Machine Learning Methods for Surge Rate Prediction: A Case Study of Yassir

Dr. Krishnendu Mukherjee

Former Application Architect, o9 Solutions

Abstract. Transportation Network Companies (TNCs) face two extreme situations, namely, high demand and low demand. In high demand, TNCs use surge multiplier or surge rate to balance the high demand of riders with available drivers. Willingness of drivers, willingness of riders to pay more and appropriate surge rate play a crucial role in maximizing profits of TNCs. Otherwise, a considerable number of trips can be discarded either by drivers or riders. This paper explains an application of a combined classification and regression model for surge rate prediction. In this paper, twenty-six different machine learning (ML) algorithms are considered for classification and twenty-nine ML algorithms are considered for regression. A total of 55 ML algorithms is considered for surge rate prediction. This paper shows that estimated distance, trip price, acceptance date and time of the trip, finishing time of the trip, starting time of the trip, search radius, base price, wind velocity, humidity, wind pressure, temperature etc. determine whether surge rate or surge multiplier will be applied or not. The price per minute applied for the current trip or minute price, base price, cost of the trip after inflation or deflation (i.e. trip price), the applied radius search for the trip or search radius, humidity, acceptance date of the trip with date and time, barometric pressure, wind velocity, minimum price of the trip, the price per km etc., on the other hands, influenced surge rate A case study has been discussed to implement the proposed algorithm.

Keywords: *Machine Learning, Surge Rate Prediction, Surge Price, Classification, Regression, Random Forest, Light GBM, XGBoost*

1 Introduction

Surge price, a spatio-temporal function of vehicle supply and user demand, is a dynamic pricing method to fulfill the unanticipated demand during peak time by incentivizing the driver's participation (Battifaranoa and Qian, 2019; Karamanis *et* al., 2020). Transportation Network Companies (TNCs) such as Uber and Lyft are using the surge multiplier or surge rate to increase the fare if the number of requests for service at a particular location exceeds the number of available vehicles. Thesurge multiplier allures drivers to relocate to serve the demand and encourages customers to postpone their rides or trips (Battifaranoa and Qian, 2019). The demand for a ride is generally influenced by traffic congestion, socio-demographic factors, weather conditions, non-availability of public transport, holidays, festivals, and personal exigency of the rider. During the low demand, known as *wild goose chase* (WGC), price falls, and idle drivers are forced to pick up distant riders at the cost of their earnings (Castillo*y et* al., 2017). TNCs maximize their profits during prime time or peak demand of riders and are compelled to pay discounts to the riders during happy hour or low demand of riders. Artificial Intelligence (AI) such as deep learning, machine learning etc are considered to be effective to predict surge rate (Ke *et al.,* 2017; Li *et* al., 2017; Wei and Chen, 2012; Battifaranoa and Qian, 2019; Silveira-Santos *et* al., 2023). Uber is using a modified Long Short Term Memory (LSTM) architecture for heterogenous timeseries to predict surge multiplier (Laptev *et* al.,2017; Zhu and Laptev,2017). Lyft, on the other hand, uses causal machine learning, reinforcement learning, and optimization to predict surge multiplier. Uber is using predictive analysis and Lyft is using prescriptive analysis for surge multiplier prediction. In this paper, decision tree-based approaches are considered.

The proposed method departs from prior work in the following ways:

1. In this paper, 55 different machine learning algorithms are considered to predict surge rate on hourly basis on a data-driven model.
2. In 2019, Battifaranoa and Qian considered separate models for each location. In this paper, a classification-regression model is considered for all locations. When to apply surge rate? What would be the value of the surge rate? – are the two most important questions for TNCs. A Binary classifier is used to

answer the first question. Regression is used to answer the second question.

3. The literature review shows that the majority of the researchers are considering case studies of Uber and Lyft. Case study of Yassir has been referred to by Belli and Ali Djoudi (2023), Nadia (2024), and Sihem (2024). However, no research paper has been found on the application of machine learning to predict the surge rate of Yassir Ride-Hailing System.

The rest of the paper is organized as follows. In section 2, the proposed methodology is discussed. In section 3, a case study of Yassir is explained lucidly and section 4 is dedicated to results and discussion.

2 Methodology

The proposed approach integrates both classification and regression. A binary classifier is used to predict whether the surge rate will be applicable or not. If it is applicable, then the regressor predicts the surge rate. Twenty-six different classifiers and twenty-nine different regressors are considered in this paper, mentioned in Table 1.

