Monthly Challenge: Lending Data

Background

This challenge builds on the groundwork laid in the Machine Learning Project course, where we explore using past loan data from Lending Club to build models that can predict if a loan will be paid off on time or not. In the first 2 missions, we remove features that leak data (e.g. if the loan was paid on time or not), contain many missing values, or not useful (e.g. first 3 numbers in the applicant's zip code).

In the third mission, we experiment with logistic regression and random forest models to try to achieve a high true positive rate (TPR) and a low false positive rate (FPR). Our best logistic regression models achieved a TPR of 67% with an FPR of 40% and a TPR of 20% with an FPR of 7%. While one model had a good TPR (67%), the other one had a good FPR (7%). One model compromises TPR for FPR while the other model compromises FPR for TPR.

Since we were looking at the problem from the conservative investor's standpoint, we were more interested in a low FPR than a high TPR. In this challenge, you'll try to build a model that narrows the compromise between TPR and FPR.

Challenge

Construct a machine learning model that achieves a **TPR greater than 50%** while maintaining a **FPR less than 7%**. In the end of the third mission, we make a few recommendations for ways to improve the models. We've expanded the recommendations here as a starting point for your exploration:

- Better handle the class imbalance
    - Try undersampling, oversampling, and mixing both undersampling and oversampling. Read more on Wikipedia.
    - Try varying the penalty values further.
- Try other models
    - Scikit-learn supports many models. Use this is an opportunity to research models you haven't worked with or learned about but the library supports.
    - Explore ensembling models.
- Improve the features the model uses
    - Explore columns we discarded to see if you can turn them into useful features.
    - Explore different the feature selection techniques that scikit-learn supports. Use this as an opportunity to research and understand the different feature selection approaches.
- Explore tuning the hyperparameters of the machine learning models.
    - Hyperparameters are any property of the model that affects how the model behaves that's independent of the dataset. The **weight_class** parameter is an example of a hyperparameter.

- Learn about the different hyperparameter tuning techniques that scikit-learn supports.

Dataset

1. Navigate to https://www.lendingclub.com/info/download-data.action.
2. In the left section (titled DOWNLOAD LOAN DATA), set the default **Year** value to **2007-2011** if it's not selected by default, and click the **Download** button.
3. Review and understand the data cleaning & feature selection we explore in the first 2 missions in the Machine Learning Project course.
4. Transform the **loan_status** column, the target column we want the model to predict:
    a. **Fully Paid** becomes **1**
    b. **Charged Off** becomes **0**