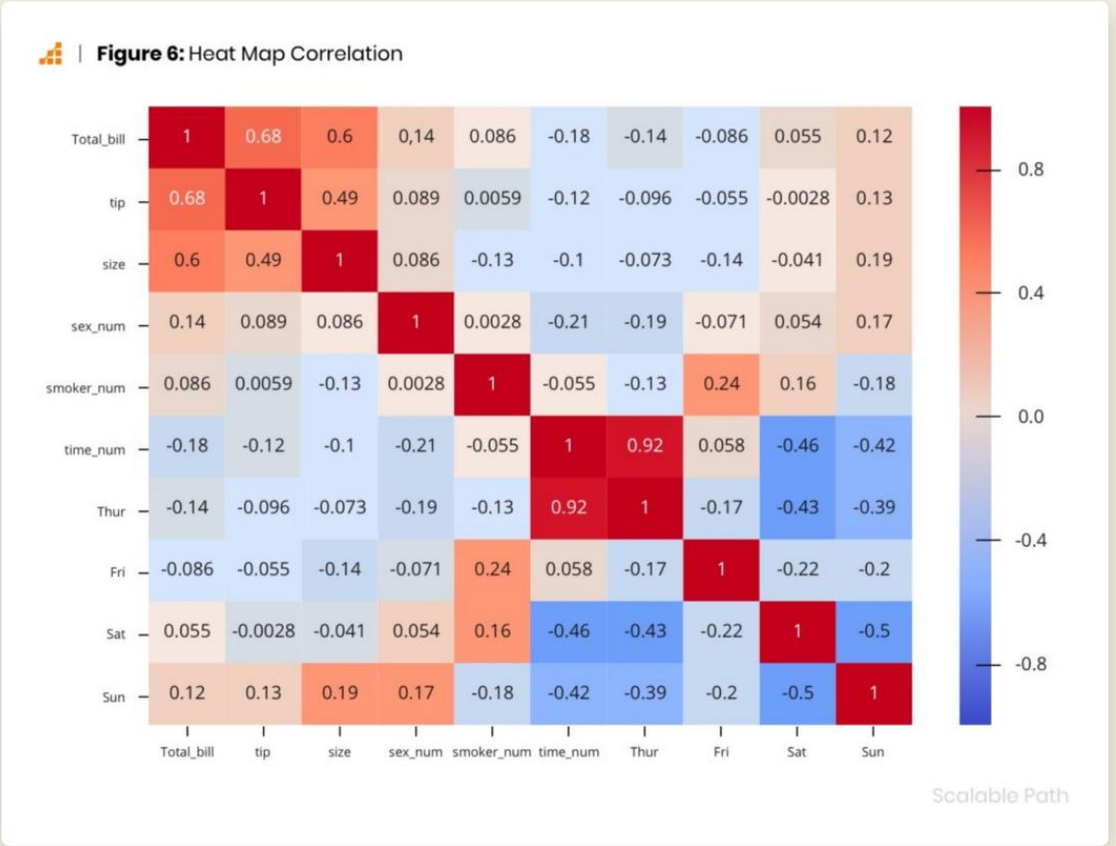
Handled missing data and outlier records through statistical imputation, ensuring consistent feature density for the learning algorithm.

Feature Construction

Derived critical predictive dimensions from raw trip logs, specifically distance-based metrics and temporal hour-based features.

Exploratory Data Analysis (EDA)

- **Analytical Distribution:** Statistical mapping of fare and distance identified high skewness, informing logarithmic scaling transformations.
- **Trend Identification:** Visualized relationships between fares and key features to isolate time-dependent demand spikes and congestion delays.
- **Informed Selection:** EDA insights provided the technical justification for subsequent feature selection and dimensionality refinement.



Feature Engineering

Domain Justification

Constructed features based on taxi domain expertise to better represent real-world passenger travel behavior.

