Cover Letter

Alex Trostanovsky, 100984702

CGSC 3999A Summer 2018 - First Report

August 16th, 2018

Dear Prof. Slobodenyuk,

The following report, *Home Credit Default Risk Evaluation*, details my Summer 2018 Co-op work term as a Software Developer at Apption, under the supervision of Imtiaz Fazal (IF), Director of Analytics.

I hereby acknowledge that this report adheres to the guidelines set for work term reports and that this report consists of solely my own work.

Carleton University

Apption

CGSC 3999A - Work Term Report 1

Home Credit Default Risk Evaluation

Author:
Alex Trostanovsky,
100984702,
alextrostanovsky *at*
cmail.carleton.ca

Supervisor:
Imtiaz Fazal, (IF),
Director of Analytics

August 16, 2018

Executive Summary

The following report details my participation in the Home Credit Default Risk Kaggle Competition.

- Context: Driven by an interest to enter the Fintech (Financial Technology) sector, Apption decided to investigate the publicly available datasets in an active Kaggle competition (Kaggle is a Data Science Portal for Public Education and Competitions).
- Project Duties:
  - To investigate the datasets and determine which criteria and features were indicative of loan applicants with a high risk of default.
  - To learn more about the algorithms behind the classifiers being used to label loan applicants with respect to risk of default levels
- Contributions:
  - A full Exploratory Data Analysis (EDA) of seven tables of data containing a combined number of more than 400 features.
  - Documentation detailing the algorithmic and implementational details of:
    - Decision Tree Classifier
    - Random Forest Classifier
    - Light Gradient Boosting Machine Classifier
- Reflections
  - A cursory introduction to foundational machine learning algorithms and project workflows
  - A concrete analysis of important feature engineering procedures imperative to Exploratory Data Analysis in the domain of Loan Default Candidate Classification

Contents

1 Introduction

The repot begins with a description of the organizational environment of Apption. The companies' responsibilities, products, and operating hierarchy are briefly explained. Next, the body of the report will detail my final project of my Co-op work term at Apption Software. I was tasked with entering a Kaggle competition (Kaggle is a forum for publicly available Machine Learning competitions, discussions, and tutorials). The objectives of my work in entering the competition are described, as well as the company's motivation in gaining the domain specific knowledge required for the competition, and what Apption hoped to learn by entering it.
An overview of the methodolgy I employed in introducing myself to the domain of machine learning, and the technical environment which I familiarized myself with in order to adapt to the common workflows employed in the industry is provided.
Next, the interim results of my Exploratory Data Analysis (EDA), feature engineering process, and algorithmic tuning are presented. Some challenges which I faced in entering this Data Science competition are raised, and the solutions which I came up with to face these challenges are discussed.
Lastly, I reflect on my participation in this competition, and consider the impact this project had on my technical skill repertoire. I consider how the completion of this project relates to my academic studies, and how it may have contributed to Apption Software.

1.1 Organizational Context

Apption (located at 1600 Scott Street, #400, Ottawa, ON, K1Y4N7) is a small-medium sized Software Development company offering data analysis solutions organizations and businesses. Apption then embeds these solutions into operational applications to maximize business performance. [1] From my experience, Apption Software projects can be divided into two main categories:

- Analytic Projects
- Modernization Projects

In my work term at Apption, I was exposed to, and completed work which was concerned with these two categories. However, this report is specifically

concerned with my experience with an exploratory foray into the domain of Loan Analytics.
The heirarchy of my workflow supervision and reporting chain of command is described in Fig. 1.

