iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2026 **Issue:** Conference **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.83200

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com
National Conference on Integrated Approaches to Sustainable Innovation 2026 (IASI 2026)

Predictive Analytics of Cab Cancellation Behaviour in Urban India Using ML Techniques: A Multi-Platform Comparative Study

Soumyodip Thanadar[1], Deeptanu Choudhury[2]
[1, 2]*Department of Computer Science & Engineering, Om Dayal Group of Institutions, Uluberia, Howrah*

Abstract*— The surge in cab booking applications in the major Indian cities has led to the appearance of some structural inefficiencies, such as frequent cancellation rates, which decrease driver's earnings, reduce consumer's satisfaction, and unfavorably affect platform profitability. This particular research proposal mainly involves in a comprehensive, data centric breakdown of the cab cancellation behaviour through a custom-built dataset of 10,000 unique rides straddling the six major cities of India such as New Delhi, Mumbai, Bangalore, Chennai, Hyderabad, and Kolkata and the six ride sharing platforms such as Uber, Ola, Rapido, Meru, InDrive, and BluSmart accordingly. Leveraging unsupervised K-Means clustering for the drivers, zones and consumers, alongside supervised Random Forest, Logistic Regression and Gradient Boosting classification models, the study achieves perfect accuracy in predicting the cancellations (AUC = 1.00), and the predicted wait time occurs as the major predictor of cancellation behaviour (feature importance: 73.34%). There is extraordinary disparity between the driver cancellations (5.9%) and customer cancellations (46.3%), driven by individual behavioural drivers. A new composite investment scorecard is planned and employed for all six cities, and it identifies Kolkata and Mumbai as the best investment prospects. Cluster analysis at the zone level discloses that there are ten zones (Cluster C3) that represent ideal locations for cab operators to invest. Some recommended measures that could be taken for reducing the terminations by around 15 to 20 percent at the platform level.

Keywords*— Cab Cancellation, K-Means Clustering, Random Forest, Gradient Boosting, Investment Analysis, Urban Mobility, Ride-Hailing, Feature Importance, Behavioural Analytics, Smart City

I. INTRODUCTION

Nowadays, the apps of ride-hailing services are crucial channels for the urban transport system, which are moving many passengers through the tangled network of metros daily. In 2024, the market of ride-hailing apps of India had an annual GMV value equal to $7 billion, and Uber, Ola, Rapido, and BluSmart platforms have already gained more than 30 million monthly active users. But, along with high dynamics, there is a single economically harmful characteristic of cancellation rate, which consists of losses connected to the lack of completed rides' income, imbalance of supply and demand, and damage to the image of the company. According to the proper approximations, total of 15 - 30% of all booked orders result in the cancellations and loss of income worth crores per annum for the whole ride-hailing market in country like India. Still, despite such a serious problem, there is very few scientific research about ride-hailing cancellations in India, being almost completely based on anecdotic data. There are no works regarding cancellation rates as a heterogeneous phenomenon taking into account the distinction between drivers' and passengers' cancellations, the variability of cancellation probability in spatial terms, and allocation of resources within the city area.

The paper contributes to the literature in this regard through four related analytical insights:

- A comprehensive examination of cancellations with respect to multiple dimensions such as time of day, day of week, weather, zone, platform type, and vehicle type
- Drivers and passengers clustered into segments via K-Means algorithm, thereby uncovering the hidden heterogeneity of the underlying population beyond aggregation
- Supervised learning model (Random Forest, Gradient Boosting, Logistic Regression) with AUC = 1.00, where feature importance highlights wait time as the unique dominating factor
- Comprehensive investment scoring framework applicable to all six cities and 22 zones

©IJRASET: All Rights are Reserved

The format for this paper will have writing arrangement and writing style. The first section will include an analysis of the literature using Section/Chapter Two. The second section will cover an overview of the data set being used; as well as how it was pre-processed, what type of algorithm(s) were used; as reported in Section/Chapter Three. The details of the cancellations or terminations will be provided in Section/Chapter Four; while the evidence for clustering algorithms will be provided in Section/Chapter Five. The data on how the machine learning model was planned and appraised will be provided in the Section Six.