Table 1. Models (or methods) selected for forecasting surge rate

Sl No	Classifier	Sl No	Regressor
1	LGBM Classifier	27	Extra Trees Regressor
2	XGBoost Classifier	28	Random Forest Regressor
3	Bagging Classifier	29	Bagging Regressor
4	Decision Tree Classifier	30	Decision Tree Regressor
5	Random Forest Classifier	31	XGBoost Regressor
6	Extra Trees Classifier	32	Extra Tree Regressor
7	Extra Tree Classifier	33	KNeighbors Regressor
8	KNeighbors Classifier	34	LGBM Regressor
9	Label Propagation	35	Histogram-based Gradient Boosting Regressor
10	Label Spreading	36	Gradient Boosting Regressor
11	SVC	37	MLP Regressor
12	AdaBoost Classifier	38	NuSVR
13	BernoulliNB	39	AdaBoost Regressor
14	Nearest Centroid	40	TransformedTarget Regressor
15	Perceptron	41	Linear Regression
16	Passive Aggressive Classifier	42	Lars
17	Quadratic Discriminant Analysis	43	Ridge
18	GaussianNB	44	Ridge CV

19	Dummy Classifier	45	Lasso Lars IC
20	Ridge Classifier	46	Bayesian Ridge
21	Ridge Classifier CV	47	SGD Regressor
22	Linear SVC	48	Orthogonal Matching Pursuit CV
23	SGD Classifier	49	Lasso CV
24	Logistic Regression	50	ElasticNet CV
25	Calibrated Classifier CV	51	LassoLars CV
26	Linear Discriminant Analysis	52	Lars CV
		53	Tweedie Regressor
		54	Orthogonal Matching Pursuit
		55	Poisson Regressor

❖ LGBM: Light Gradient Boosting Method, XGboost: Extreme Gradient Boosting, CV: Cross-validation, SVC: Support Vector Classifier, SVR: Support Vector Regressor, SGD: Stochastic Gradient Descent, MLP: Multi-layer Perceptron, LaasoLarsIC: LASSO model fit with LARS with Akaike information criterion (AIC) or Bayesian information criterion (BIC)

According to Elliott *et* al. (2015), *"forecast combination offers one approach for dealing with the effects of estimation error, model uncertainty, and instability in the underlying data generating process. By diversifying across multiple models, combinations typically deliver more stable forecasts than those associated with individual models."* In this paper, thus, a combination of multiple ML algorithms is considered. Data preprocessing or exploratory data analysis (EDA), baseline model selection, hyperparameter tuning, selection of metrics for classification and regression, model validation, and post-processing are considered in the proposed approach. The given dataset contains 18,184 rider's requests from 1st May 2023 to 2nd May 2023 from Algeria, Tunisia, and Senegal. The weather dataset is further merged with the given dataset to study the exogenous effect. The merged dataset contains 18,184 rows and 49 columns. Considering the limited size of the dataset, only machine learning algorithms are considered. Fig 1 shows the total number of trips discarded in pickup country and its corresponding pickup daira. Oran, Cheraga, Dar El Beida, Bar Mourad Rais, and Sidi M'Hamed are the top 5 daira in Alegria or Algerie where trip has been mostly discarded.