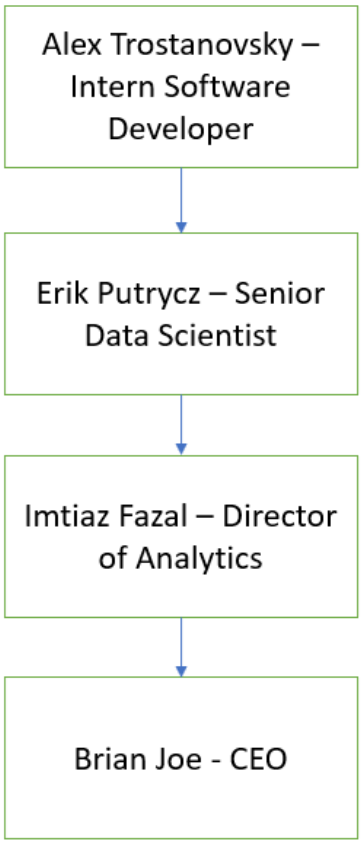

Figure 1: Work Heirarchy

2 Work Experience

Apption was interested in exploring the possibility of entering the Loan Analytics and Credit Scoring sector. This segment of the Financial Technology (Fintech) sector consists of processing potential loan recipient applications, and using the parameters which are available (or which are derived) from candidate applications to construct a score which could inform the businesses that provide said loans about the likelyhood with which a

candidate could default (i.e. be late on payments, or default completely). Apption was interested in developing a knowledge base of the Loan Application and Analytics Marketplace. The company was interested in learning more about the criteria which Loan Providers are looking for when assessing loan applicants' probability of default (i.e. how likely it is that a consumer which has been approved for a loan will either a) be late on repayment of said loan, or b) be unable to repay said loan). Doing so is important for loan providers because:

1. Customers which are more likely to default/be late on loan repayments are risky candidates which could cost Loan Companies revenue (whether on lost income, or repossession fees). However, assigning a rating to each applicant can be beneficial to such a business. For example, consider two applicants: The first has been assessed by the Loan Provider in being likely to be late on 60% of loan repayments, and the other has been assessed comparatively at an 80% risk rate (will be late on 80% of payments). Rating the potential customers as such allows the business to decline the loan to the latter customer, while approving a loan for the former, but at a much higher interest rate than that which a less 'volatile' customer would receive.

2. From discussion with Loan Analytics businesses, Apption has found that many small-medium banks and Loan Portfolio Management Companies currently employ a method which consists of a series of checks which inform the final score of candidates. This approach produces a categorical ranking for potential consumers based on a handful of parameters (such as credit scores and Debt to Income Ratios).
   Consider the rigidity of such a model to the addition of parameters. For example, say that Company X wishes to add 'Presence of Loan Delinquents in Client's immediate Social Circle' as an informative parameter. (This criteria is actually becoming more present in the screening process of Loan Financing Businesses. When screening an applicant, a social media inquiry searches for individuals in the Client's social circle (e.g. Facebook or LinkedIn friends), queries these individuals to see if they have applied / or are currently being financed for a loan by the business. If these individuals have a history of payments being paid late, then this fact will negatively impact the client's application.)
   Adding such a criterion to an existing logical model is a difficult task which requires reworking the conditional logic of the existing grading

system. On the other hand, if Company X adapts a machine learning model to inform Applicant Ratings, adding new criteria is as simple as gathering the data (i.e. performing the social media search), and training the algorithm on the newer (expanded) dataset. (I discuss what 'training' an algorithm refers to in a high-level manner, below.)

2.1 Goal

I was instructed to enter the Home Credit Default Risk Competition. The objective of this competition was to use historical loan application data to predict whether or not an applicant will be able to repay a loan. Home Credit is a global consumer financing service which operates mainly in CIS (Commonwealth of Independence States) and SEA (Southeast Asia) countries. In entering this competition Apption hoped to:

- Learn more about the data which global loan financers require and work with when rating loan applicants
  - Specifically, the data sources which are available to Loan Businesses (whether primary, or third party)
  - The data types (income, demographic, geographic, etc.) being used to inform approvals and rejections
- Learn more about the machine-learning/artificial intelligence (AI) algorithms behind the classifiers employed by Loan Companies

Learning more about these factors will eventually allow Apption to offer potential business solutions to Small-Medium Banks and Loan Financiers which could improve the businesses' productivity and optimize their revenue performance.