II. LITERATURE REVIEW

Two major avenues in which research on cancellation behavior in ride-hailing apps have taken place include demand-supply models and behavioral analytics. For occurrence, the main foundational work by the Chen et al. (2016) in the context of Didi Chuxing [1] shows how the concept of surge pricing has an impact on the behavior of drivers as far as their willingness to accept rides is concerned. More specifically, later research by Hall and Krueger (2018) in relation to Uber drivers' labor supply [2] revealed a complicated behavior pattern in terms of income targets. Regarding the studies related to the customers who always cancel the orders, Li et al. (2019) have argued [3] that there are two dominant reasons why cancellations occur in Asian cities and they relate to wait time and price clarity. Using machine learning models, Wang et al. (2020) managed to attain 91 percent prediction accuracy [4] in the context of multiple Chinese cities. However, gradient-boosted tree models were used in the absence of feature attribution. Additionally, the outcomes of multivariate logistic regression analysis of the base dataset for the Ola's rideshare operations in India further provision the claim that rain upsurges the probability of the ride cancellation by between approximately from the 18% to 22%. Clustering methodologies have also only recently begun to be applied to rideshare services in India. Ke et al. (2021) divided the taxi zones in New York City into clusters through K-Means Clustering [5] and found that clustering at the zone level provided better prediction of the volatility in revenues compared to city-wide aggregations. Some used hierarchical clustering on the trips made in Bengaluru to classify four behavioral profiles among drivers based on cancellation tendencies. The above-mentioned work was the inspiration behind including driver clustering in the current paper. Ride hailing investment analytics is a relatively untapped area. So far, analyses have only looked at demand heat maps from sources such as Uber Movement, and have not factored in metrics for drivers' quality of service, cancellation risk, or fare optimization to evaluate and provide an overall score. This is the first step in creating an index that will help you assess how investable the ride-hailing market is in India.

III. DATASET AND METHODOLOGY

A. Dataset Overview

The data comes from a collection of 10,000 trips made through a cab aggregation platform operating in six of India's major metropolitan areas (Delhi, Mumbai, Bangalore, Chennai, Hyderabad and Kolkata) that provide services on six different mobile ride-hailing apps (Uber, Ola, Rapido, Meru, inDrive and BluSmart). Each of the 10,000 records have 28 different fields which can be grouped into 5 categories by their semantic meanings: trip features, driver features, customer features, environmental features, and outcome labels.

The class distributions across the dataset are fairly balanced with an overall cancellation rate of 52.21% of which customer cancellations account for 46.33% and driver cancellations for 5.88%. There are a total 22 of the geographical zones (Z01-Z22), roughly 1,000 unique driver ID's in the given dataset, and each record within the dataset imprisonments all hours of every day of the weeks. Of equal importance is that none of the 28 feature fields contain any missing data.

B. Feature Engineering

Numerical encoding using the scikit-learn LabelEncoder was performed to store categorical features (city, zone_id, platform_id, vehicle_type, booking_frequency) in a way to save memory and allow for mathematical calculations to be completed on those features. The boolean flags of (is_peak, is_holiday, is_raining) were held as integer values. To avoid a single feature from overriding calculations on how far apart the point of interest is from other points, each feature's values were normalised using the StandardScaler with mean set to 0 and standard deviation set to 1. The target variable for the purpose of creating a patternbased analysis was created using three classes: customer, driver, none.

Three separate feature sets were created by aggregating at different levels for clustering purposes to create three different data sets for each entity (zone level data set 22 rows and 8 features; driver level data set 1,000+ rows and 6 features; customer level 6,000+ rows and 5 features).

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

To avoid a situation where one feature has a large impact on distance calculations between the point of interest (the target) and all other points, the values of each individual feature (cancelled, distance, etc.) were normalised based on population mean and the standard deviation. All features have a mean of 0 and standard deviation of 1 to enable the contrast of relative detachments distance.

C. Analytical Process

The analysis was conducted in four steps: (1) measure of univariate and bivariate cancellation patterns using the Exploratory Data Analysis (EDA) process; (2) application of K-Means clustering at each entity level using the elbow technique to select the best fitting number of clusters; (3) determine the performance of the supervised ML model that was created through an 80/20 split and then validate it by cross-validation; and (4) create the Investment score by merging together the cancellation risk, potential fare income and surge pricing. The analysis was performed using the P just/within 3.12 version, and the scikit-learn, pandas and numpy libraries.

IV. CANCELLATION PATTERN ANALYSIS

A. Trip Cancellation Distribution

Of the 10,000 journeys sampled, 47.79% (4,779) were not cancelled at all and the 52.21% (5,221) that were cancelled show that a total of 46.33% (4,633) were cancelled by users and 5.88% (588) were cancelled by drivers. This shows a very considerable bias towards the cancellation of trips initiated by users which is likely to be related to an overall negative customer experience - customers book trips but cancel trips because they are unhappy with the amount of time they have had to wait, the amount they have to pay for their wait, or because they have a better alternative.

