OLA - Ensemble Learning Business Case

❖ Topic: Ensemble Learning in Driver Attrition Prediction
❖ Duration: 1 week

Why this case study?

From the company’s perspective:

- Ola is a leading ride-sharing platform, aiming to provide reliable, affordable, and convenient urban transportation for everyone.
- The constant challenge Ola faces is the churn rate of its drivers. Ensuring driver loyalty and reducing attrition are crucial to the company's operation.
- Analyzing driver data can reveal patterns in driver behavior, performance, and satisfaction. This would help in foreseeing potential churn, allowing proactive measures.
- By leveraging data science and ensemble learning, Ola can predict driver churn, which would be pivotal in its driver retention strategy.

From the learner’s perspective:

- Engaging with this case offers a practical understanding of how ride-sharing giants operate and the challenges they face.
- Ensemble learning, comprising various algorithms and techniques, can significantly improve prediction accuracy over standalone methods.

- This exercise hones the participant's skills in imputation, dealing with class imbalances, feature engineering, and model evaluation.

- Additionally, learners gain hands-on experience in tackling real-world problems, transforming raw data into actionable insights that can guide business strategies.

Dataset Explanation: ola_driver.csv

1. MMMM-YY: Reporting month and year.
2. Driver_ID: A unique identifier for every driver.
3. Age: Age of the driver.
4. Gender: Driver's gender. Male: 0, Female: 1.
5. City: City code representing the city the driver operates in.
6. Education_Level: Education level of the driver, categorized into 0 for 10+, 1 for 12+, and 2 for graduate.
7. Income: Average monthly income of the driver.
8. Date Of Joining: The date when the driver joined Ola.
9. LastWorkingDate: The most recent or final day the driver worked with Ola.
10. Joining Designation: Designation of the driver at the onset of their journey with Ola.
11. Grade: A grade assigned to the driver at the reporting time, likely denoting performance or other metrics.
12. Total Business Value: The total monetary value (business) a driver brings in a month. Negative values might indicate cancellations, refunds, or other financial adjustments.
13. Quarterly Rating: Rating assigned to drivers on a quarterly basis. Ratings range from 1 to 5, with 5 being the best.

What is Expected?

Assuming you are a data scientist at Ola, you are entrusted with the responsibility of

analyzing the dataset to predict driver attrition. Your primary goal is to utilize ensemble learning techniques, evaluate the performance of your models, and provide actionable insights to reduce driver churn.

Submission Process:

Upon concluding the case study...

- Document your insights, methodology, and results in a Jupyter Notebook.
- In your notebook, ensure you:
    - Showcase the Python code for all data analysis, feature engineering, model building, and performance evaluation.
    - Incorporate visualizations like bar plots, correlation heatmaps, ROC AUC curves, and more, supporting your analysis.
    - Conclude with significant insights derived from the dataset and propose actionable recommendations for Ola to enhance driver retention strategies.
- Transform your Jupyter Notebook into a PDF (Utilize the Chrome browser's Print command for this).
- Adhere to the submission guidelines and upload the PDF on the specified platform.
- Note that once your work is submitted, there's no provision to edit or modify your submission.

General Guidelines:

This scenario mirrors real-world challenges and embodies the tasks data scientists frequently grapple with. Embrace this opportunity to dive deep and simulate a professional experience.

During the course of this study, it's possible to face hurdles or even feel daunted:

- Re-evaluate the problem statement periodically to assure alignment with objectives.
- Deconstruct multifaceted tasks into simpler, achievable steps.
- If faced with code errors or issues, turn to online forums or official documentation. Problem-solving acumen is indispensable for data scientists.
- Collaborate with colleagues. Engaging in the discussion forum can offer diverse perspectives, aiding in overcoming obstacles or sparking new ideas.
- Revisit lectures or explore external resources for topics you're unsure about.

- For any overarching issues or if the problem statement appears ambiguous, don't hesitate to contact your Instructor.

Remember, every challenge faced is an opportunity to grow. Approach this case with enthusiasm, diligence, and an open mind.

What does 'good' look like?