2.2 Nature of the work

2.3 Experience

Home Credit is a service which provides lines of credit (or loans) to people with low exposure to banking services. Such businesses as HC (Home Credit) are always interested in improving the prediction algorithms they use to determine whether or not a client is likely to repay a loan on time

or have difficulty in repaying a loan [2]. This is the impetus for HC in hosting this Kaggle Competition: to see what sorts of models the machine learning community can help develop, and how their own (HC) models may be improved.

2.3.1 Data

The data in a classification task is usually divided into:

1. **Training** - The data which is labelled with the classification the algorithm is expected to make. The algorithm will be 'trained' on this data.

2. **Testing** - The data on which the algorithm / model will be tested against upon completion of training. This data is not labelled, even though the label is known to the data scientist. In this way, the data scientist can 'test' the performance of the algorithm. (Is the algorithm mostly correct / incorrect in classifying novel loan candidates?)

Specifically, with regards to this task, training the model consisted of:
Presenting a chosen/developed machine learning algorithm with the training data: a list of parameters per past loan application (e.g. income, age, length of employment at current job, etc.) along with the the binary classification of whether or not this loan was classified as a risky loan or not. To HC, a target variable with value 1 meant 'will have difficulty repaying loan on time', and 0 meant 'will repay loan on time'.

Therefore, this task can be described as a standard supervised classification task:

Supervised: the labels are included in the training data and the goal is to train a model to learn to predict the labels from the features.

Classification: The label is a binary variable. (As was described above).

2.3.2 Submission Judging Metric

This competition judges submissions using a common classification metric known as **Receiver Operating Characteristic Area Under the Curve**

(AUROC)[2].

Once a model is trained on the data, and we have classified the testing data, **The Receiver Operating Characteristic (ROC)** curve graphs the true positive rate verses the false positive rate.
True positive rate measures the ratio of actual positives that are correctly identified as such (i.e. Loan Applicants which were classified as 'will have difficulty repaying loan on time' which were, in reality, late on a certain amount of repayments.)
False positive rate measures the opposite. In this case, this rate will measure the amount of times the classifier has classified a loan candidate as likely to 'have difficulty repaying the loan on time', whereas this client, in reality, did not default on any loan payments.
It should be stated that businesses such as HC are very interested in minimizing this False Positive Rate. That is, if an applicant is incorrectly classified as a risky candidate for a loan, this candidate may be rejected, and HC, in consequence, will lose a client (and a source of potential revenue). Figure 2. represents the performance of three classifiers using the AUROC metric [2].

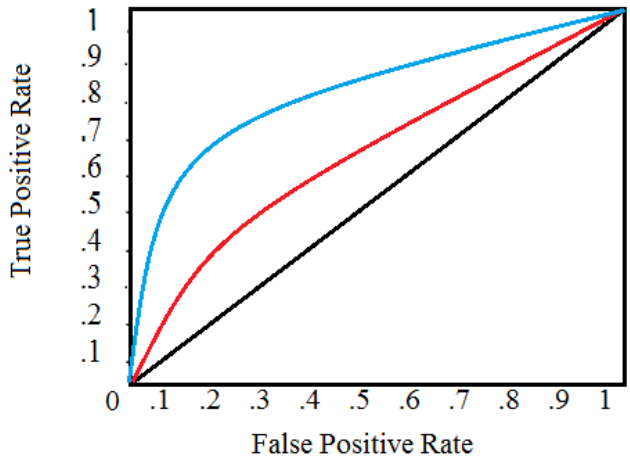

Figure 2: Reciever Operating Characteristic Area Under the Curve

A line that is to the left and above another line indicates a better model. Therefore, the blue line represents a more accurate classifier than the red line, which is more accurate than the black line. The diagonal black line represents a naïve, random guessing model.

In turn, the AUROC metric attempts to maximize the Area Under the Curve of the ROC curve. That is, the naïve guessing model represented by the black line will have an AUC of 0.5, while the AUC of the blue line will be a number less than 1, but larger than 0.5.