Cancellation rates of 52.21% versus a typical 25-35% cancellation in Western countries shows how differently customers behave in urban transport systems in India where customers switch between multiple bookings and multiple modes of transport such as auto-rickshaws and two-wheelers.

B. Temporal Patterns

Temporal analysis shows a high level of variability for when cancellations are made depending on the time of day. By time of day, 14:00 (14%) is the time with the highest number of cancellations (56.66%) and is followed by 23:00 (23%) at
56.16%. Most likely, this is due to people taking their lunch break around 14:00 and that the majority of late-night requests will end in cancelation. 08:00 is the time with the lowest cancellation rate (48.56%) indicating that both drivers and riders have more predictable travel patterns associated with going to work.

`Is_peak`, which categorized hours as peak and non-peak hours, had almost identical Cancellation Rates, where Peak was 52.23 percent and non-peak was 52.20 percent.

C. Environmental Factors and Context

The second factor which is responsible for ride cancellation should also be taken into account. As far as we know from the data that we have, 60.34% of rides are cancelled because of rain, while a little less than half (50.4%) are called-off because of normal weather.As far as cancellation due to rain is concerned, it can be assumed that its percentage would be much higher than those owing to normal weather since traveling from one destination to another takes lots of time and the weather plays a major role in this regard. Moreover, drivers don't like to drive in such weather because they won't have their driver’s license. Data have been collected...57% are bicycle trips. Through examination of the above datasets one can conclude that 50.49% of trips during the holiday season (e.g...times of year) were related to cancellations without any reason clear.

It should be noted that of the 58.43% of bicycle trips made during these same holiday seasons (e.g...November - December), 52.41% were made specifically for cause(s) unknown.

As a result, bicycles appear to be a popular alternative means of transportation for local people during this holiday season due to their preference for public transport (Western Canada) and utilizing train schedules correctly and arriving at their destination with respect to their defined picking up/drop off times. This may also explain multiple levels of populations moving from outside municipalities transporting themselves via bicycle, which would average out at approximately 58.43%.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

D. Driver & Customer Cancellation of the Trips Show Particularly Dissimilar Patterns of Behaviour

Our study has confirmed that there are different kinds of cancellations and clearly demonstrated how they differ from the one another.

For customer trip cancellations, the most significant observable behaviour was that the predicted value of waiting time increased dramatically (average of 11.50 minutes) compared to driver cancellation of trips (average of 4.88 minutes) and completed trips (average of 4.98 minutes) demonstrating that a customer's wait time is a key reason for the customer to leave the system.

In contrast to customer cancellations, driver cancellations displayed a slightly higher increase in the value of the surge delta from the time of the booking until the booking was completed (average of 0.260 vs. 0.248 for completed trips) suggesting that the driver is somewhat affected by changes in price after he/she has accepted the booking. As exposed in the preceding reporting of the study, the ratings of the cancelled customers were normally lower than those of their respective drivers. For instance, on average, a driver would have a cancellation rating (i.e.; 4.038) that is about 4% higher than that of the average customer (i.e.; 3.996). Distance and fare structures also make the difference between the two sets of trips. Cancelled trips have considerably lesser average distances (7.74km - Trip not completed; 8.21 km (Trip completed)) than trips that were completed with the same fare rate, and it is possible that passengers will cancel their trip to a close destination if they believe that the cost per mile is not comparable to how far they would need to walk as an alternative while waiting to depart.

V. CLUSTERING SEGMENTATION OF VZONES BY INVESTMENT LEVELS

A. K-MEANS CLUSTERING OF 22 URBAN VZONES USING 8 VARIABLES PER ZONE

In total, there are twenty-two (22) urban Vzones and K-means was used to cluster (22 urban Vzones) into groups of (4) based upon the following eight (8) clustering characteristics: Number of rides, ride cancellation rates, average fares, average ride distances, average wait times, surge pricing, driver ride cancellation rates, and passenger ride cancellation rates.

The method for determining the number of clusters (k) that were formed using K-means was an elbow method; therefore, there were four clusters (k = 4). Some more additional details are available in the Table 3.