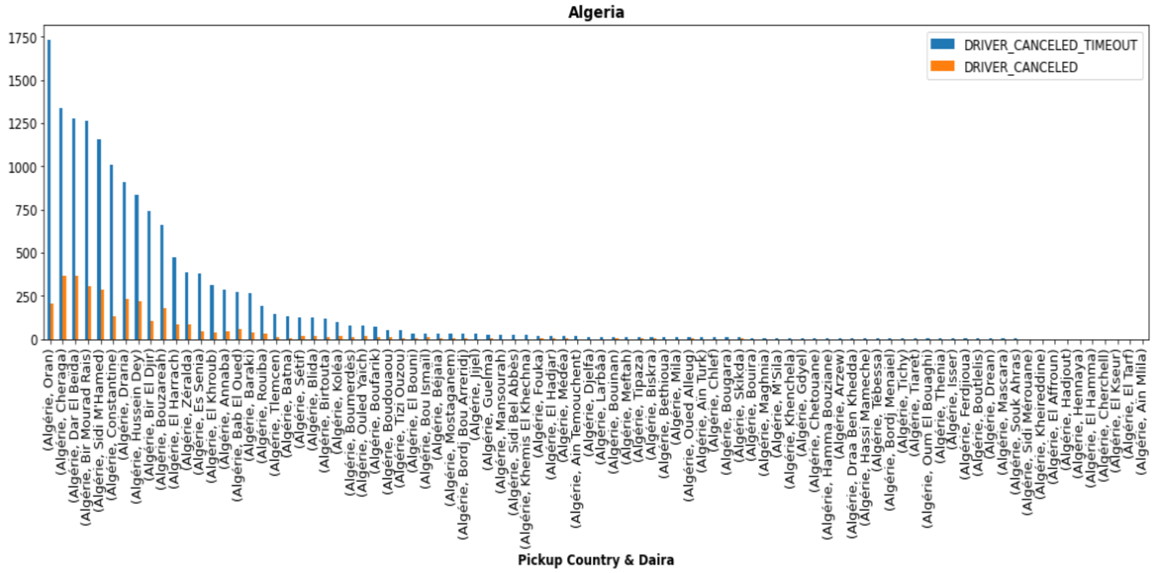


Fig. 1. Cancelled trips in Algeria

Large numbers of rider requests came on Thursday and Friday in Algeria, Senegal, and Tunisia, shown in Fig 2 and Fig 3. Oran, Alger, Constantine, Blide etc are some of the most demanding pickup wilaya in Algeria, shown in Fig 2. The average ride duration is longer on Thursday than on Friday in Senegal and Algeria. Moreover, the average ride duration is more in Senegal than Algeria irrespective of more rider's requests in Algeria, as a considerable number of ride requests have been discarded in Algeria, as shown in Fig 3. Exploratory data analysis also reveals that the total number of trips with surge rate is influenced by time and place. In Alger, from 6:19 AM to 8:20 AM, the total number of trips with surge is not considerably high. From 10:00 AM to 17:30 PM, surge trips remain high in Alger. In Oran, from 12:00 PM to 18:30 PM, the total number of trips with surge remains high. In Blida, the total number of trips with surge rate increased linearly from 13:53 PM to 15:15 PM, shown in Fig 4. In Fig 5, pickup wilayas of Algeria, Senegal, and Tunisia are shown for better understanding of the ride-hailing network.

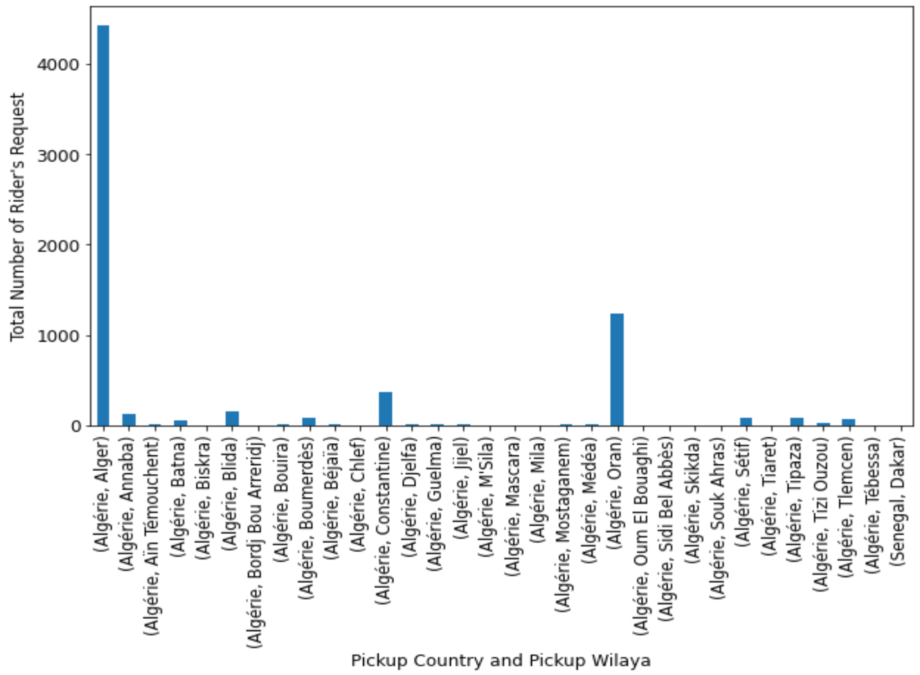
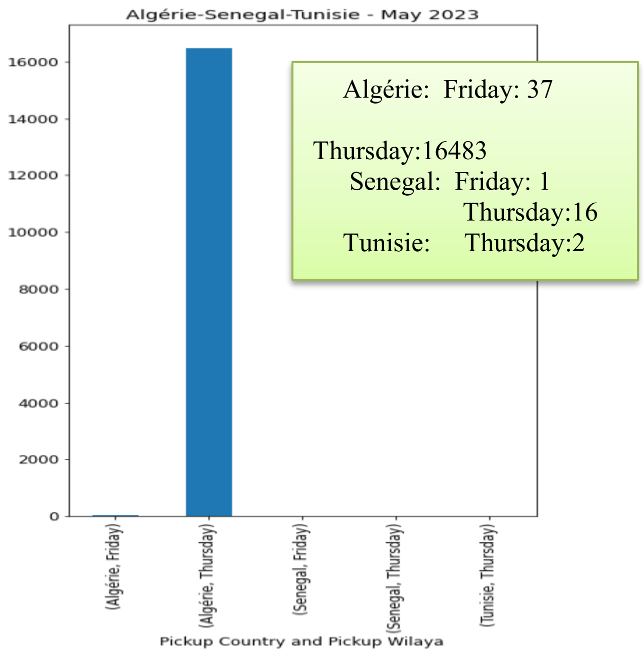