2.3.3 Technical Environment

Many machine learning projects benefit from the widely (and freely) available frameworks and libraries which programming languages such as Python support. As I have gained a familiarity with this programming language this past work term, (and as it is one of the most commonly used languages for Machine Learning projects today), I decided to build a model using this language.
I developed this model in the widely used Jupyter Notebook environment. This application is used in Data Science projects because it provides a development environment in an open sourced web application. This environment allows for inline documentation, execution, inspection, and visualization of code without needing to leave the environment. This means that many of the processes which are imperative to any data science project (data sterilization and visualization, statistical modelling, and building and training machine learning models) can be done in a sequential and fluid manner.
Furthermore, Python supports the following frameworks freely available for installation and import:

1. `numpy` - scientific computing with python (containing many useful linear algebra operations essential for machine learning algorithms)
2. `pandas` - open source data structures and data analysis tools
3. `sklearn` - a python machine learning library (containing models and algorithms)
4. `matplotlib` - a python plotting library (useful for data visualization)

2.3.4 Exploratory Data Analysis (EDA)

I have learnt that when presented with a new dataset for a machine learning project, one of the most crucially important (and one of the first) phases is an EDA. This step consists of the calculation of informative statistics and building of figures and graphs to identify trends, anomalies, and

relationships within the data [2]. These discoveries can be used to inform the choices we make about which classifying algorithms to use, and which features are informative in the dataset.

2.3.5 Training the Model

I will now focus on the implementational details of the Random Forest classifier. While this model is relatively basic, it serves as the basis for algorithms which consistently win Kaggle competitions. However, before delving into the process of training a Random Forest classifier, and how it classifies novel instances, I must explain the algorithmic foundation on which Random Forests are built: Decision Tree Classifiers.

2.3.6 Decision Trees

Consider the Decision Tree in Fig. 3.

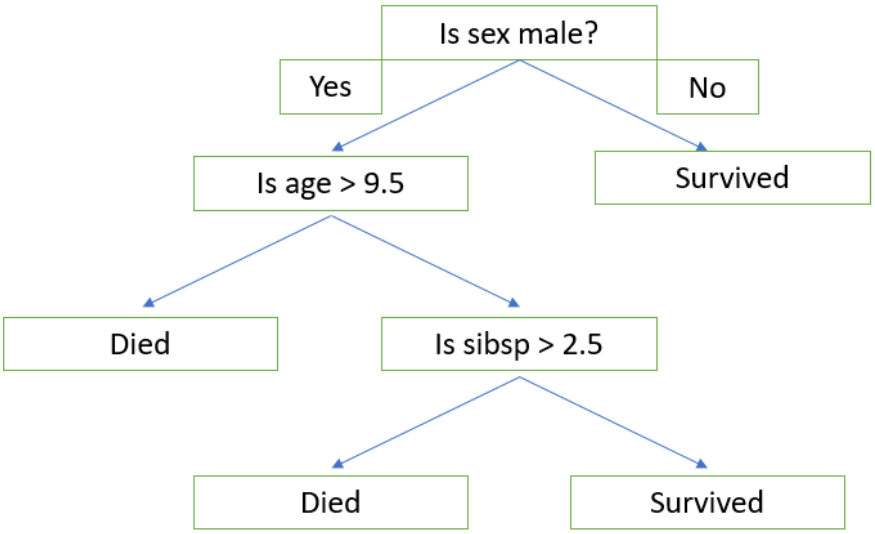

Figure 3: Decision Tree Classifier trained on the Titanic Dataset

Fig. 3 represents a Decision Tree Classifier trained on a (well-known) machine learning dataset named the Titanic dataset. In it, passengers of the titanic are represented by features such as gender, age, and number of siblings/spouses (the `sibsp` feature). In general, a decision tree classifier

is trained using the **Classification and Regression Tree (CART)** algorithm. The algorithm splits the training data in two subsets using a single feature and a threshold for that feature. e.g. (is sex = M?). The CART algorithm chooses these features and thresholds by computing which of the features (and corresponding threshold) can divide the dataset into the purest subsets (weighted by size). In this context, 'pure' refers to a subset of the data in which a maximal amount of the datapoints are of the same Target attribute (i.e. alive/dead). Concretely, the Decision Tree classifier learnt that to separate the dataset into the purest and largest subsets, the first discriminating feature should be sex. For each of the produced subsets, this algorithm will recurse and perform the same operations to split the corresponding subsets into further subsets.