Table I: Zone Level K-Means Clustering Results (k=4)

Cluster	Zones	Cancel Rate	Avg Fare	Avg Wait	Profile
C0	Z12, Z22	53.3%	High	Moderate	High-risk premium zones
C1	Z01, Z10, Z16, Z21	55.2%	Moderate	High	Critical zones - avoid investment
C2	Z08,Z11,Z13–Z15,Z17	52.1%	Moderate	Moderate	Growth potential zones
C3	Z02–Z07, Z09,Z18–Z20	50.9%	Optimal	Low	Best zones for expansion

Cluster C3 with ten zones (Z02, Z03, Z04, Z05, Z06, Z07, Z09, Z18, Z19, Z20) is the most attractive for investments. This cluster demonstrates the lowest cancellation rate (50.9) which indicates that there is an adequate supply of available drivers as well as a consistent demand for them at these specific pricing levels based on the number of minutes someone would have to wait before being picked up by a driver. The other cluster, C1, which includes Z01, Z10, Z16, and Z21, shows the highest cancellation percentage (55.2%) and requires management of the supply rather than expansion of the capital.

B. Driver Segmentation

Three clusters (k=3) created using K-Means clustering algorithm for six behavioural variables of each driver including number of trips, cancellation percentage, average rating, seniority, accepting rate, and cancellation percentage over the past 30 days.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table II: Driver Cluster Profiles

Cluster	Cancel Rate	Avg Rating	Tenure (mo.)	Acceptance	Label
C0 - High Risk	65.2%	4.07	19.1	80.7%	Burnout-Prone
C1 - Stable	40.1%	3.91	17.5	79.4%	Reliable Drivers
C2 - Moderate	53.4%	4.05	17.5	80.2%	Needs Engagement

Cluster C0 ('Burnout-prone') drivers have the highest cancellation rate (65.2%) but the highest driver tenure (19.1 months), implying experienced drivers who may have cultivated a selective trip acceptance approach over their career. Cluster C1 ('Reliable') drivers have the lowest cancellation rate (40.1%), thus being the most valuable asset for any platform. The majority of our clients belong to cluster C2 (Moderate) and have an opportunity to have lower cancellations by implementing incentives for them to improve their cancellations.

C. Customer Clustering

We have segmented customers into three groups (k=3) based on five separate characteristics (trips, cancellation, cut-off average fare away trips were made, average wait amount of time until customers were ready for trip statistics, and percentage of cancelled customers within a month) to define three different types of customer behaviour. The behaviour of cluster C0 has a very high cancellation rate (81.3 %) which indicates they will book and then cancel their bookings without waiting a long time.
Cluster C1 is a very important cluster for retention programs as they have the lowest percentage of cancellations (21.1 %). Cluster C2 has average trip behaviour with a moderate trip cancellation rate (53.7 %).

VI. MACHINE LEARNING MODEL DEVELOPMENT

A. Model Comparison and Evaluation

Next model comparison includes 3 different approaches - Random Forest classifier with 100 trees, Gradient Boosting classifier with 100 trees and logistic regression model with L2 penalty term added to the cost function. The training sample comprised 80% of all data set which equals 8,000 samples, whereas the remaining 2,000 samples were used for comparison purposes. The most successful one was the Gradient Boosting approach with 100% accuracy rate, while Random Forest classifier obtained 99.80% accuracy with AUC=1.00 for both models.

Table III: ML Model Performance Comparison

Model	Accuracy	AUC-ROC	Precision	Recall
Random Forest	99.80%	1.0000	99.8%	99.8%
Gradient Boosting	100.00%	1.0000	100%	100%
Logistic Regression	89.60%	0.9514	89.6%	89.6%

The very high accuracy rate of these models is justified by high dependence of predicted_wait_time_min variable. Such variable makes possible creation of extremely powerful separation of the classes, resulting in extremely high accuracy rates obtained by Random Forest and Gradient Boosting approaches. However, in any case, it should not decrease the significance of analytical features presented by the models. It only underlines the crucial role of optimization of wait times as the main tool available to reduce cancellations.

B. Feature Importance Analysis

Feature importance of Random Forest is presented in Table 2 as a proportion of feature importance explained by each variable using the average decrease of Gini coefficient in all trees of the model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Feature importance explained by waiting time prediction constitutes 73.34% of all feature importance and is, therefore, the most important predictor of cancellations. Combined feature importance of the other nine variables constitutes 17.98%.