Fig. 2. Total number of rider's request



Algérie: Friday: 37

Thursday:16483

Senegal: Friday: 1

Thursday:16

Tunisie: Thursday:2

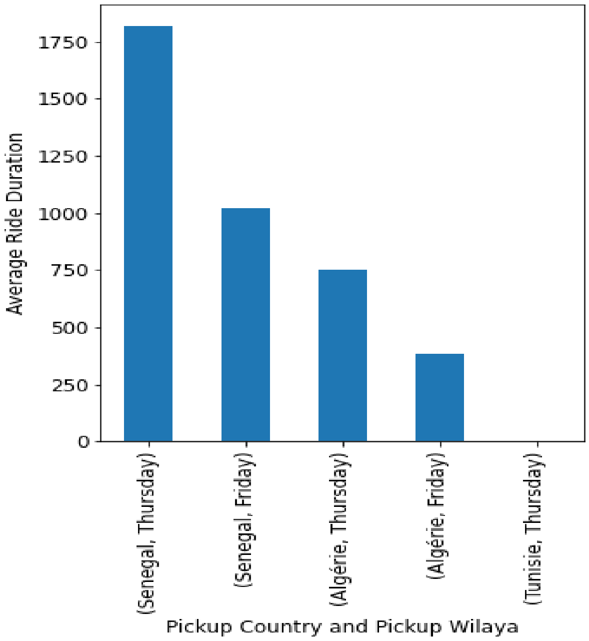


Fig. 3. (a) Total number of requests received in Algeria, Senegal, and Tunisia (b) Average ride duration on Thursday and Friday