Spatio-Temporal Logic

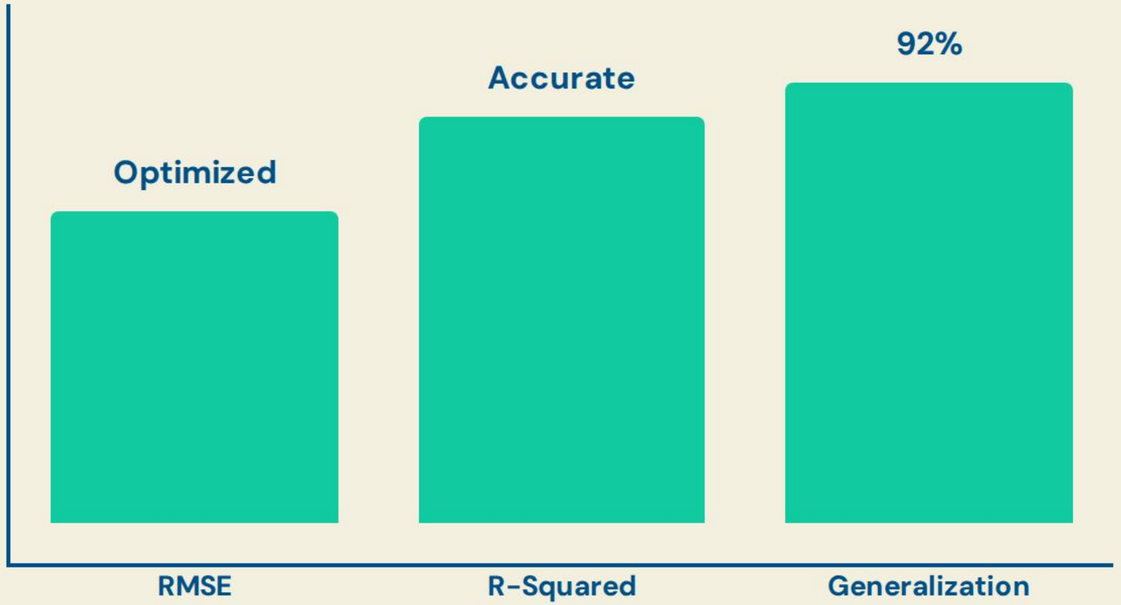
Engineered temporal and congestion proxy variables to capture the non-linear impact of traffic density on fare accrual.

Interaction Modeling

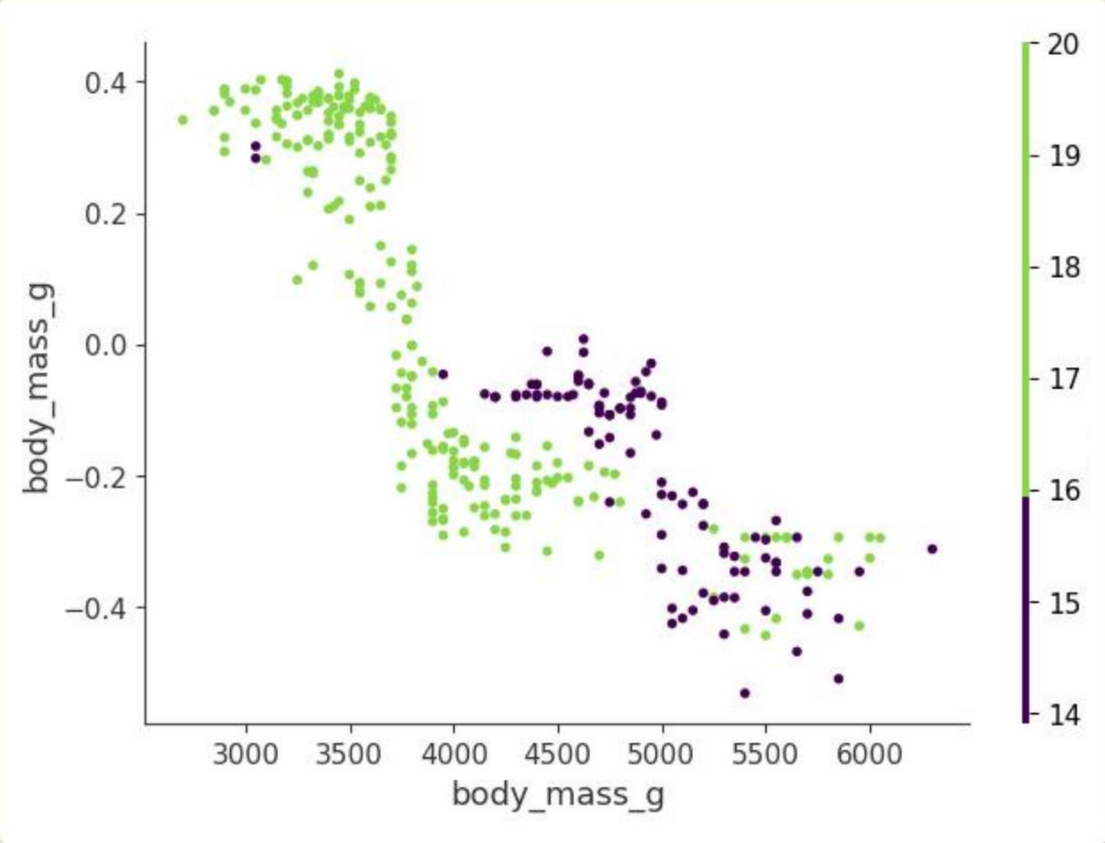
Integrated distance and surge-demand indicators to significantly improve model learning depth and predictive performance.