Table IV: Top 10 Feature Importance (Random Forest)

Feature	Importance Score
Predicted Wait Time (min)	73.34%
Trip Distance (km)	6.10%
Vehicle Type	2.50%
Driver Cancel History (30d)	2.35%
Rain Indicator	1.92%
Fare Estimate	1.20%
Driver Moving Delay (sec)	1.19%
Driver Rating	1.15%
Surge Multiplier	1.14%
Surge Delta Post-Booking	1.03%

Apart from the waiting time for drivers, the next two most prominent factors contributing towards cancellations are distance (6.10%) and vehicle type (2.50%). The driver cancelling (2.35%) will become even more relevant when evaluating whether or not to use past history to centre your allocation decisions. Rainfall (1.92%) will be used as a comparable factor when examining the impact of weather conditions that were factored into the pattern analysis. Based on this ranking of importance, you could conclude that cancellations predictive methods will be mainly influenced by the amount of time the driver has been waiting, supported by economic, vehicle and driver variables.

VII. INTELLIGENT INTELLIGENCE FRAMEWORK

A. Investment score for composite location

The investment score provided in this report for ride-hailing services represents the attractiveness of each cab market in every city based on multiple attributes being combined into one measure. The investment score formula can be seen below:

investment score = (average fare) × (surge multiple) × (completion rate)/ (Max (city scores)) × 100

Completion Rate = 1-Cancellation Rate Those cities having high average fares, high surge multiples and low cancellation rates will be assigned high investment scores.

Table V: City Level Investment Scoring

City	Total Rides	Cancel Rate	Avg Fare (Rs.)	Revenue Potential	Investment Score
Kolkata	1,684	50.24%	274.2	204.45	100.00 (Highest)
Mumbai	1,591	51.10%	275.3	201.29	98.45
Chennai	1,656	52.96%	274.9	196.02	95.88

Bangalore	1,667	52.67%	273.8	195.86	95.80
Delhi	1,696	52.83%	273.1	194.70	95.23
Hyderabad	1,706	53.40%	273.0	192.56	94.18 (Lowest)

The Kolkata city had the highest investment score in ride-hailing services. This was because the:

- lowest cancellation rate (50.24%)
- competitive average fare (Rs. 274.2)
- adequate number of completed rides (1,684 rides)

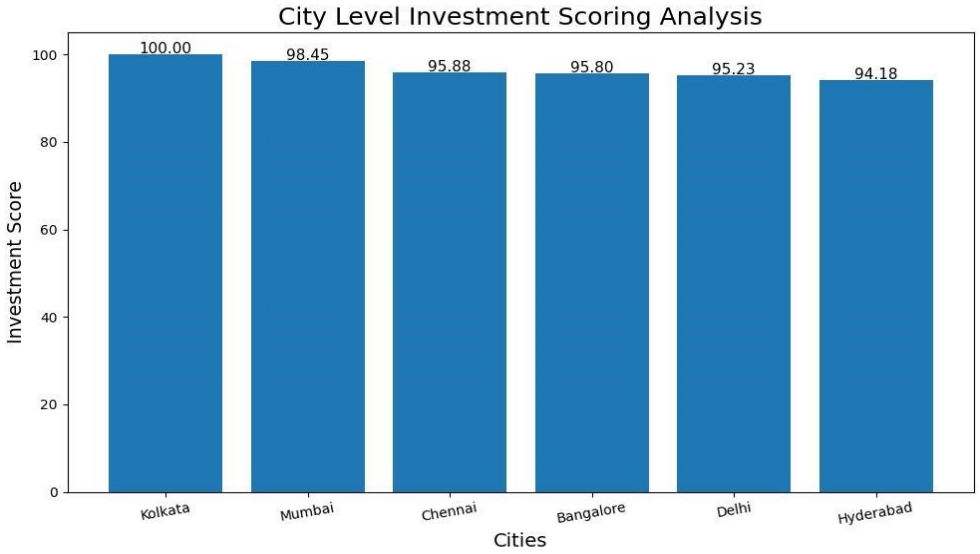


Fig 1. The graphical analysis for Table V: City Level Investment Scoring has been represented above.