2.3.7 Random Forests

Random Forests are an ensemble method of classification. **Ensemble methods** are classifiers which utilize different classification algorithms on the same training data. The separate classifiers classify the testing data, and the resultant classifications are aggregated to produce a more confident classification.
Specifically, Random Forests are ensembles of Decision Trees. In essence, each decision tree is trained on either:

1. A random subset of data
2. A random subset of features
3. A random subset of features and data entries

This logic can be summarized as the *'wisdom of the crowd'* . Often, aggregating the predictions of a group of weak classifiers (in this case, each tree being trained on a incomplete subset of the dataset), will result in better predictions than with the best individual predictor. [3]

2.3.8 Scripting

After exploring the data provided by HC and constructing some informative features, I trained a powerful classification and ranking machine learning framework named LightGBM on the training data. The scope of this report does not allow for an explanatory analysis of LightGBM, but I

have learnt that the classification algorithm is based on using a technique called Gradient Boosting on Decision Tree Algorithms. Next, I adapted a script from [4] and constructed a Jupyter notebook which:

1. Preprocesses the data from HC
    - Sterilizes any missing / outlying entries in the data
2. Constructs new features based on aggregations of existent numerical features
3. Trains a LightGBM classifier on the augmented dataset

The construction of this notebook results in a clear documentation of the machine learning pipeline (from receiving the data to the production of classification model). Therefore, this script allows future potential users to reutilize the code, and adapt it to future usecases and machine learning projects.

2.3.9 Results

After training the LightGBM classifier, I submitted the testing data to be classified. This classified testing set was then submitted to the Kaggle servers to be rated. I assume that this rating is based on HC comparing the 'correct' classifications of the testing data (1 or 0 for each loan applicant) to the classifications produced the LightGBM model. Following submission of the classified testing set, a true positive/false positive proportion is calculated for the classification of the model, and this rate constitutes the rank of my submission.
The highest such rank I achieved (using the LightGBM model) was 0.792 accuracy. To reiterate, this meant that the proportion of the true positive rate to the false positive rate of the LightGBM classifier was **79.2%**. In comparison, the winning teams (at the time of writing this report) achieved a grade of **81%**.
In addition, Fig. 4 presents a table of feature importances generated during the training of the LightGBM model. At a high level, these features can be thought of as the features which are most discriminatory in the dataset. i.e. They are the features which 'split' the dataset into the purest subsets.

