Final: Loan Defaults

# Can you predict loan default?

You’ve recently joined a major financial institution and are tasked with developing and comparing several machine learning models to predict “loan status”. Specifically which loans are likely to default. Since we now work in a regulated industry, we need to ensure that our models are both explainable and that the predictions that we create can be explained; and most importantly we can document what we did!

## Models to be Evaluated.

You will train and evaluate different models.

* **Logistic Regression**: Serves as a baseline for performance comparison.
* **Random Forest**: An ensemble method known for its robustness and ability to handle complex data structures.
* **Gradient Boosting Machine (GBM) OR XGBoost**: Advanced ensemble techniques known for their predictive power.
* **Neural Network**: An approximation method known for it’s ability to identify non-linear relationships.
* [**StackingClassifier**](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.StackingClassifier.html) **~~or AutoGluon Weighted Ensemble~~**.

## Methodology / Approach

Here you want to document your methodology, here is the standard approach:

**Data Exploration and Preprocessing**

1. **~~Exploratory Data Analysis (EDA)~~**~~: Conduct an initial analysis to understand the data's characteristics, including distribution of the target variable, missing values, and potential outliers.~~
2. **~~Data Preprocessing~~**~~: Address missing values, deal with extreme values, encode categorical variables, and prepare the data for modeling.~~

**Model Development**

1. **~~Model Training~~**~~: Develop models using Logistic Regression, Random Forest, and GBM/XGBoost, Neural Network and an AugoGluon or SKLEARN stacking ensemble model on the training data.~~
2. **~~Parameter Tuning~~**~~: Optimize model parameters to enhance performance – for scikit-learn based models.~~

**Global Model Explanations**

1. **~~Model Comparison~~**~~: Compare the models based on their performance and feature importance scores to identify the most effective model on both the train and test sets. What metric should you choose to pick the best performing model and why?~~
2. **~~Feature Importance~~**~~: compare the feature importance from one model to another, why would they be different?~~
3. **~~Roc Curve & PR Curves on TEST set~~**~~. you need to plot all of your models with a ROC curve and a PR curve, explain what they mean and how to interpret them.~~
4. **~~Partial Dependance Plot~~**~~: Pick your best performing model and generate partial dependance plots explaining the AVERAGE impact to predictions in a way that a business person can understand.~~
5. **~~Operational Strategy at 2% and at 5% FPR~~**~~: Propose a strategy to achieve and maintain a 2% and also 5% false positive rate, detailing its implications on recall and precision. What does this mean for the business in plain language?~~

**Local Explanations**

Using your best performing model, make predictions on your test set and identify the TOP 10 best True Positives, your TOP 10 False Positives and your Top 10 False Negatives. Eyeball the data, or use plots or trees or some other method to explain why you think you gave a prediction a particular score. Is there a pattern or common theme that you notice for wrong predictions?

**Final Deliverables**

* **Executive Summary (2 pages max)** 
  + ~~The Firm believes that credit score is the most important predictor, your write up should discuss why or why not.~~
  + ~~Finally, the firm wants to operate at a 5% false positive rate, how can they do that? based on your best performing model what is the rule with predicted probably of fraud threshold that will give them a 5% false positive rate, what is the recall and precision at that threshold? what does it mean to operate at 5% false positive rate?~~
  + ~~Firm would like to understand in “plain language” how the firm should use the PR Curve and ROC AUC to operate their model, what are these curves and why are they important.~~
* **Model Report(unlimited pages but be sensible)**: A comprehensive WELL-ORGANIZED write-up ~~detailing the EDA, Methodology, Global Explanations, Local Explanations.~~ 
  + **~~EDA & Preprocessing~~**
  + **~~Methodology~~**
  + ~~Model Training & HPO~~
  + ~~Global Explanation: model performance, charts, variable importance, partial dependance plots.~~
  + ~~Local Explanations: explain top your top 10 correct predictions of loan default (loan default = 1 aka True Positive), your top 10 predictions of loan default = 0(true negative); and your top 10 incorrect predictions both false positive and false negative. You can simply do this by eyeballing the data or can you use some more impressive method?~~
* **Predictions on Holdout Set**: Apply the best-performing model to the **fraud\_holdout.csv** dataset and prepare a submission file with predictions.

You have been provided two datasets:

* fraud\_training.csv – use this one to train and evaluate your model
* fraud\_holdout.csv – use this one to score your hold out set submit your predictions with your report. (just include ID and P\_DEFAULT in your CSV)

## Deliverables for Grading

For this project, you will complete and submit the following.

1. **Executive Summary (2 pages)**
2. **Model Report** (unlimited just be organized) - The report will detail the data munging/shaping, data understanding, data preparation and modeling phases of this project. In the report template (*project\_template.docx*) provided to you in project 2.
3. **Notebook** – I’ll want your notebook used to produce results in the report. Your code should be appropriately commented so I can tell what is going on – ideally, I should be able to repeat your analysis.
4. **Predictions file** – We want you to apply your model to the loan\_holdout.csv data set. You will submit this file with your report – note this is competitive you must beat the benchmark to get any points!
   1. *Word of caution* ***–*** *it is incredibly easy to identify the same submissions. I’d hate to have to deal with honor code violations at the end of this term!*

## Required Tasks

You will need to write an Executive Summary and Model Report which will contain your detailed analysis. And you will need to submit your predictions to us as well.

## Executive Summary

What problem are you challenged with? What were 3 or 4 key findings (things you found interesting that influenced the model or things that were odd about the data). What was result of your model, and 2-3 recommendations that you’d make to make your model actionable.