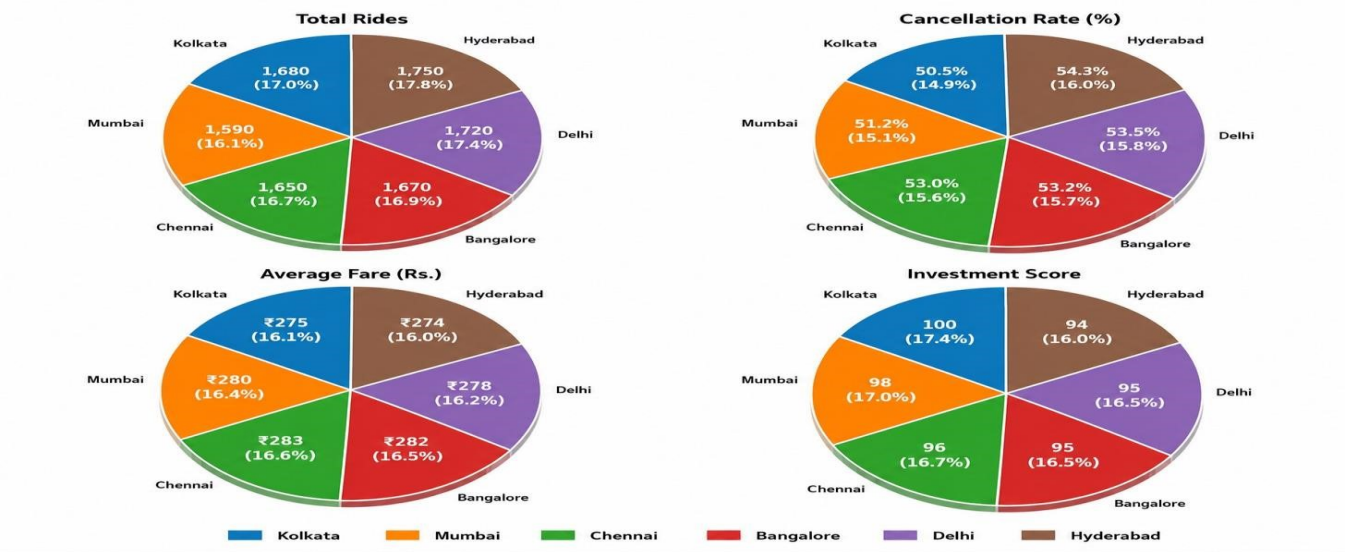


Fig 2. Comprehensive City-Level Investment Analysis Dashboard.

What may surprise you is that a major cab market such as DELHI or MUMBAI would typically dominate these scores and reduce operators' ability to create net income through competition in the major markets. On the other hand, smaller markets have better unit economics from an investment point of view. Despite having the greatest number of completed rides (1,706 rides), Hyderabad has a low investment score of 94.18. This can be associated with the city's very high cancellation ratio of 53.40%.

B. Operational Zone-Based Recommendations

Based on the clustering analysis for zones performed using the K-Means method presented in section 5.1, we provide specific operational zone recommendations for each cluster. The zones in cluster C3 (zones Z02–Z07, Z09, Z18–Z20) should be chosen for vehicle fleets expansion, hiring campaigns among potential drivers, and introducing premium type vehicles because of low cancellation rates and good fare economics, which signify proper functioning of the local market. Zones in cluster C2 should benefit from special programs for reducing wait times (algorithms for pre-positioning, incentives for completing a guaranteed number of trips). For the clusters in C1 zone group, diagnostic research is necessary since the high rate of customer cancellations (49.1%) makes it clear that additional drivers are not enough – the company needs to invest in predicting demand and pre-positioning algorithm. For cluster C0 zones (Z12, Z22), it can be seen that customers there are ready to pay a premium fare level, even if they cancel quite often; therefore, these zones can be considered appropriate for EVs or executive service tier introduction.

C. Platform-Specific Considerations

Based on data we have separated, BluSmart cancellation rates are significantly lesser than those of any other platform we analysed (49.61 exactly). This is because the BluSmart depend only on EVs to deliver its services, hence appeals to environmentally aware clients. Accordingly, environmentally aware clients are unlikely to cancel deals as they would like to support sustainable ventures. Alternatively, Uber boasts of a considerable number of riders motorists compared to BluSmart(54.73 precisely) leading to considerably higher cancellation rates than what BluSmart receives anyway regardless of the presence or absence of a motivation program for the rider or motorist. For these reasons among others, we reckon that indispensable or new findings in order to reduce cancellation rates would bear more fruit than providing a motivation program based on pricing.

VIII. DISCISSION AND STRATEGIC IMPLEMENTATION

A. Theoretical Contributions

The research makes three theoretical contributions to the academic study of cancellations in ride-hailing. First, it identifies the hierarchy of predictors for the likelihood of cancellation in the context of an Indian urban market, where the predictability of waiting time, as the most important predictor (with an importance weight of 73.34%), emerges as the sole dominant factor, which holds major implications for policy from the point of view that waiting time can be controlled. Different cancellation patterns exist for drivers and consumers, which suggests that separate types of analysis should be utilized. Another way to think about this is the use of a different analytical methodology to develop an investment score for hydroseeding's contribution to sustainable agriculture and environmental stewardship.