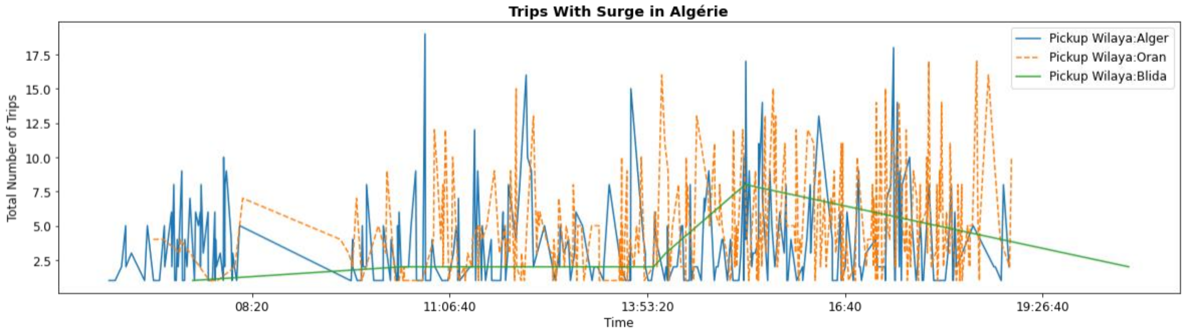


Fig. 4. Total number of trips with surge rate vs time

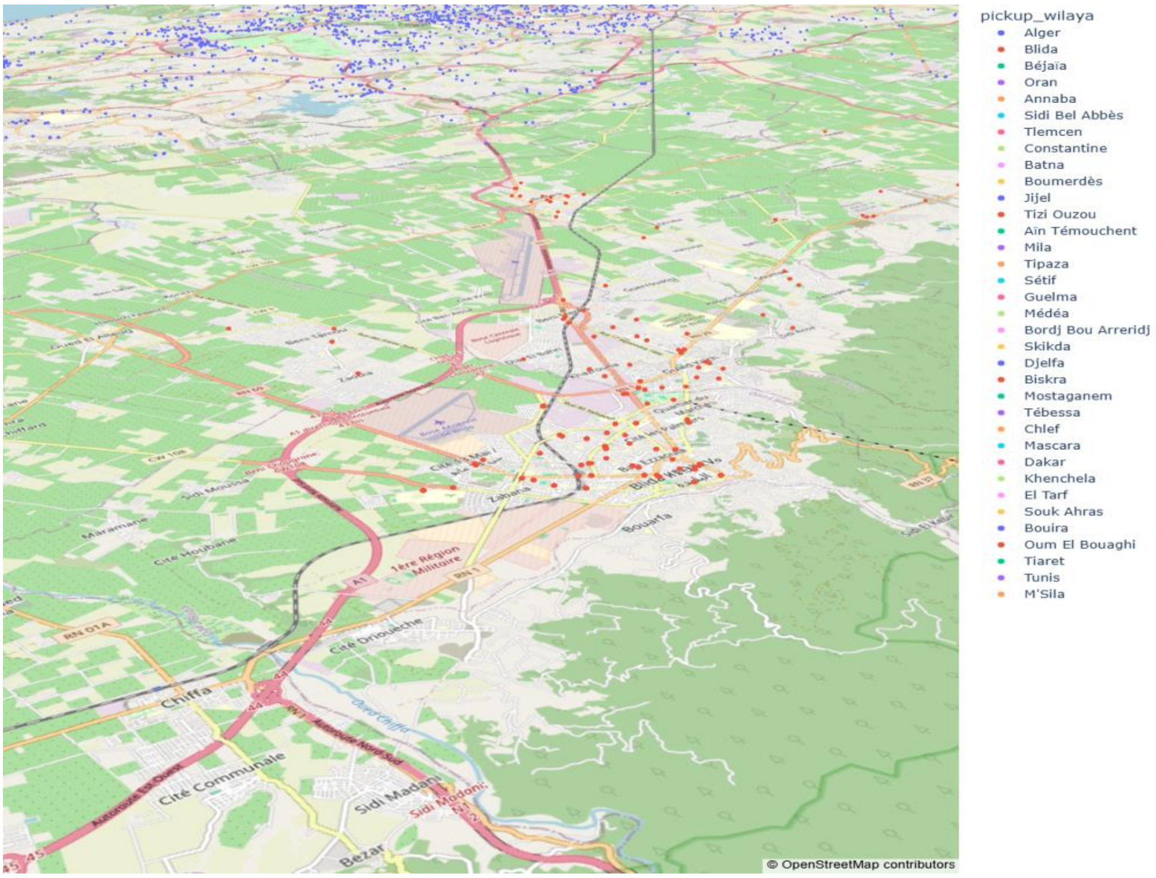


Fig. 5. Pickup Wilayas in Algeria, Senegal, and Tunisia

The best classifier is selected based on the suitability of bigdata application, accuracy, ROC AUC, F1 score, and computational time as shown in table 2. The Light Gradient Boosting Method (LGBM) is much faster than the Extreme Gradient Boosting (XGBoost) method. Both LGBM and XGBoost are boosting methods, which are a popular ensemble method in machine learning. Random Forest is also an ensemble method. Ride-hailing system generally needs bigdata platform such as Spark. Spark-ML has Random Forest, Light GBM, XGBoost etc. Further, CUDA and Spark can be combined to implement XGBoost at lightning speed. Light GBM is selected for classifier and XGBoost is selected for regressor.

Table 2. Twenty-six different algorithms for classifier

Classifier Name	Accuracy	ROC AUC	F1 Score	Computational Time
LGBM Classifier***	0.992701	0.982422	0.992673	0.095381
XGBoost Classifier	0.991241	0.977771	0.99119	1.456928
Bagging Classifier	0.990511	0.975445	0.990447	0.140591
Decision Tree Classifier	0.985401	0.972415	0.985401	0.031249
Random Forest Classifier***	0.988321	0.972254	0.988254	0.453947
Extra Trees Classifier	0.987591	0.969929	0.987507	0.282125
Extra Tree Classifier	0.984672	0.966304	0.984598	0.015626
KNeighbors Classifier	0.963504	0.923467	0.963221	0.187645
Label Propagation	0.961314	0.905134	0.960411	1.020696
Label Spreading	0.959854	0.902376	0.958918	1.601646
SVC	0.892701	0.748988	0.88664	0.424916
AdaBoost Classifier	0.879562	0.707128	0.869799	0.234214
BernoulliNB	0.766423	0.700594	0.788517	0.015626
Nearest Centroid	0.737226	0.685171	0.765873	0
Perceptron	0.835036	0.663687	0.830646	0.015696
Passive Aggressive Classifier	0.786131	0.651717	0.796231	0.015617
Quadratic Discriminant Analysis	0.448905	0.582311	0.505665	0.031241
GaussianNB	0.213869	0.529981	0.153527	0.015552
Dummy Classifier	0.843066	0.5	0.77128	0
Ridge Classifier	0.843066	0.5	0.77128	0.016129
Ridge Classifier CV	0.843066	0.5	0.77128	0.016009
LinearSVC	0.841606	0.499134	0.770555	0.221897

SGD Classifier	0.840876	0.498701	0.770192	0.031651
Logistic Regression	0.837956	0.49697	0.768737	0.06241
Calibrated Classifier CV	0.837226	0.496537	0.768372	0.156731
Linear Discriminant Analysis	0.836496	0.496104	0.768007	0.03207

❖ CV: Cross-validation, SVC: Support Vector Classifier, *** Selected Classifier for Analysis

