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.

Your task is to build and compare 3 different models, predicting (loan\_default); identify any outliers and explain them; 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 have been provided two datasets:

* loan\_train.csv – use this one to train and evaluate your model.
* loan\_holdout.csv – use this one to “score” and post to Kaggle.

## Deliverables for Grading

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

1. **Executive Summary (20 points)**
2. **Model Report** (25 points) - 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 1.
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 to Kaggle to see if you beat the benchmark – note this is competitive you must beat the benchmark to get any points on your Kaggle submission!
   1. *Word of caution* ***–*** *it is incredibly easy to identify the same Kaggle submissions. Up to this point I haven’t made an issue of this – however I will be diving into these 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 Kaggle

## Executive Summary (15 points)

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 (30 points)

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, LogLoss, Precision, Recall, 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? F1 score?
* 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)

## Kaggle Submission (5 points)

This final Kaggle 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.