One dataset, many users.
Ten use cases and motivations for building a machine learning model using the Lending club dataset.

By
Sri Krishnamurthy

As per Wikipedia[1], "LendingClub is a US peer-to-peer lending company. The company claims that $15.98 billion in loans had been originated through its platform up to December 31, 2015. Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee."

With the release of Lending club data [2], a lot of researchers and practitioners are leveraging this dataset to understand credit risk, borrowing patterns, investment opportunities and to understand consumer choice and behaviors. This dataset is a gold mine to teach data science [3] and a lot of schools are using it to teach data science.

In this case, we introduce 10 fictional characters who may be interested in working with Lending club. They have hired your consulting team to analyze the Lending club dataset to achieve their specific goals. In this guided exercise, we will ask each team to understand the needs of each persona and provide them a model that caters to their specific needs. Here is the cast of characters.

1. Rick the Investor (Risk averse)
2. Tola the Investor (Risk taker)
3. Taz the Borrower (Good credit)
4. Pip the Borrower (Bad credit)
5. Bipa the Portfolio manager (Who lends to multiple people of different risk profiles)
6. Arc the Arbitrager (https://www.lendacademy.com/social-lending-arbitrage-good-or-bad-idea/)
7. Slick the Data scientist
8. Irs the Professor
9. Dat the Data vendor who sells data and insights
10. Mar the Regulator who regulates how the model can be used/not used

Your team is very familiar with the Lending club data, but you need to educate your client on the company and the insights you draw from the data.

Task 1:
Framing: Understanding your client, available data and exploratory data analysis:

Your first task is to understand your client, the Lending club platform, the loan statistics and the investor performance.

1. Research and summarize what your client's needs are going to be.
2. Explore the data and comment on data quality, features and get a feel for the data. For the purpose of this assignment, we will restrict the data to https://www.kaggle.com/wendykan/lending-club-loan-data
3. Lending club has generated many graphs for you:
   - https://www.lendingclub.com/info/statistics.action
   - https://www.lendingclub.com/info/statistics-performance.action
   - https://www.lendingclub.com/info/demand-and-credit-profile.action

Write a 2-page summary highlighting the key takeaways from these charts and what insights do you think your client could gain from these charts.

Task 2: Data preparation
Data cleansing, Pre-processing, Feature engineering

1. Put together a plan for data cleaning, pre-processing and do feature engineering
2. Use Featuretools (https://github.com/featuretools/featuretools/) to do the same.
3. Write a note on your observations on featureTools vs manual Feature engineering

Task 3: Prediction

1. Your goal is to build 3 models (Regression, Random forest, Neural Networks) to predict interest rates. Write a design spec for your methodology for designing each model. Discuss your Independent and Dependent variables.
2. Try all the 3 models and use MAPE as a criteria to choose the best model. Summarize MAPE for both Training and Testing data.
3. Try 5-fold cross validation. How does the model performance change?

Task 4: Hyper-parameter optimization

1. Hyperparameter tuning. You will try Hyperparameter tuning to see if you could do better. Write a design on which hyper parameters you will try to optimize. We are providing a few hyper parameters
   a. Regression: Try L1, L2, Elasticnet regularization
   b. Neural networks: Change epochs, optimizers, learning rate
   c. Random forest: No of trees, Tree depth

   Try these and try other ones of your choice.
   Tune your hyper parameters using. https://scikit-learn.org/stable/modules/grid_search.html
   Write a report that discusses the effect of Hyper parameter tuning
2. AutoML is a craze everyone is behind. Use
   a. TPOT
   b. AutSKLearn
   c. H2o.ai

And summarize the MAPE for each model

3. Discuss your manual model vs AutoML approaches with respect to:
    a. Interpretability
    b. Reproducibility

Task 5: Analysis

Build out test cases to test out use cases that are important for your client. Run these tests and write a 1-page report to your client on why they should/shouldn't use your model to predict interest rates and how they could use your model.

References:

1. https://en.wikipedia.org/wiki/Lending_Club
2. https://www.lendingclub.com/info/download-data.action
3. https://www.liebertpub.com/doi/full/10.1089/big.2018.0092

Notes:

- Each task will be graded. They have equal points.
- Deadline for delivery: March 22nd, 11.59pm
- You will put together a 10 minute presentation to be presented on March 23rd.