Predicting New York City Taxi Fares Using Machine Learning

May 28, 2025

Executive Summary

This report outlines a data-driven solution designed to predict taxi fares accurately in real-time using machine learning. The project focuses on NYC Yellow Taxi data from January 2019 and explores how modern AI techniques can enhance fare estimation, customer transparency, and operational efficiency.

Business Context and Motivation

In a highly competitive ride-hailing and taxi service market, accurate fare estimation is critical for:

- **Improving Customer Trust:** Transparent, consistent pricing increases customer satisfaction and reduces fare disputes.
- **Dynamic Pricing and Operational Efficiency:** AI-powered fare prediction enables better decision-making during high-demand periods, improving vehicle utilization.
- **Strategic Planning:** Insights derived from trip data can support route optimization, driver allocation, and revenue forecasting.

This initiative supports the integration of artificial intelligence into core business operations, positioning the company as a tech-enabled, future-ready transport provider.

Approach Overview

Our solution follows a structured end-to-end machine learning pipeline:

1. **Data Analysis and Cleansing:** Cleaned over 6.9 million records by removing duplicates, handling missing values, and filtering out outliers.
2. **Feature Engineering:** Created over 10 new features such as trip duration, time-of-day segments, average speed, and airport trip indicators to enrich predictive power.

3. **Model Development:** Trained multiple algorithms including Gradient Boosting, k-Nearest Neighbors, and Neural Networks for both fare amount prediction (regression) and fare class prediction (classification).
4. **Evaluation:** Gradient Boosting Regressor emerged as the best-performing model with an R^2 score of 0.96, while the Bagging Classifier showed excellent F1-score performance (0.95) in classifying fare categories.
5. **Operationalisation:** Designed a practical deployment strategy covering real-time model serving, monitoring, retraining, and scalability using cloud-native tools (e.g., Docker, Kubernetes, REST APIs).

Key Results and Value Proposition

- **Highly Accurate Fare Prediction:** Our top model delivers over 95% predictive accuracy on real trip data.
- **Cost Efficiency:** Real-time fare estimation can reduce billing disputes, overcharges, and manual oversight.
- **Business Intelligence:** Clustering and correlation analyses provide actionable insights on customer patterns, fare structures, and high-revenue segments.

Implementation Plan

Deployment into production can be achieved in three phases:

- **Phase 1 – Pilot:** Integrate the model into internal dispatching tools to assess real-world performance.
- **Phase 2 – Customer Integration:** Embed predictive pricing into the mobile app or booking platform to offer fare previews.
- **Phase 3 – Full Automation:** Enable dynamic pricing based on predicted demand, trip type, and congestion conditions.

The model is containerized and built for scalability, supporting thousands of fare predictions per second with minimal latency.

Strategic Recommendation

We recommend the immediate piloting of the model in one urban zone, followed by staged scaling. This AI-driven fare prediction tool is not just a technological upgrade—it is a strategic asset for modernizing operations, improving rider experience, and achieving competitive advantage.

This project represents a tangible and low-risk opportunity to integrate artificial intelligence into the core of our service. The tools and methods are production-ready, the benefits are measurable, and the long-term ROI is substantial.