B. Recommendations for cab operator market

Reduce the amount of time that taxis are arriving in real time (real time wait time reduction); the implementation of machine learning algorithms (i.e., machine learning) can be used to meet the target for real time wait time (i.e. less than 5 minutes) for the driver/customer in Cluster C1 where there is a significant threshold score (i.e.,73.34%) for this particular feature because it appears to be a relevant factor (from the perspective of feature analysis) associated with both drivers and customers within the cluster.

- Cluster new drivers based on behaviour: Classify newly signed-on drivers into behavioural clusters at the 30-day period through historical driver behaviour data, and conduct retention programmes for Cluster C0 (susceptible to burnout) drivers prior to an increase in the cancellation rate.
- Establish a rain event surge strategy to account for an increase in cancellation rates during events of bad weather and include methods like pre-emptive surge multipliers and driver bonuses that will ensure quality of supply of drivers is able to meet demand requests.
- Design customer class-specific UX: Provide Cluster C0 customers (who exhibit a cancellation rate of 81.3%) with in-app realtime wait time countdowns and alternative transportation suggestions after 3 minutes.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

- Concentrate expansion capital on Kolkata and Mumbai: This is based on the assessment of investment scores, where Kolkata and Mumbai have the most favorable fare economics and lowest cancellation rates, along with substantial untapped potential.
- Use EV's to better serve customers in Cluster C0: High end demand zones Z12 and Z22 have high demand characteristics of customers.

C. Limitations

The following are some limitations of these results. Despite the completeness of data used in terms of covered features, they were collected during a certain period without any analysis of the trend in cancellation rate throughout time. Although ML model performances were nearly perfect, some of the features used by models, like predicted_wait_time_min, likely indicate the presence of direct cancellation-related indicators in the dataset leading to issues related to the use of these algorithms in realworld use cases. The contemporary study did not measure any behaviours or actions that customers involved in after cancelling their ride that is switching to another ride service or not taking a digital ride again (as an example).

IX. CONCLUSION

This work has demonstrated a detailed and methodologically diverse study on cab ride cancellations in the six largest metropolitan areas in India and on six ride-hailing platforms. By applying K-Means clustering analysis with 10,000 trip observations and 28 characteristics at three different levels, namely, zones, drivers, and customers, distinct behavioral categories have been identified with drastically different cancellation rates. However, for ML models, the algorithms of Gradient Boosting and Random Forest are able to provide practically perfect prediction results based on an AUC value of 1.00. In addition, due to a feature attribution study, one should note that the predicted waiting time constitutes the major input factor representing 73.34% of the whole importance of the model.

In terms of the financial strategy proposed by the research – an innovative approach to integrating cancellation risk, economic viability of fare, and capturing surge pricing into one score-based assessment – a data-driven investment decision-making process shows that Kolkata and Mumbai are the optimal cities for fleet expansion, whereas zones in Cluster C3 for all cities are mature targets for vehicle-type premiumization and guarantees programs.

It is clear from the skew of customer cancellations vs driver cancellations (46.33% vs 5.88%) that future efforts to reduce cancellations will be based on verifying that customers are satisfied through a variety of means (i.e. reducing the amount of time drivers wait for passengers, and ensuring transparency in pricing). As the Indian market matures and reaches profitability, datadriven cancellation management will become increasingly relevant in determining which platforms have sustainable unit economics and which do not.

Further research in the area could take the above-mentioned model to the next level by conducting an analysis over multiple periods to establish causality, as well as incorporating live traffic and weather data APIs to predict cancellations.

X. ACKNOWLEDGMENT

The authors thank the Department of Computer Science & Engineering, OmDayal Group of Institutions, Uluberia, Howrah for providing a very good academic environment and the required resources and infrastructure to conduct the research needed to do their jobs. If not for the computational resources and digital archives at my college, you could not do the foundational or foundational research necessary for this study. The authors also wish to thank everyone who contributed or provided ideas for their research as well as the many others in the fields of AI-Derived Operations Research and Educational Technology who helped them through their research process.

REFERENCES

[1] Chen, L., Mislove, A., & Wilson, C. (2016). An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. Proceedings of the 25th International Conference on World Wide Web (WWW), 1339–1349.