1. Define Problem Statement and perform Exploratory Data Analysis

	Hint	Approach
a. Definition of problem	Start by crystallizing the problem statement. What's the objective of Ola? Why is predicting driver attrition so important?	The main aim is to predict potential driver churn using multiple attributes to maintain a consistent driver base and ensure business continuity.
b. Observations on Data	A thorough understanding of the dataset structure is key. Observe the shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), missing value detection, statistical summary.	a. Use functions like data.info(), data.describe(), and data.shape in Python. Identify numeric versus categorical attributes. Convert categorical data types using astype('category') if needed. b. Since we don't have a pre-defined target variable, we will create one based on certain conditions or behavior observed in the data. c. Aggregating on 'driverid' might be essential to identify consistent behaviors or patterns.
c. Univariate Analysis	Begin the analysis with individual variables. For continuous attributes, use distribution plots, and for categorical ones, use bar	For continuous variables, use histograms or density plots. For categorical variables, use countplots. Tools like Seaborn make these

	or count plots.	visualizations straightforward. This helps in understanding the distribution of individual variables.
d. Bivariate Analysis	Dive into relationships between two variables. (Relationships between important variable)	Employ scatter plots for continuous-continuous relationships, boxplots for categorical-continuous correlations, and crosstab or stacked bar plots for categorical-categorical correlations. For instance, understanding how a driver's ratings correlate with the frequency of rides can offer key insights.
e. Illustrate the insights based on EDA	Every graph and table should deliver an insight	Take notes on surprising distributions, high correlations, or peculiar behaviors seen in the bivariate analysis.
f. Comments on range of attributes, outliers of various attributes	Range and outliers can greatly influence model performance	Use box plots and IQR to detect and comment on outliers. Understand the business context to decide if outliers should be handled or left as they are.
g. Identify normal vs skewed distributions and understand why.	Identify normal vs skewed distributions and understand why.	For continuous variables, comment on the skewness. For relationships, comment on positive or negative correlations, clusters, or other patterns noticed.
h. Comments for each univariate and bivariate plots	Just plotting isn't enough, explain them.	Every visualization should come with a 2-3 line commentary. For instance, "Most drivers have a rating between 4.5 and 5. However, a scatter plot between frequency of rides and ratings shows that higher-rated drivers tend to have more rides." This makes findings more digestible and actionable.

2. Data Preprocessing

	Hint	Approach
a. Duplicate value check	Duplicate rows can skew the results and create redundancy.	Investigate your dataset for any duplicate entries. It might be useful to first examine duplicates based on a subset of features rather than the entire row, as sometimes complete rows might not be identical, but a subset of attributes could still have repeated patterns.
b. Missing value treatment	Missing values can influence model training.	a. Identify columns with missing values. b. Decide the best strategy: imputation using central tendencies, deletion, or more advanced methods depending on the importance and type of the variable. c. More focus on smartly imputing the data.
c. Outlier treatment	Outliers can skew model outcomes.	a. Visualize data for detecting outliers using graphical tools b.Choose an appropriate technique for handling outliers: capping, transformation, or removal. The choice should be backed by a logical reason.
d. Feature engineering	Crafting new variables or modifying existing ones can enhance the model's predictive power.	a. Flag Creation: Attributes such as feedback_count, ride_frequency, etc., might benefit from binary flags under specific conditions. b. Date-related features: Extracting day, month, or year can expose temporal patterns. c. Geolocation Insights: Deriving city or district from GPS data can highlight geographic trends. d. Categorize the "Age" feature into bins such as: 'Young' (18-30), 'Middle-aged' (31-50), 'Senior' (50+). This can help in

		identifying patterns based on age groups. e. Feature Mining from Aggregated Data: ● Derive new features from aggregated data, such as variations in expenses, shifts in ride frequency, or feedback count changes. ● Change in Rating (Example): After aggregating, find the difference between successive quarterly ratings to capture performance trends.
e. Data preparation for modeling	Preparing data in a format suitable for modeling is crucial.	a. Depending on the model's requirements, consider scaling the features. The choice of scaling technique would vary based on data distribution and model sensitivity. b. Different encoding techniques are suitable for different types of categorical variables: I. Label Encoding: Use for ordinal categories with a natural order (e.g., Low, Medium, High). II. One Hot Encoding: Best for nominal categories without inherent order. III. Target Encoding: Good for high cardinality features. Replaces categories with the mean of the target variable for that category. Beware of overfitting; consider regularization or smoothing.
f. Identify normal vs skewed distributions and understand why.	After creating new feature, For themIdentify normal vs skewed distributions and understand why.	For continuous variables, comment on the skewness. For relationships, comment on positive or negative correlations, clusters, or other patterns noticed.

