# Capstone Project Reflection - Lending Club

## Team TARP - James Anderson, Jon Harris, Vincent Ji, and Joe Lu

**1. After finishing your project, are you able to fulfill all objectives of your proposal?**
Please list the steps from your proposal and respective achievement rates and shortcomings. If you achieved more than what you had proposed originally, please add it and describe your major achievements.

| Steps/items | % comp. | Comments |
|---|---|---|
| Extract Transfer Load (ETL) | 100% | |
| Exploratory Data Analysis | 100% | |
| Feature Engineering | 100% | |
| Train logistic regression model | 90% | Produced reasonable results but short of the 5% false positive goal we set |
| Train xgboost model | 90% | Produced better results than logistic regression but there is still room to improve |
| Finalize Results | 100% | |
| Presentation | 80% | Wrapping up the deck |

**2. What are the major obstacles your team has faced to fulfill your project objectives?**
Please analyze it through several tangents.

| | |
|---|---|
| **Data collection and preparation** | The dataset is too large for Jupyter notebook to do EDA and machine learning. We had to use two different methods to sample 5% of total data (~2 million observations). Also, we leveraged SQL (all data) to sanity-check if the sampling results are accurate. |
| **Business insights generation** | Scoping the project was challenging because there were many angles to approach the Lending Club dataset.<br>We took a broad approach at the beginning by answering 40 |

| | |
|---|---|
| | questions in Aiko’s project proposal as well as some new questions the team created. We were able to derive insights from the data and make low-hanging-fruit recommendations but the process took quite a few days. |
| **Proposal design** | Our original idea of running a simple model early wasn’t practical because we couldn’t understand the importance of all the features to feed into it. We ended up running two models in parallel in the second half of the project. |
| **Team organization and management** | - Working remotely certainly has been a major challenge for the team. Productivity takes a hit when you are not working side-by-side.<br>- We tried to divide the work evenly so everyone has exposure to ETL, EDA, feature engineering, and modeling. This is good for educational purposes and keeping the team informed. On the other hand, there is some redundancy in the work that was done.<br>- During the scoping phase, we were spinning our wheels a bit trying to figure out the topic to focus on. |
| **Technical strength/weakness** | Weakness<br>- The team didn’t have a lot of experience dealing with unbalanced dataset.<br>Strength<br>- Everyone was able to code/analyze data and contribute to the overall goal. |
| **Unforeseen issues** | - EDA took longer as we tried to gather all the info before creating a story<br>- It takes a long time to tune hyperparameters for Random Forest model |

**3. Please list the contributions of individual team members in the projects and their roles.**
Write in % the overall contribution of each team member.

| Team members | % contr. | Comments |
|---|---|---|
| James Anderson | 25% | |
| Jon Harris | 25% | |
| Vincent Ji | 25% | |
| Joe Lu | 25% | |

**4. In finishing your project, summarize what your team, both at an individual level and at a group level, have learned in the process.**

Throughout the bootcamp, the lecturers have stressed the importance of EDA. The lending club project was no exception, with EDA taking up a majority of our time. Previously mentioned, each team member took part in the EDA, and as a result, each greatly improved their Pandas and Matplotlib skills, with some members additionally incorporating visualization packages such as Seaborn and Plot.ly. Specific to the project, the team was able to gain experience with imbalanced datasets and logistical regression/xgboost classification. While several members were able to leverage their financial background in the current project, Jon Harris used this project to immerse himself in financial jargon and better understand inputs and outputs of investments. Perhaps most importantly, the team gained more experience in working remotely, collaboration using  github, and learning from each other.

**5. If the capstone project starts over again, what would you do differently to address the issues you have identified and encountered.**

Due to the number of features (n=150),  the team quickly rushed to pick apart every feature, determine what each represented, and which *could be* most relevant - prior to selecting a project topic. In hindsight, the team should have first investigated Aiko's 40 background questions prior to selecting a project topic, which would have familiarized the team to the data and tangentially help us identify which features may be most relevant. We wasted valuable time guessing what could do with the data. Only after a few days did we realize this was a wasted effort because the project goal only required an understanding of the ~30 features available to the investor on the Lending Club website.

**6. Can you think of other industries and topics on which you could apply the same technique and methodology you used in your capstone project?**

1. Education - use it for admission of applicants
2. Marketing - predict which potential customer targeted via ads will make a purchase
3. Healthcare - analyze which patient has certain disease

**7. Please list the names of instructors and TAs you have discussed with for your project and estimate the amount of time you have spent with them.**

| Staff members | Time | Comments |
|---|---|---|
| Thomas | 1.5hrs | Project proposal meeting and office hour |
| Aiko | 3hrs | Three one-hour meetings from project scoping to machine learning |
|  |  |  |