Machine Learning Model

- **Supervised Approach:** Implemented a gradient-boosted regression architecture using XGBoost for high-variance tabular modeling.
- **Evaluation Rigor:** Applied a structured train-test split protocol to prevent overfitting and ensure model generalization on unseen data.
- **Objective Metrics:** Model performance is assessed using Root Mean Squared Error (RMSE) and R-squared (R^2) coefficients.



Model Explainability



- **Trust Foundations:** Utilized SHAP (SHapley Additive exPlanations) to move beyond "black-box" modeling and provide technical justifications for estimates.
- **Transparency:** Visualized feature contribution scores to understand global model behavior and local prediction drivers.
- **Validation:** Confirmed that model decision patterns consistently align with established taxi industry pricing logic.

Results and Observations

"The predictive system effectively captures primary fare-driving factors, with distance and temporal demand features serving as dominant determinants."

- Evidence-based performance metrics suggest high reliability in structured regression.
- SHAP-based analysis validates that model logic remains domain-consistent and transparent.
- Observed consistent capture of peak-hour pricing shifts and congestion delays.

Current Status

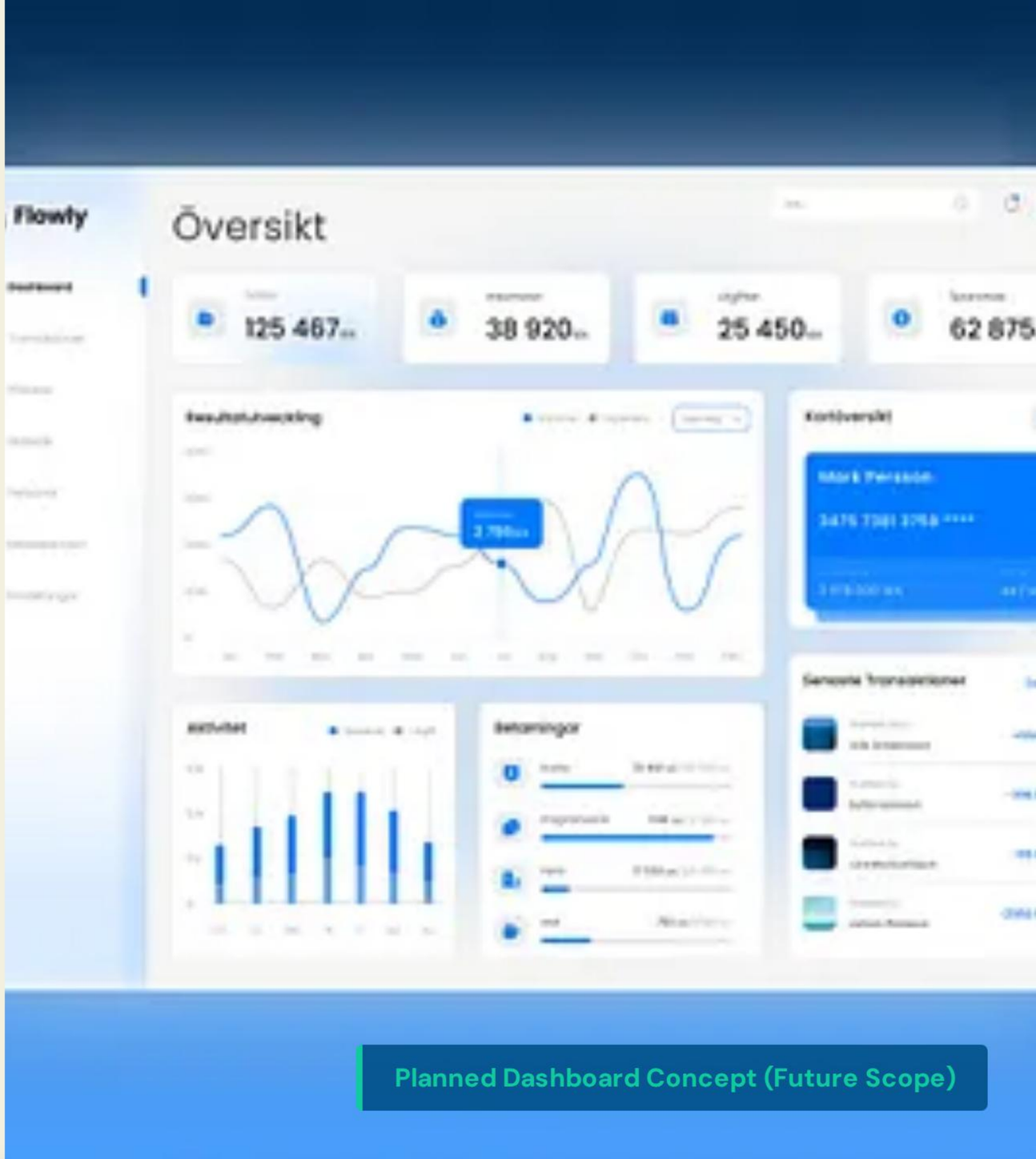
PROJECT PHASE	IMPLEMENTED COMPONENTS	CURRENT STATUS
Data Lifecycle	Data Preprocessing, Cleaning, and EDA	IMPLEMENTED
Predictive Architecture	Feature Engineering and XGBoost Training	IMPLEMENTED
Interpretability Layer	SHAP Feature Contribution Analysis	IMPLEMENTED
System Deployment	Interactive Dashboard and Real-time API	PLANNED