3. Model building

	Hint	Approach
a. Data Splitting	Before model building, split the dataset into training and validation sets.	Typically, a 70-30 or 80-20 split ratio is employed. If the target variable shows class imbalance, consider a stratified split.
b. Addressing Class Imbalance	Imbalanced datasets can lead the model to predict mostly the majority class. This can lead to misleadingly high accuracy but poor generalization.	a. Exploratory Data Analysis (EDA): Check distribution of target variable using plots. b. Oversampling with SMOTE: Use the SMOTE method to generate synthetic samples for the minority class. c. Algorithmic Approach: Use algorithms that can set `**class_weights**`, giving more importance to the minority class.
c. Ensemble Learning: Bagging	Bagging reduces variance by averaging multiple low-bias, high-variance models. Random Forest is a common example.	a. Use BaggingClassifier or RandomForestClassifier from sklearn.ensemble. b. Adjust hyperparameters like the n_estimators, max_depth, max samples, max_features. c. Fit the classifier on training data and evaluate on validation data.
d. Ensemble Learning: Boosting	Boosting builds multiple models in a sequential manner where each new model attempts to correct the errors of the previous ones. Gradient Boosting, AdaBoost, XGBoost are examples.	a. Start with 'GradientBoostingClassifier' from 'sklearn.ensemble'. b. For advanced models, consider using XGBoost or LightGBM. c. Tune hyperparameters like learning_rate, n_estimators, and tree-specific parameters. d. Evaluate model on validation set to check its performance.
e. Feature Importance with Boosting Algorithms	Boosting algorithms like XGBoost can provide feature importance scores which can be beneficial in understanding which predictors have the most influence.	a. Once model is trained, extract feature importance. b. Plot feature importances in descending order. c. Analyze top features and their

		influence on the model's predictions.

4. Results Interpretation & Stakeholder Presentation

	Hint	Approach
a. Understand the Business Context	The primary concern for Ola is to ensure a quality driving experience for its users and retaining efficient drivers.	a. Understand Ola's objectives in evaluating its drivers: Are they looking to reward top performers? Address underperformers? Improve customer satisfaction? b. Recognize the challenges faced by drivers and the factors affecting their performance.
b. Interpreting Model Coefficients	Understanding which features have the most impact on driver performance can help Ola in strategizing training and incentives.	a. Analyze the coefficients from the model. Features with higher coefficients might indicate significant influencers on driver performance. b. Understand the relationship direction from the sign of coefficients. For instance, does higher education level positively or negatively correlate with performance?
c. Visual Representations	Visuals can often convey information more effectively than numbers alone.	a. Use plots to represent the distribution of top drivers by city or education level. b. Showcase the correlation

		between driver ratings and factors like age, income, and education through scatter plots or heatmaps.
d. Trade-off Analysis	Decisions about driver incentives, training, or recruitment may come with trade-offs.	a. Discuss the implications of recruiting more educated drivers versus the costs associated with it. b. Analyze the benefits of investing in driver training compared to potential increases in customer satisfaction.
e. Recommendations	Recommendations should align with Ola's business goals and be based on data-driven insights.	a. Suggest specific strategies like targeted training programs, improved recruitment processes, or incentive schemes based on model insights. b. Provide evidence from the data analysis, such as cities with the most potential for growth or key age demographics to target.
f. Feedback Loop	The transportation industry is dynamic. It's essential to ensure continuous monitoring and adaptability..	a. Propose setting up a periodic review process to assess the model's relevance and performance. b. Recommend surveys or feedback mechanisms to collect data on new trends, driver concerns, and customer feedback to refine the model in the future.

Questionnaire (Answers should present in the text editor along with insights):

1. What percentage of drivers have received a quarterly rating of 5?
2. Comment on the correlation between Age and Quarterly Rating.
3. Name the city which showed the most improvement in Quarterly Rating over the past year
4. Drivers with a Grade of ‘A’ are more likely to have a higher Total Business Value. (T/F)
5. If a driver's Quarterly Rating drops significantly, how does it impact their Total Business Value in the subsequent period?
6. From Ola's perspective, which metric should be the primary focus for driver retention?
    1. ROC AUC
    2. Precision
    3. Recall
    4. F1 Score
7. How does the gap in precision and recall affect Ola's relationship with its drivers and customers?
8. Besides the obvious features like "Number of Rides", which lesser-discussed features might have a strong impact on a driver's Quarterly Rating?
9. Will the driver's performance be affected by the City they operate in? (Yes/No)
10. Analyze any seasonality in the driver's ratings. Do certain times of the year correspond to higher or lower ratings, and why might that be?