* ~~State the problem~~
* ~~3-4 Key findings~~
* ~~Model Performance & Interpretation of it.~~ 
  + ~~Be sure to address precision / recall vs AUC~~
* ~~2-3 actionable Recommendations~~ 
  + ~~Data quality is not a recommendation~~

Helpful hint: do not attempt to draft an executive summary until after you’ve done the analysis and written the modeling report. The executive summary should be the last thing you should do.

## Detailed Analysis (BE ORGANIZED)

This is the meat of your report, you should follow the steps provided in the report\_template.docx document for details and steps. Your Detailed Analysis is not about getting you to write a long report really it is about communicating your understanding and findings.

* ~~Explorations relative to the target~~
* ~~Anomaly detection – can you identify and explain 5-10 anomalous records~~
* ~~Minimum of 3 different models trained and compared make sure you address the following~~
  + ~~Table of performance, ROC, Precision, Recall, F1 (What is F1)~~
  + ~~Roc Chart w. Score Threshold and Description of what it means. Ex. If you select a score of 0.5 where does that put your FPR and TPR on the Test set?~~
  + ~~Precision Recall Chart – if you select a score of 0.5 what does that translate into as far as precision and recall?~~
  + ~~Operating Table – where would you recommend operating your model?~~
* ~~How did you choose hyper parameters?~~
  + ~~Did you do hyper parameter tuning, did you have sufficient experiments to tune?~~
* ~~Global explanations of your best model:~~ 
  + ~~Variable importance~~
  + ~~Partial dependency plot of top variables~~
* ~~Local Explanations -~~ 
  + ~~TP – top 10 true positives, loan default = 1 and ordered by pred\_1 score DECENDING~~
  + ~~FP – top 10 false positives, loan default = 0 and ordered by pred\_1 score DECENDING (high scoring but actually didn’t default)~~
  + ~~FN - top 10 true negatives, loan default = 1 and ordered by pred\_1 score ASCENDING (low scoring that did default)~~

## Submission (5 points)

This final predictive submission, points will be awarded based on your position on the leaderboard ABOVE the benchmark, if you are in the top 10% you will get 5 points, in the next 10% you’ll get 4.5 points. IF you fail to beat the benchmark, you’ll get -1 point, if you fail to submit a prediction your get -5 points.

|  |  |
| --- | --- |
| decile above benchmark | Points |
| 1st decile (first 10%) | 5 |
| 2nd | 4.5 |
| 3rd | 4 |
| 4th | 3.5 |
| 5th | 3 |
| 6th | 2.5 |
| 7th | 2 |
| 8th | 1.5 |
| 9th | 1 |
| 10th | .5 |
| failure to beat benchark | -1 |
| no submission | -5 |

### Some expectations:

* Make sure your report is visually appealing and easy to read. Style is important.
* Make sure you back up assertions with numbers!
* Make sure every chart you create has:
  + Title, x & y axis labels, and **a description of why someone** would care to look at it.
* Make sure any table you create has
  + A description of how to interpret it.
* Make sure you always evaluate multiple models and compare them.
* Make sure you spend time understanding what your code does.

Info About Lending Club Data

The Lending Club dataset is a popular dataset used for analyzing peer-to-peer lending data. It contains detailed information about loans issued through the Lending Club platform, including borrower information, loan details, and payment information. While I can provide a general overview of what you might expect to find in a Lending Club data dictionary, please note that the specifics can vary depending on the version of the dataset you're using.

Here's an overview of some of the fields you might encounter in a Lending Club data dictionary:

1. **loan\_amnt**: The listed amount of the loan applied for by the borrower.
2. **funded\_amnt**: The total amount committed to that loan at that point in time.
3. **funded\_amnt\_inv**: The total amount committed by investors for that loan at that point in time.
4. **term**: The number of payments on the loan. Values are typically in months and can be either 36 or 60.
5. **int\_rate**: Interest Rate on the loan.
6. **installment**: The monthly payment owed by the borrower if the loan originates.
7. **grade**: LC assigned loan grade.
8. **sub\_grade**: LC assigned loan subgrade.
9. **emp\_title**: The job title supplied by the Borrower when applying for the loan.
10. **emp\_length**: Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
11. **home\_ownership**: The home ownership status provided by the borrower during registration. Values are: RENT, OWN, MORTGAGE, OTHER.
12. **annual\_inc**: The self-reported annual income provided by the borrower during registration.
13. **verification\_status**: Indicates if income was verified by LC, not verified, or if the income source was verified.
14. **issue\_d**: The month which the loan was funded.
15. **loan\_status**: Current status of the loan (e.g., Fully Paid, Charged Off, Default, In Grace Period, Late).
16. **purpose**: A category provided by the borrower for the loan request (e.g., debt consolidation, credit card refinancing, home improvement).
17. **title**: The loan title provided by the borrower.
18. **zip\_code**: The first 3 numbers of the zip code provided by the borrower in the loan application.
19. **addr\_state**: The state provided by the borrower in the loan application.
20. **dti**: A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
21. **earliest\_cr\_line**: The month the borrower's earliest reported credit line was opened.
22. **open\_acc**: The number of open credit lines in the borrower's credit file.
23. **pub\_rec**: Number of derogatory public records.
24. **revol\_bal**: Total credit revolving balance.
25. **revol\_util**: Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
26. **total\_acc**: The total number of credit lines currently in the borrower's credit file.
27. **initial\_list\_status**: The initial listing status of the loan. Possible values are – W, F.
28. **application\_type**: Indicates whether the loan is an individual application or a joint application with two co-borrowers.
29. **annual\_inc\_joint**: The combined self-reported annual income provided by the co-borrowers during registration for joint applications.
30. **dti\_joint**: A ratio calculated using the co-borrowers’ total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the co-borrowers’ self-reported monthly income for joint applications.