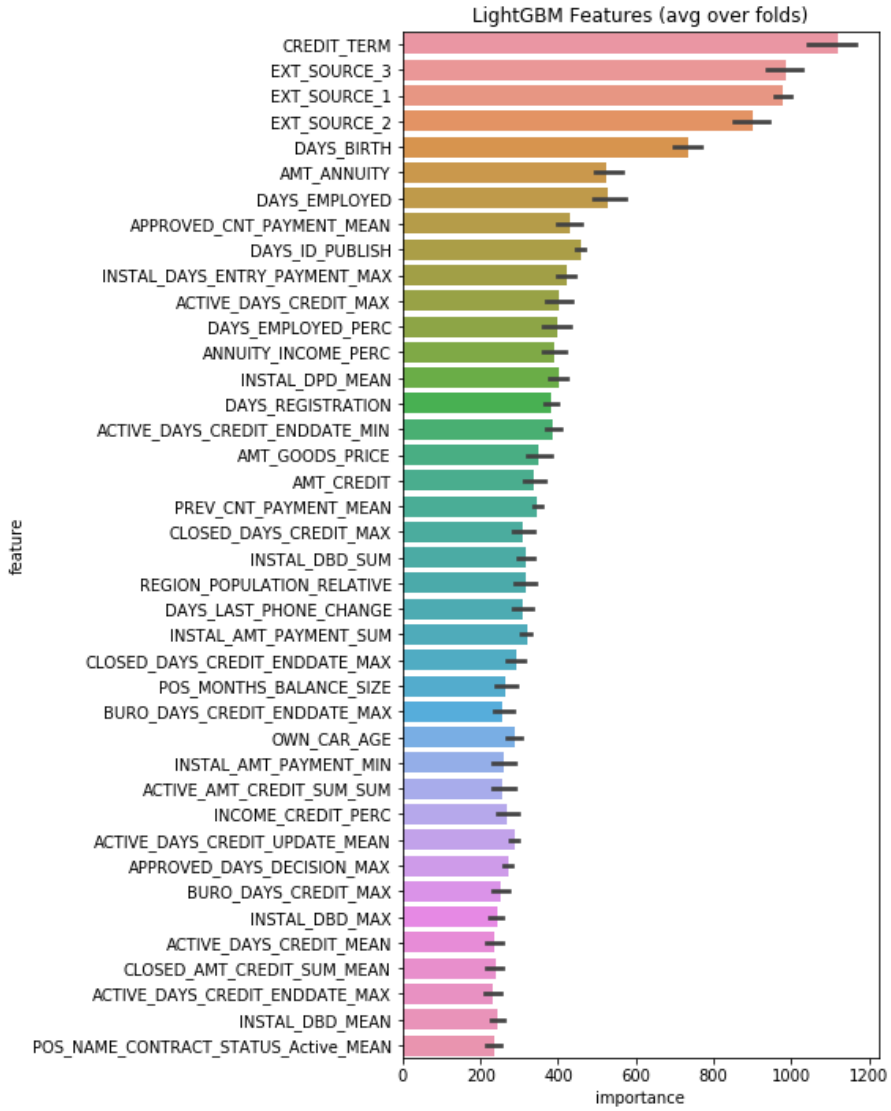

Figure 4: Feature importances generated the trained LightGBM classifier

2.3.10 Discussion

The difference of 1.8% between the submission metric of the LightGBM classifier and the winning algorithms (at the time of subsmission) is quite informative. From this minute difference, I deduced that these competitions are won by algorithms and models which are used in succession of one another

(similar to ensemble methods) and fine-tuned by machine learning and Data science experts.

2.4 Challenges and solutions

2.4.1 Challenges

One of the main challenges I faced during this project was the level of inexperience I had with machine learning and data science. Faced with the level of expertise shown by the top teams in the competition, I quickly realized the immense breadth of knowledge that the field of machine learning contains. In the past, I had been interested in the applications of machine learning algorithms to solve novel problems, and had attempted to educate myself with online tutorials.
Luckily, the resurgence of popularity of machine learning in the computer science and tech industries has produced many understandable and approachable resources such as online tutorials, courses, and textbooks. These provide both an introductory overview to the algorithms and techniques in machine learning projects, as well as, in-depth analyses of the cutting edge technologies which, in some cases, win Kaggle Data Science competitions.

2.4.2 Solutions

When following some of the tutorials available from the Kaggle community, I quickly realized that many of the algorithmic details important for understanding the methodology of machine learning classification models are codependent on knowledge of other basic computer science, statistical, and mathematical subject matter. That is, when faced with an algorithm with which I was unfamiliar, I had to refer to its implementational definition (which was usually based on other simpler, but related, concepts) and refresh/relearn the information associated with it.
For example, before I could familiarize myself with Random Forests, I had to refamiliarize myself with Decision Trees. This meant that I had to constantly refer to educational aids (as referenced above) during the development of this project. I believe that this constant familiarization has instilled in me a solid understanding of the basic foundations of machine learning.

3 Reflections on Work Experience

3.1 Contributions

I had been very excited to participate in this competition, but quickly realized that the domain of Data Science Competitions is filled with expert computer scientists utilizing complex algorithms to answer difficult business questions. Using such resources as [3], I gained a foundational understanding of the basic toolkit possessed by any successful data scientist. With this knowledge, I created a reusable machine learning framework for Apption. In addition, the EDA I performed produced a couple of key features which I empirically demonstrated to be informative in the evaluation of a candidate in terms of future loan repayment delinquency. Specifically, I had found that these features:

1. The total length of credit repayment term (i.e. how long will the candidate take to fully repay the loan)

2. The Mean Days Past Due (DPD) value averaged over past loans (if applicable) (i.e. in the case where candidates applied for previous loans, did they incur any DPD, and if so what was each loan's maximal DPD value for each loan?)

were both informative features that improved the performance of the classifiers I had experimented with. Concretely, this meant that Apption had gained that knowledge, and when approaching possible clients with a value proposition regarding the hypothetical implementation of a Loan Analytics Product (i.e. an algorithm/model which could classify loan applicants with respect to risk factors), it would be clear that adding these features to any data model can provide important information about loan applicants.