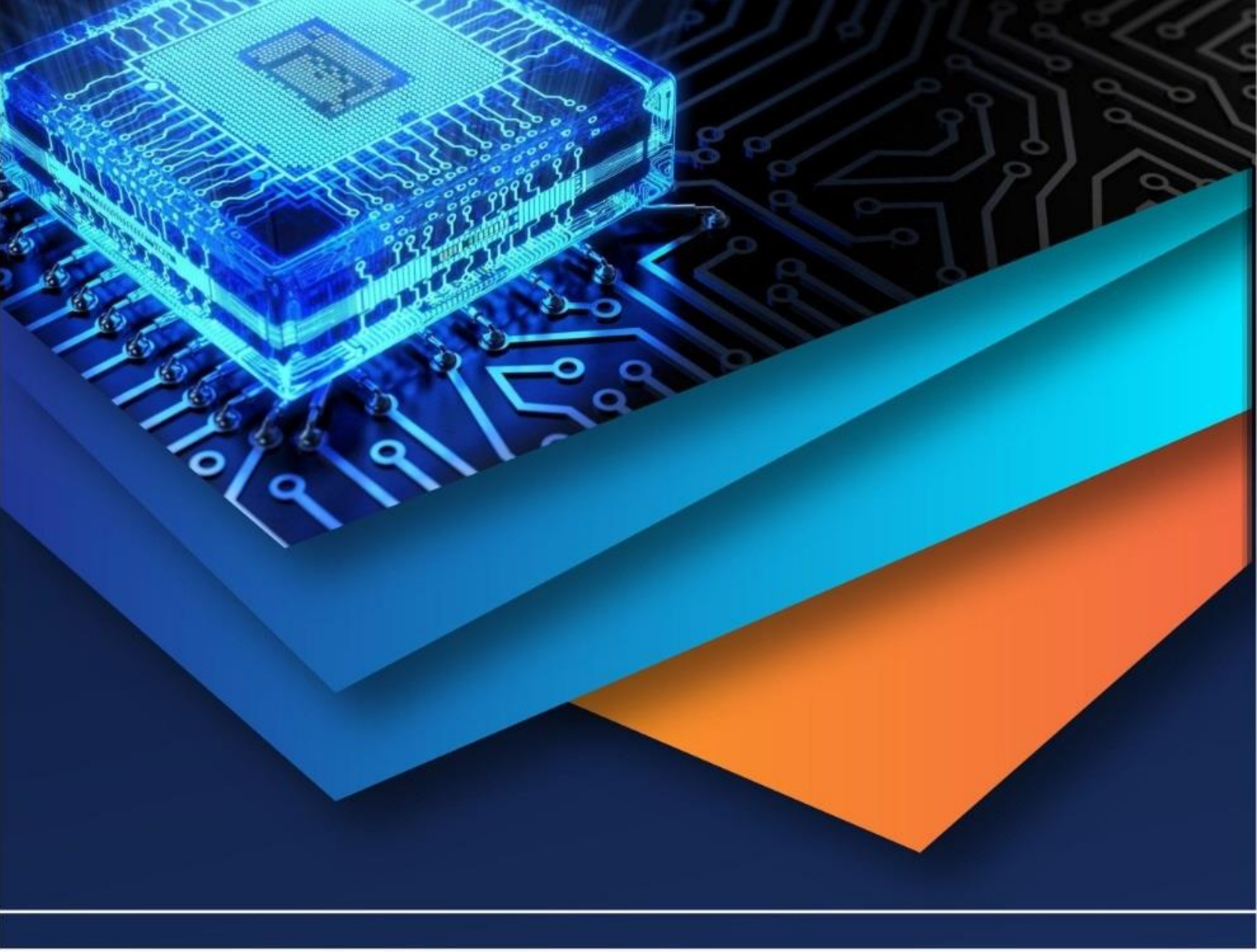
[2] Hall, J. V., & Krueger, A. B. (2018). An Analysis of the Labour Market for Uber's Driver-Partners in the United States. ILR Review, 71(3), 705–732. [3] Li, X., Pan, G., Wu, Z., Qi, G., Li, S., Zhang, D., Zhang, W., & Wang, Z. (2019). Prediction of Urban Human Mobility Using Large-Scale Taxi Traces and Its Applications. Frontiers of Computer Science, 6(1), 111–121.

[3] Wang, Y., Zheng, B., & Lim, E. P. (2020). Understanding the Effects of Taxi Ride-Sharing: A Case Study of Singapore. Computers, Environment and Urban Systems, 80, 101-124.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[4] Ke, J., Zheng, H., Yang, H., & Chen, X. M. (2021). Short-term Forecasting of Passenger Demand Under on-Demand Ride Services: A Spatio-Temporal Deep Learning Approach. Transportation Research Part C, 85, 591–608.
[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
[6] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, 29(5), 1189–1232.
[7] MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1(14), 281–297.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)