Table 3. Twenty-nine different algorithms for regressor

Regressor	Adjusted R-Squared	R-Squared	RMSE	Time Take
Extra Trees Regressor***	0.902939218	0.903790006	0.026148764	0.377552509
Random Forest Regressor	0.891564033	0.892514531	0.027638594	0.880922079
Bagging Regressor	0.890034378	0.890998284	0.027832854	0.093739748
Decision Tree Regressor	0.884809215	0.885818923	0.028486437	0.015816212
XGBoost Regressor***	0.865745731	0.866922539	0.030753413	0.130310297
Extra Tree Regressor	0.864863407	0.86604795	0.030854304	0.015708685
KNeighbors Regressor	0.798075895	0.799845866	0.037715802	0.046948195
LGBM Regressor	0.730968101	0.733326306	0.043534259	0.103543997
Histogram-based Gradient Boosting Regressor	0.726361486	0.72876007	0.043905394	0.337969065
Gradient Boosting Regressor	0.516647965	0.520884798	0.058352712	0.359773874
MLP Regressor	0.371877922	0.37738374	0.066519856	0.301112652
NuSVR	0.366484522	0.372037616	0.066804833	7.889201403
AdaBoost Regressor	0.342494206	0.348257587	0.068057981	0.046776533
TransformedTarget Regressor	0.085465116	0.093481492	0.080265591	0.021704674
Linear Regression	0.085465116	0.093481492	0.080265591	0.015771389
Lars	0.085465116	0.093481492	0.080265591	0.015999317
Ridge	0.085461573	0.09347798	0.080265746	0.01220274
Ridge CV	0.085429877	0.093446562	0.080267137	0.010226011
LassoLarsIC	0.085198449	0.093217162	0.080277292	0
Bayesian Ridge	0.085118224	0.09313764	0.080280812	0.015955687
SGD Regressor	0.08297759	0.09101577	0.080374678	0.00506115
Orthogonal Matching Pursuit CV	0.082613488	0.09065486	0.080390632	0.015626192
Lasso CV	0.081789652	0.089838245	0.080426721	0.07181716
ElasticNet CV	0.081738559	0.0897876	0.080428958	0.062970877
LassoLars CV	0.081368804	0.089421086	0.08044515	0.015675306
Lars CV	0.081368804	0.089421086	0.08044515	0.027891397
Tweedie Regressor	0.064021844	0.072226181	0.08120114	0.011476994
Orthogonal Matching Pursuit	0.047781029	0.056127725	0.081902599	0.015622616
Poisson Regressor	0.001692965	0.010443648	0.083861251	0

❖ CV: Cross-validation, SVR: Support Vector Regressor, LaasoLarsIC: LASSO model fit with LARS with Akaike information criterion (AIC) or Bayesian information criterion (BIC), SGD: Stochastic Gradient Descent, *** Selected regressor

3 Case Study

In 2017, Noureddine Tayebi conceived an idea to support every Algerian while he was in Palo Alto, California. His idea geminates the first ride-hailing service in Algeria, popularly known as Yassir. In 2019, Yassir raised 13M USD. In 2021, Yassir got the attention of Y Combinator, the world's largest incubator, and raised 30 M USD. In 2022, Yassir raised 130M USD from world class investors such as BOND and Y Combinator and opened their Tech Hub in Europe. Yassir has about 6 M users and 1,30,000 partners throughout 45 cities in Algeria, Morocco, South Africa, Senegal, Canada, France, and Tunisia. Such an exponential rise will encourage several techies and entrepreneurs to dream and contribute for our society.

Surge pricing or demand pricing is time-based pricing and its sole objective is to adjust price as per the changing demand. Generally, the surge pricing system considers market activities such as driver's utilization rate, supply etc and determines the price as per predefined business rules. Artificial intelligence is well suited for extracting patterns of changing demand and predicting the surge prices as per the hypothesis function. Yassir Pricing System is moving from a manual system to an AI-based system. It is using Geohash, a public domain geocode system invented in 2008 by Gustavo Niemeyer, and predicting the surge price as per the optimal pricing rules at Geohash level. Interested readers can refer [5] and [14] for Geohash. The optimal pricing system should maximize demand and total driver earnings to maximize Yassir earnings.

A classification model is combined with a regression model to predict surge rates by considering the exogenous effect of the weather on surge price. The proposed model is shown in fig 6.

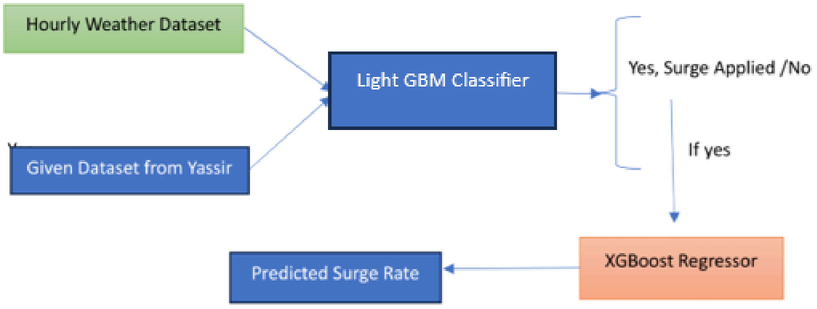


Fig.6. Simple classification-regression model for surge rate prediction

4 Results and Discussion

The study shows that estimated distance, trip price, acceptance date and time of the trip, finishing time of the trip, starting time of the trip, search radius, base price, wind velocity, humidity, wind pressure, temperature etc determine whether the surge rate or surge multiplier will be applied or not, shown in fig 7. Random Forest (RF) classifier, a supervised learning algorithm, was used to achieve 98.3% accuracy with feature engineering and hyperparameter optimization. Light GBM (LGBM), on the other hand, achieved 99.27% accuracy. Fig 7 shows that LGBM gave the highest priority to 'estimated distance' and second-highest priority to 'trip_price' whereas RF gave the highest priority to 'trip_price' and second-highest priority to 'estimated_distance'. LGBM outperformed all classifiers for the given dataset.

