### **Literature Review**

**Predictive Analytics for Rideshare Pricing**

**Anjala Boyapati**

### **Introduction**

Ride-hailing platforms such as Uber and Lyft leverage dynamic (or surge) pricing to balance supply and demand in real time. While this practice benefits companies and drivers, it presents an interesting challenge for predictive modeling: prices vary widely based on factors such as time of day, trip distance, local events, and weather conditions. Researchers, data scientists, and enthusiasts alike have attempted to unravel the determinants of these prices and forecast them using both classical machine learning (ML) and deep learning approaches.

This literature review focuses on works that have examined factors affecting Uber and Lyft’s pricing—particularly the influence of weather—and on data-driven methods to predict cab fares and surge multipliers. It begins by describing Kaggle-based exploratory analyses on a publicly available “Uber & Lyft Cab Prices” dataset, then expands to include academic and industry research discussing predictive models. This review concludes with a discussion of research gaps and how the present project aims to address them.

### **Body**

#### **1. Kaggle Analyses of Uber & Lyft Pricing Data**

**1.1 Explorations Using Random Forest and Basic ML Pipelines**Several Kaggle users have tackled the *Uber & Lyft Cab Rides* and *Weather* datasets, aiming to predict fares or surge multipliers.

* **Gaurang Gupta’s Surge Price Prediction Notebook (Kaggle)** demonstrated a workflow of data merging (cab rides with hourly weather features), data cleaning, and feature engineering (e.g., extracting day of week and hour). A RandomForestClassifier was then employed to classify the surge multiplier category (1.0 vs. higher values). The result was a high accuracy (over 95%) but required techniques like SMOTE to handle severe class imbalance for multipliers above 1.0.
* **Ravi Munde’s Random Forest Model** took a similar approach—merging ride data with weather features (temperature, cloud coverage, humidity, etc.) to predict both surge multipliers and actual prices. The use of a RandomForestRegressor to predict fare amounts and a RandomForestClassifier for surge multipliers exhibited good performance, with *distance* consistently reported as one of the most influential predictors. These Kaggle analyses underline the importance of data preprocessing, feature importance ranking (distance, time, weather attributes), and the interplay between spatiotemporal factors and price outcomes.

While these notebooks offer valuable proofs of concept, they mostly rely on structured ML workflows (Random Forest, label encoding, standard hyperparameter tuning). There is limited exploration of deep learning methods, advanced hyperparameter search, or more granular spatiotemporal modeling.

#### **2. Academic Research on Dynamic Pricing and Demand Prediction**

**2.1 Predicting Surge Pricing with Statistical and Machine Learning Methods**

* *Battifarano and Qian (2019)* studied real-time surge pricing in Pittsburgh, modeling Uber’s surge multipliers using regression (Lasso) and non-linear ML methods. They emphasized the importance of spatiotemporal clustering (segmenting locations and times of day) and found that historical surge multiplier values and real-time traffic conditions were crucial. Their work also showed how events and social factors contributed to localized price spikes, though they reported lower importance for weather than for traffic.

**2.2 Deep Learning for Short-Term Demand**

* *Chen, Thakuriah, and Ampountolas (2021)* proposed a CNN-based deep learning model (“UberNet”) to forecast short-term ride-hailing demand (i.e., pickups). Although their primary output was ride requests (rather than price), the paper highlighted the utility of spatial convolutions and temporal dilations for capturing complex patterns. Weather features played a role but were secondary to day/time variables, demographic data, and historical demand.

**2.3 Machine Learning for Surge Rate Prediction: International Use Cases**

* *Mukherjee (2024)* examined surge rate prediction for Yassir, a ride-hailing service in North Africa. The study tested a wide range of ML algorithms (lightGBM, XGBoost, etc.), showing that trip distance, base fare, and time variables were key drivers of surge. Weather parameters—humidity, wind velocity—also influenced surge rates but to a lesser degree. Their highest accuracy model (LightGBM) reached over 99% classification accuracy to detect whether surge applies.

Across these academic works, consistent themes emerge: (1) **distance and time** of day remain paramount features; (2) **historical price or demand** trends significantly bolster predictions; (3) **weather** factors can improve model performance but are often secondary to time-varying supply-demand signals. Moreover, advanced methods (CNNs, LSTMs, or ensemble tree models with hyperparameter tuning) often outperform classical regression or baseline models.

#### **3. Gaps and Opportunities**

Despite the progress in both Kaggle-based notebooks and formal research studies, several gaps remain:

1. **Comprehensive Data Integration**: Many studies partially merge or drop certain weather attributes and do not always handle missing data robustly.
2. **Handling Outliers and Imbalanced Classes**: Surge multipliers above 1.0 frequently appear in small proportions. While SMOTE and similar oversampling methods help, more sophisticated approaches for outlier detection and class imbalance can be explored.
3. **Rich Feature Engineering**: Feature interactions (e.g., the interplay between rain intensity, hour, location, local events) are not always thoroughly captured in existing models.
4. **Comparative Analysis of ML vs. Neural Networks**: Although some papers compare classical and advanced models, direct comparisons with consistent hyperparameter tuning remain limited.
5. **Scalable Deployment Considerations**: Most works remain exploratory, without detailing real-time architectures or computational constraints for large-scale, frequent predictions.

### **Conclusion**

Existing literature underscores the complexity of predicting ride-hailing fares and surge multipliers, highlighting the interplay of temporal patterns (peak commute hours, weekends), trip characteristics (distance, pickup/drop-off locations), and environmental factors (weather, local events). Kaggle notebooks demonstrate valuable exploratory pipelines and straightforward Random Forest approaches, while academic papers underscore more rigorous feature engineering, advanced ML models (Lasso, XGBoost, CNNs), and their predictive strengths.

**Uniquely, the present project** proposes to (1) merge the *Uber & Lyft Cab Rides* and *Weather* datasets comprehensively, (2) conduct thorough data cleaning, outlier management, and feature engineering, (3) evaluate both traditional ML (Random Forest, XGBoost) and deep learning models (potentially MLP or CNN architectures) with systematic hyperparameter tuning, and (4) provide robust validation methods (cross-validation, additional error metrics) for performance assessment. By expanding on prior works’ successes—particularly in feature engineering and model comparison—this project aims to offer a deeper understanding of how weather interacts with demand- and supply-side constraints to shape dynamic ride-hailing prices.

### **References**

Battifarano, M., & Qian, Z. (2019). Predicting real-time surge pricing of ride-sourcing companies. *Transportation Research Part C: Emerging Technologies, 107*, 444–462. https://doi.org/10.1016/j.trc.2019.08.019

Chen, L., Thakuriah, P., & Ampountolas, K. (2021). Short-term prediction of demand for ride-hailing services: A deep learning approach. *Journal of Big Data Analytics in Transportation, 3*(2), 175–195. https://doi.org/10.1007/s42421-021-00041-4

Mukherjee, K. (2024). Machine learning methods for surge rate prediction: A case study of Yassir. *Munich Personal RePEc Archive (MPRA), 122151*. https://mpra.ub.uni-muenchen.de/122151/

Gauranggupta123. (n.d.). Surge price prediction [Kaggle Notebook]. Kaggle.<https://www.kaggle.com/code/gauranggupta123/surge-price-prediction/notebook>

RaviMunde. (n.d.). Starter: Uber & Lyft ride prices: Random forest [Kaggle Notebook]. Kaggle.<https://www.kaggle.com/code/ravi72munde/starter-uber-lyft-ride-prices-random-forrest>

Uber & Lyft Cab Prices Dataset. (n.d.). *Kaggle Datasets*. https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma