LoanTap Logistic Regression Business Case

❖ Topic: Logistic Regression in Credit Underwriting
❖ Duration: 1 week

Why this case study?

From the company's perspective:

- LoanTap is at the forefront of offering tailored financial solutions to millennials.
- Their innovative approach seeks to harness data science for refining their credit underwriting process.
- The focus here is the Personal Loan segment. A deep dive into the dataset can reveal patterns in borrower behavior and creditworthiness.
- Analyzing this dataset can provide crucial insights into the financial behaviors, spending habits, and potential risk associated with each borrower.
- The insights gained can optimize loan disbursal, balancing customer outreach with risk management.

From the learner's perspective:

- Tackling this case offers practical exposure to real-world financial data and its challenges.
- Logistic Regression, a foundational algorithm, is pivotal in binary outcomes like loan decisions.
- Participants will hone skills in data preprocessing, model evaluation, and understanding trade-offs, essential in the data science realm.
- The case emphasizes actionable insights, fostering the ability to drive data-informed strategies in financial sectors.

Dataset Explanation: LoanTapData.csv

1. loan_amnt: Amount borrower applied for.
2. term: Loan duration (36 or 60 months).
3. int_rate: Interest rate on loan.
4. installment: Monthly repayment amount.
5. grade: LoanTap assigned loan grade (Risk ratings by LoanTap.)
6. sub_grade: LoanTap assigned loan grade (Risk ratings by LoanTap.)
7. emp_title: Borrower's job title.
8. emp_length: Duration of borrower's employment (0-10 years).
9. home_ownership: Borrower's housing situation (own, rent, etc.).
10. annual_inc: Borrower's yearly income.
11. verification_status: Whether borrower's income was verified.
12. issue_d: Loan issuance month.
13. loan_status: Current status of the loan.
14. purpose: Borrower's reason for the loan.
15. title: The loan's title provided by the borrower.
16. dti (Debt-to-Income ratio): Monthly debt vs. monthly income ratio.
17. earliest_cr_line: Date of borrower's oldest credit account.
18. open_acc: Number of borrower's active credit lines.
19. pub_rec: Negative records on borrower's public credit profile.
20. revol_bal: Total credit balance.
21. revol_util: Usage percentage of 'revolving' accounts like credit cards.
22. total_acc: Total number of borrower's credit lines.
23. initial_list_status: Loan's first category ('W' or 'F').
24. application_type: Individual or joint application.
25. mort_acc: Number of borrower's mortgages.
26. pub_rec_bankruptcies: Bankruptcy records for borrower.
27. Address: Borrower's location.

What is Expected?

Assuming you are a data scientist at LoanTap, you are tasked with analyzing the dataset to determine the creditworthiness of potential borrowers. Your ultimate objective is to build a logistic regression model, evaluate its performance, and provide actionable insights for the underwriting process.

Submission Process:

Once you've completed the case study...

- Draft your findings and the entire process in a Jupyter Notebook.

- In the notebook, ensure that you:
    - Display the Python code for all your analysis, model building, and evaluation.
    - Provide visualizations like plots, heatmaps, ROC AUC curves, and more, which support your findings.
    - List down valuable insights derived from the data analysis, and suggest actionable recommendations for LoanTap to enhance their underwriting process.

- Convert your Jupyter Notebook into a PDF (Save as PDF using the Chrome browser's Print command).

- Upload the PDF on the platform as per the submission guidelines.

- Remember, once submitted, you won't have the option to edit your submission.

General Guidelines:

- This is a real-world scenario and represents the type of tasks many data scientists deal with daily. So, make sure you take this opportunity to immerse yourself fully.

- It's natural to encounter challenges or even feel overwhelmed at certain stages:

    - Revisit the problem statement to ensure you're on the right track.
    - Break down complex tasks into smaller, manageable chunks.
    - If you encounter errors, don't hesitate to search online or refer to documentation. Remember, problem-solving is a crucial aspect of a data scientist's job.
    - Engage with your peers. The provided discussion forum link can be a valuable resource. Share your challenges and insights, and you might just find the spark you need.
    - Don't hesitate to review lecture materials or external resources if you feel uncertain about any topic.
    - If you believe there's a significant issue or need clarity on the problem statement, reach out to your Instructor.

What does 'good' look like?

1. Define Problem Statement and perform Exploratory Data Analysis

	Hint	Approach
a. Definition of problem	Begin with a clear and concise problem definition. What is the objective of LoanTap? Why is determining loan eligibility crucial?	Determine the creditworthiness of potential borrowers using various attributes, to ensure that the loans are given to those who are most likely to repay them
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.	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.
c. Univariate Analysis	Begin by understanding individual variables. (distribution plots of all the continuous variable(s) barplots/countplots of all the categorical variables)	For continuous variables, use histograms or density plots. For categorical variables, use countplots. Tools like Seaborn make these visualizations straightforward. This helps in understanding the distribution of individual variables.
d. Bivariate Analysis	Dive into relationships between two variables. (Relationships between important variable)	Scatter plots for continuous vs. continuous relationships, boxplots for categorical vs. continuous, and crosstab or stacked bar plots for categorical vs. categorical relationships. For example, you can use a box plot to compare the loan amounts between those approved and not approved.

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.	Each plot should have an accompanying 2-3 line comment or observation. For example, "The majority of loan applicants are in the age range of 30-40. Age vs Loan Amount scatter plot indicates that older individuals generally apply for higher loan amounts."

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. Creation of Flags: For attributes like Pub_rec, Mort_acc, and Pub_rec_bankruptcies, consider creating binary flags based on certain conditions. b. Extracting month and year from date-related variables can help in capturing time-related patterns. c. Deriving state or region from address fields can narrow down the geographical spread and indicate location-based trends. d. Mapping the emp_length might involve converting textual employment lengths into numerical values or categories,

		making them more useful for modeling.
e. Multicollinearity and Feature Selection	Reducing multicollinearity and choosing relevant features can lead to better model performance.	a. VIF (Variance Inflation Factor): Calculate VIF for each predictor variable. A VIF above a certain threshold (often 5 or 10) might indicate multicollinearity. b. RFE (Recursive Feature Elimination): This technique helps in selecting the most important features by recursively reducing the number of attributes and evaluating the model performance.
f. 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.
i. Identify normal vs skewed	Identify normal vs skewed distributions and understand why.	For continuous variables, comment on the skewness. For

distributions and understand why.		relationships, comment on positive or negative correlations, clusters, or other patterns noticed.

3. Model building

	Hint	Approach
a. Build the Logistic Regression model	Logistic Regression is suitable for binary classification problems.	Prepare the data for training and validation, initiate the logistic regression algorithm, and fit the model on the training data
b. Hyperparameter Tuning	Hyperparameters are the parameters of the model that are not learned automatically and need to be set manually or iteratively. Tuning them can enhance the model's performance.	a. Use techniques like GridSearchCV or RandomizedSearchCV to perform hyperparameter tuning. b. Define the grid of hyperparameters. For Logistic Regression, parameters like regularization strength (C), penalty (l1 or l2) can be tuned.
c. Handling Class Imbalance	In many real-world datasets, especially in cases like fraud detection or loan default prediction, the number of instances of one class can heavily outnumber the other, leading to a model that performs poorly on the minority class.	a. Exploratory Data Analysis (EDA): During EDA, check the distribution of the target variable. If one class heavily outnumbers the other, consider it an imbalance b. Resampling Techniques Oversampling: Increase the number of instances in the minority class by replicating data or generating synthetic samples. Libraries like SMOTE (Synthetic Minority Over-sampling Technique) can

		be beneficial. c. Algorithmic Approach: Use algorithms that allow setting class weights, making the model pay more attention to the minority class. In LogisticRegression from sklearn, you can set the class_weight parameter to 'balanced'.
d. Display model coefficients with column names	Coefficients provide insight into the importance and relationship of predictors with the outcome.	Once the model is trained with the best hyperparameters, extract coefficients and map them against predictor names to understand the weight/importance of each feature.

4. Results Interpretation & Stakeholder Presentation

	Hint	**Approach**
a. Understand the Business Context	Your model is only as good as the actionable insights it can provide.	Familiarize yourself with the business objectives, challenges, and KPIs. This will enable you to frame your findings in the context that resonates with business stakeholders.
b. Interpreting Model Coefficients	Knowing which features significantly affect the target variable can provide actionable insights.	Review the coefficients from the logistic regression. Higher absolute values indicate greater importance. Consider the sign (+/-) to understand the direction of the relationship.

c. Visual Representations	Visuals can often convey information more effectively than numbers alone.	Utilize well-labeled charts, plots, and graphs (like ROC AUC curve and Precision recall curve) to visualize and emphasize key insights
d. Trade-off Analysis	Business decisions often involve trade-offs. It's crucial to understand the balance between risk (false positives) and opportunity (financing more individuals).	Highlight the implications of false positives and false negatives. Discuss strategies to strike a balance, such as adjusting the classification threshold.
e. Recommendations	End with actionable recommendations derived from your analysis	Recommend strategies to improve loan approval processes, mitigate risks, or capitalize on opportunities. Support these recommendations with evidence from your analysis.
f. Feedback Loop	Continuous improvement is key in analytics.	Propose methods for continuously monitoring the model's performance over time and iterating based on new data or changing business needs.

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

1. What percentage of customers have fully paid their Loan Amount?
2. Comment about the correlation between Loan Amount and Installment features.
3. The majority of people have home ownership as _____.
4. People with grades ‘A’ are more likely to fully pay their loan. (T/F)
5. Name the top 2 afforded job titles.
6. Thinking from a bank's perspective, which metric should our primary focus be on..

    1. ROC AUC
    2. Precision
    3. Recall
    4. F1 Score
7. How does the gap in precision and recall affect the bank?
8. Which were the features that heavily affected the outcome?
9. Will the results be affected by geographical location? (Yes/No)