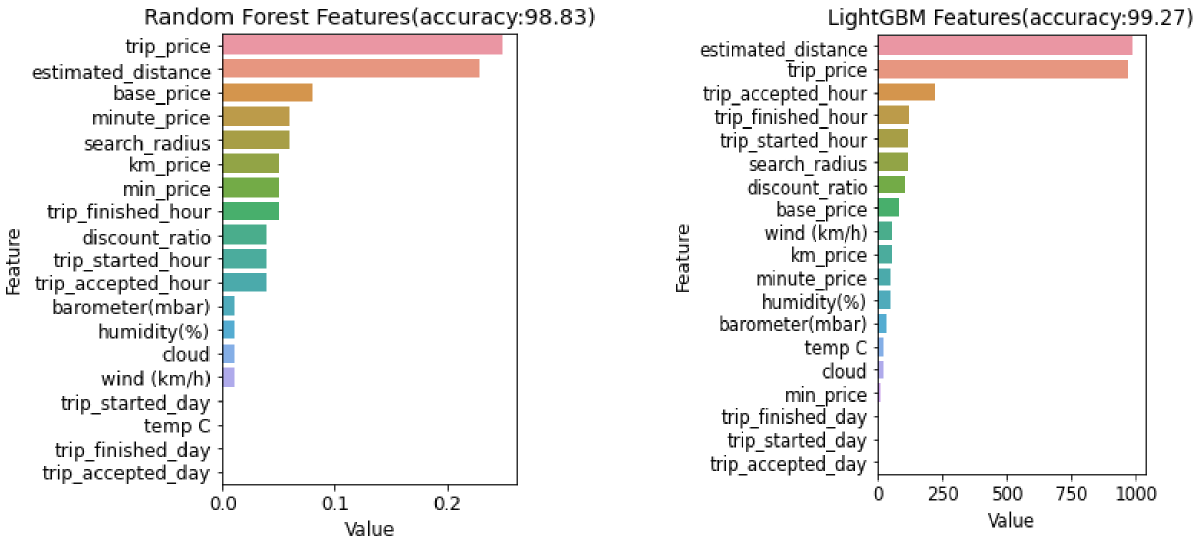


Fig.7. Features influencing ‘surge applied’

A simple correlation plot, shown in fig 8, was used to remove the multicollinearity effect.

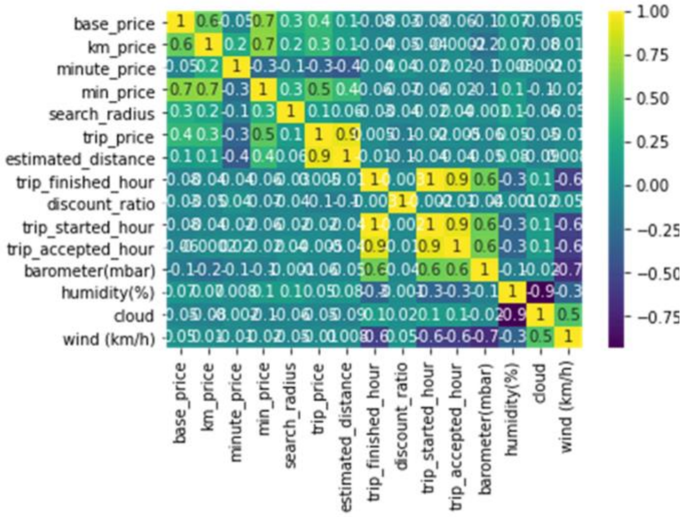


Fig.8. Correlation plot

OPTUNA, an open source hyperparameter optimization framework, was used for optimizing hyperparameters of XGBoost to achieve 0.031 RMSE. OPTUNA gave highest importance to alpha or L1 regularization to prevent overfitting, as shown in fig 9. It is pertinent to mention that Battifaranoa and Qian (2019) also used L1 regularization to predict surge price.