3.2 Relation to academic studies

As a Cognitive Science student, I had been interested in the potential of Machine Learning and Artificial Intelligence to inform human decision making from my first introduction to the topic. Specifically, I had been interested in the prospect of training Machine Learning models to analyze and produce art (specifically, music) and create models which could generate novel musical pieces. [5] I believe that the prospect of this endeavour could:

1. Inform us about the creative process itself in humans. That is, training a machine learning model to compose Bach music entails a very different form of Exploratory Data Analysis. (What are the features which are important in making a Bach piece sound like Bach?)

2. If we are able to develop algorithms which are able to imitate (and perhaps, emulate) an innately human creative process, we'll gain a deeper understanding of the 'inner workings' of such a process. That is, training a machine learning model to be *creative* can surely shed some light on what *creativity* is.

Finally, in retrospect, I believe that enrolling in the COMP4107 Neural Networks course before my Co-op work term would've provided me with a background in Deep Learning which would have allowed to explore different avenues of model training surely relevant to the objective of this project.

3.3 Career Development

My participation in this Kaggle competition has provided me with an essential foundation on which to continue accumulating the technical and practical knowledge required to succeed in the technological industry. During my studies in the Institute of Cognitive Science in the Cognition and Computation specialization, I had become increasingly interested in the prospect of machine learning and artificial intelligence in optimizing human decision making (whether business oriented or day-day). As this was my first Co-op work term, I had applied to many positions with an emphasis on Data science. I believe that after completing this project, I now possess the skillset to make me a viable and competent candidate for these types of technical occupations.

4 Summary

This report began by covering the organizational framework at Apption. Next, the objective of the project which was tasked to me - a Kaggle competition concerned with developing an algorithm to predict how capable a loan applicant is in repaying a loan - is explained. I described the basic evaluation metric, technical environment, and data schema which is commonly utilized in Supervised Classification tasks. The machine learning project

pipeline was explained: from the introductory stages of Exploratory Data Analysis, to training the machine learning model (and specifically, which types of model(s) I had chosen to use). Lastly, the results of the model training are presented, and I reflect upon the challenges I faced in completing this project, as well as the contributions it may have had to Apption's business goals, and the development of my career and academic studies.

4.0.1 Future Work

As was briefly mentioned, I believe that the addition of a neural network to the pipeline of this project may have produced a higher score in the competition. I base this belief on certain discussions in the forums of Kaggle, where it was mentioned that a deep neural network can be used to reduce the dimensionality of a high-dimensional dataset (which this competition possessed). It was proposed that running an algorithm such as LightGBM on a (reduced dimensional) dataset will achieve a higher AUROC rating. I therefore suggest to investigate the possibility of improving classification accuracy using a neural network.

References

[1] http://www.apption.com/about/

[2] Koerhrsen, Will. https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

[3] Géron, A. (2017). Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems. O'Reilly Media, Inc.

[4] Aguiar. https://www.kaggle.com/jsaguiar/updated-0-792-lb-lightgbm-with-simple-features

[5] Hadjeres, G., Pachet, F., & Nielsen, F. (2016). Deepbach: a steerable model for bach chorales generation. arXiv preprint arXiv:1612.01010. https://arxiv.org/abs/1612.01010

5 List of Abbreviations

1. AI - Artificial Intelligence
2. AUROC - Receiver Operating Characteristic Area Under the Curve
3. BI - Business Intelligence
4. EDA - Exploratory Data Analysis
5. FI - Financial Technology
6. HC - Home Credit
7. ROC - Reciever Operating Characteristic (curve)