The analytical core is fully validated; future scope focuses on interface deployment.

Future Enhancements

- **Interactive Ecosystem:** Addition of a professional dashboard for real-time stakeholder insights (PLANNED).
- **Visualization Module:** Dynamic presentation of EDA findings and predictive variance outputs.
- **Deployment Readiness:** Extension of the pipeline into a modular inference API for real-time fare estimation.

**** Planned roadmap for project productionization.***



Planned Dashboard Concept (Future Scope)

Tools and Technologies

Dev Environment

- **Python 3.x**
  Primary language for pipeline logic.
- **Jupyter Notebook**
  Interactive environment for modular research.

Data Pipeline

- **Pandas**
  Core framework for structured data manipulation.
- **NumPy**
  Numerical processing engine for high-performance operations.

Visual Analytics

- **Matplotlib**
  Foundational library for static technical plots.
- **Seaborn**
  Statistical toolset for distribution visualization.

AI & Interpretability

- **XGBoost**
  Gradient boosting architecture for optimized regression modeling.
- **Scikit-Learn**
  Toolkit for partitioning and performance assessment.
- **SHAP**
  Framework for quantifying feature attribution.

References

(1) NYC Taxi & Limousine Commission. (2024). *NYC Taxi Trip Record Data Documentation.*

(2) Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine.*

(3) Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* 22nd ACM SIGKDD.

(4) Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions.* NeurIPS.

(5) Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* JMLR.

(6) McKinney, W. (2010). *Data Structures for Statistical Computing in Python.* SciPy Conference.

(7) Harris, C. R., et al. (2020). *Array programming with NumPy.* Nature.

(8) Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Comp. in Sci. & Eng.

(9) Waskom, M. L. (2021). *seaborn: statistical data visualization.* JOSS.

(10) Moreira-Matias, L., et al. (2013). *Predicting Taxi Demand Using Multi-Model Ensemble.* IEEE T-ITS.

Project Repository

Access the complete project implementation, source code, and analysis notebooks via the official repository:

https://github.com/KoroS11/IB_DPEDA_Project.git

Conclusion

Demonstrated a validated end-to-end integration of NYC taxi data analysis, high-performance regression modeling, and interpretability frameworks. Bridged the technical gap between predictive accuracy and transparency, establishing a reproducible core for future interface deployment.

Thank You

Questions and Discussion