Project Report

Uber/Lyft Price Prediction

Date: September 18, 2025
Author: Mohamed Mahmoud & Hassan Abdul-razaq
Report Status: Final

Executive Summary

This project addresses the critical business challenge of **dynamic pricing** in the ride-hailing industry. Using a dataset of Uber ride estimates and weather data from Boston, we developed and evaluated machine learning models to predict ride fares. The objective was to identify key drivers of price fluctuations and build a model capable of accurate fare forecasting.

Two models were developed:

- **Linear Regression** achieved a strong baseline with an R^2 of **0.9208** and a MAPE of **11.64%**.
- **Artificial Neural Network (ANN)** significantly outperformed it, achieving a MAPE of **7.68%**, equivalent to **92.32% accuracy**.

This project serves as a **Proof-of-Concept (PoC)** demonstrating that ride prices can be modeled with high accuracy using features such as trip distance, service type, and location. The insights and methods developed here can be directly adapted to our company's application for **dynamic nurse/patient matching**, enabling optimized scheduling, improved revenue forecasting, and transparent pricing. Scaling with broader datasets will further strengthen the model.

1. Introduction

1.1 Problem Definition

Ride-hailing platforms like Uber and Lyft use dynamic pricing, where fares fluctuate based on demand, supply, and contextual factors. Predicting these fares is essential for maximizing efficiency and profitability.

1.2 Business Motivation

In our company’s context (dynamic nurse/patient matching), accurate fare prediction allows us to:

- **Optimize nurse allocation**: ensure fair compensation and availability during peak demand.
- **Improve financial planning**: forecast revenues and costs reliably.
- **Enhance user experience**: build trust through transparent pricing.

1.3 Project Goals

- Analyze factors influencing ride prices.
- Develop and evaluate predictive machine learning models.
- Provide a foundation for a scalable dynamic pricing system.

2. Dataset Description

The dataset combines Uber/Lyft ride data and Boston weather information, collected between late November and early December 2018.

2.1 Features

Key features included:

- **distance**: trip distance in miles.
- **cab_type, name**: service provider and product type (UberX, Black, etc.).
- **destination, source**: pickup and drop-off locations.
- **price**: target variable.

Dropped features:

- **time_stamp** (invalid, all 1970-01-01).
- **surge_multiplier** (output of pricing system, not predictive).
- **id** (unique identifiers, no predictive value).

2.2 Preprocessing & Feature Engineering

- Filtered for **Uber-only rides**.
- Removed rows with missing price (~55k).
- Final dataset size: **330,568 records**.
- **One-Hot Encoding** applied to categorical variables.

3. Exploratory Data Analysis (EDA)

- Most trips were short-distance rides, consistent with typical ride-hailing patterns.
- Short-to-medium trips dominated, making accuracy in this range especially critical.
- The dataset showed strong influence from trip distance, service type, and source/destination.

4. Modeling

Two regression models were tested.

4.1 Linear Regression

- Implemented with a pipeline (ColumnTransformer + LinearRegression).
- Served as baseline.

4.2 Artificial Neural Network (ANN)

- Architecture:
  - Dense(64, ReLU) + Dropout(0.3)
  - Dense(32, ReLU) + Dropout(0.3)
  - Dense(16, ReLU) + Dropout(0.2)
  - Output: Dense(1, Linear)
- Optimizer: Adam (lr=0.001).
- Loss: Mean Absolute Error (MAE).

5. Results

Model	R^2 Score	MAPE	Accuracy (approx)
Linear Regression	0.9208	11.64%	~88.36%
Artificial Neural Net	N/A	7.68%	~92.32%

- **Linear Regression:** achieved **R^2 = 0.9208** with MAPE = 11.64%.
- **ANN:** achieved **MAPE = 7.68%**, outperforming the baseline with stronger predictive power.

6. Discussion & Limitations

6.1 Key Insights

- **Distance** is the strongest predictor of price.
- **Service type** (UberX, Black, etc.) significantly impacts fares.
- **Location (source/destination)** strongly influences dynamic pricing.
- ANN captured complex non-linear relationships, outperforming linear regression.

6.2 Proof-of-Concept (PoC) Limitations

- **Short timeframe:** ~2 weeks of data, no seasonal/long-term effects.
- **Geographic scope:** Boston only.
- **Weather & events:** not integrated in the final model.
- **Simulated estimates:** API-collected fare estimates, not completed trip fares.

7. Future Work

- **Expanded data collection:** Gather longer timeframes, integrate traffic APIs, hospital shift data, and event calendars.
- **Advanced models:** Test XGBoost, LightGBM, and LSTM for time-series dynamics.
- **Hyperparameter optimization:** Tune ANN architecture for better performance.
- **Deployment:** Containerize model and deploy as an API for real-time pricing in our app.

8. Conclusion

This project demonstrates that **ride-hailing prices can be predicted with high accuracy** using machine learning. The ANN achieved ~92.32% accuracy, validating our approach.

For our company's nurse/patient matching app, this PoC confirms the feasibility of building a **dynamic, data-driven pricing system** that ensures:

- Fairness and transparency.
- Optimized resource allocation.
- Strong revenue management.

By scaling with broader, real-world data, this system can evolve into a **production-ready dynamic pricing engine**, strengthening our market position and delivering greater value to users.