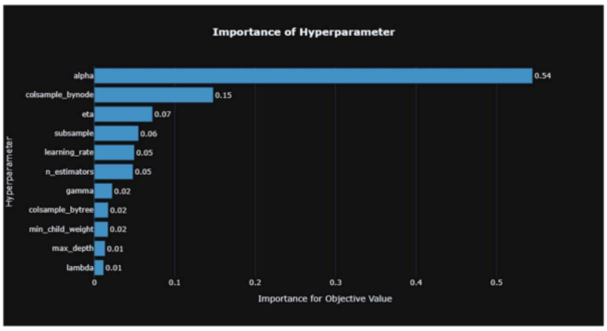


Fig.9. Hyperparameter optimization of XGBoost using OPTUNA

The study reveals that the price per minute applied for the current trip or minute price, base price, cost of the trip after inflation or deflation (i.e. trip price), the applied radius search for the trip or search radius, humidity, acceptance date of the trip with date and time or trip accepted hour, barometric pressure, wind velocity, minimum price of the trip, the price per km or km price etc influenced surge rate, shown in fig 10.

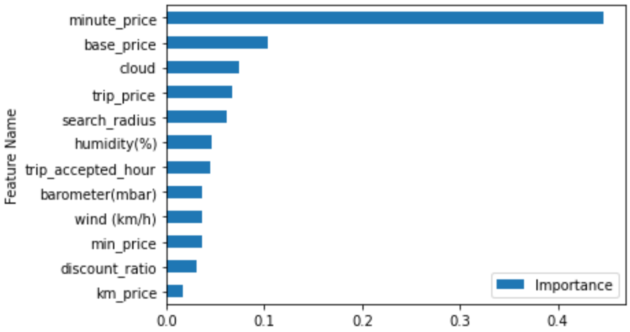


Fig.10. Features influencing the surge rate

In the premise, this paper shows the application of ML algorithms to predict the surge rates or surge multipliers of Yassir, selection of the best ML algorithms for classification and regression from 55 different ML algorithms, the effect of weather, and the effect of several other factors on surge rate.

Acknowledgement

Special thanks to Yassir for sharing real-time data.

References

1. Battifaranoa, M. and Qian, Z.: Predicting real-time surge pricing of ride-sourcing companies, Transportation Research Part C: Emerging Technologies, 107, 444-462, (2019).
2. Belli, Z., and Ali Djoudi, M.: Startup companies in tourism sector– Wijha company as a model, Journal of Business and Trade Economics, 8(2), 378-395, (2023).
3. Castillo*y,* J. C., Knoepfle*z,* D. and Weyl, E. G.: Surge Pricing Solves the Wild Goose Chase, In: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 241-242, Association for Computing Machinery, Cambridge Massachusetts USA (2017).
4. Elliott,G., Gargano, A.,Timmermann, A.: Complete subset regressions with large-dimensional sets of predictors, Journal of Economic Dynamics and Control,54,86-110 (2015).
5. Geohash Homepage, http://geohash.org/, last accessed 19/9/2024.
6. Karamanis, R., Anastasiadis, E., Angeloudis, P., and Stettler, M.: Assignment and Pricing of Shared Rides in Ride-Sourcing using Combinatorial Double Auctions, IEEE Transactions on Intelligent Transportation Systems, 22(9),5648-5659 (2021).
7. Ke, J., Zheng, H., Yang, H. and Chen, X. M.: Short-term forecasting of passenger demand under ondemand ride services: A spatio-temporal deep learning approach, Transportation Research Part C: Emerging Technologies**,** 85(October), 591–608 (2017).
8. Laptev, N., Yosinski,J., Li,L.E., and Smyl,S.: Time-series Extreme Event Forecasting with Neural Networks at Uber, In: Proceedings of ICML, (2017).
9. Li, Y., Wang, X., Sun, S., Ma, X. and Lu, G.: Forecasting short-term subway passenger flow under special events scenarios using multiscale radial basis function networks, Transportation Research Part C: Emerging Technologies*,77*, 306–328 (2017).
10. Nadia, S.: Revolutionizing Transportation: Exploring Yassir App's Features, Challenges, and User Insights, El-Manhel Economy,7(1),1095-1110 (2024).
11. Sihem, M.: Contribution of Emerging Technologies (Artificial Intelligence, Big Data) in the Development of Intelligent Public Transportation Systems - A Case Study of YASSIR Company in Algeria, Journal of contemporary economic research, 6(2),453-476(2024).
12. Silveira-Santos, T., Papanikolaou, A., Rangel, T., Manuel Vassallo, J.: Understanding and Predicting Ride-Hailing Fares in Madrid: A Combination of Supervised and Unsupervised Techniques, Applied Sciences, *13*, 5147-5163 (2023).
13. Wei, Y. and Chen, M. C.: Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks, Transportation Research Part C: Emerging Technologies 21(1), 148–162 (2012).

14. Wiki Geohash, https://en.wikipedia.org/wiki/Geohash , last accessed 19/9/2024.
15. Zhu, L., and Laptev, N.: In: Proceedings of 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE, New Orleans, LA, USA (2017).