**Problem Statement**

The problem to be solved in the project is actually a machine learning problem. There are more and more people registering for a credit card to use money in advance, which brings the bank a question that whether their credit card customers are reliable to return the money in time. Thus the need for evaluating credit card customers’ ability to refund in time is given birth. In the evaluation processing, there are several factors to take in account according to the competition description, such as delinquency variables, spend variables, payment variables and so on. Factors will be taken together to put into evaluation models and come out a result same as the real condition. Additionally, metric M is given out, which is influenced by both Normalized Gini Coefficient G and default rate captured at 4%, D, according to the competition description.

This project is seemingly aimed to evaluating customer’s credit, however, the knowledge and technology behind this is tightly relative to machine learning and its models. It is because since every bank owns large amount of client information, including their credit record and also their consumption habits(spending, refund, income and so on). By putting these present data into training and testing models, a useful model is probably made out for credit scoring. And this method has been proved valid. Actually, different models in machine learning have various functions and applications nowadays, for instance, when talking about what the project is going to solve, there are also other similar finance problems, such as to predict future stock tendency, to evaluate the value of a house by analyzing given data, to detect fraud and cheating and so on. Apparently, successfully applying machine learning to those things largely protects people’s transaction safety, gives advice to investors and protects banks from dishonest customers, which is of great importance to the people’s wealth and finance operation. In addition, machine learning evaluation models are utilized also in medical area, mobile phone apps recommending algorithms, etc. With the help of evaluation models, we human could predict future more accurately on a big data basis. When confront with the same credit evaluation problem, a well-trained model could help to decide whether the customer have access to spending money ahead of time. By doing so, banker and financial experts can save a lot of time.

**Application**

From a relatively narrow aspect only taking credit scoring into consideration, banks are indeed using machine learning models to evaluate customers, and so that decide their credit scores and lending risks. It can be seen as a way bank avoid their risks of giving loans to those risky people under a low credit level.

A typical example of using machine learning to give customers credit scores instead of assigning bankers to do so is the Zest.ai Corporation. They exhibit their AI products on the website, explaining their function is to evaluate credit level using hundreds of variables to build up relationships, thus decide whether or not the applicant is a compliant candidate for loans or credit card. (from their website: [MMS Build (zest.ai)](https://www.zest.ai/product/mms-build)). In addition, they make the credit score easy to understand by using visible app design. Credit tendency can be seen clearly. Things like logistic regression, big data and credit score frequently appear and almost become inevitable factors in their machine learning products.

More and more companies invest resources into machine learning, by building and training models, they put forward various financial tools helping finance institution avoid and reduce risks of lending money to bad credit customers.

Although lots of companies are doing this, credit scoring by AI is still a small part of financial usage of machine learning. In fact, machine learning and well-trained models are also needed in fields like entrepreneur risk management and financial crime analytics. A software IDK named Simudyne realize such functions and provide related service. A quotation from their product modelling is as following ---- *The modelling core uses objects called agents to mirror the real world at every level of detail. These agents communicate with each other by sending messages. The agents and messages mimic real world interactions, as well as the effects of those interactions over time. Agents are defined by the data they hold and the actions they take, either internally or when interacting with other agents. More on agents, networks and interactions can be found in Agents & Network/Graph.* As it has described, the core modelling is on an agent basis.

When it comes to societal significance of researching machine learning in finance area, from where I am standing, each small progression of financial machine learning could have an important impact on people’s daily finance activity security, financial institutions management and even a country’s finance safety. For instance, with the assistance of financial models of machine learning, people are more likely to pick up a investment product which fits their needs most and earns most successfully. Machine learning models could analyze both hundreds of people’s financial habits and finance product maximum benefits to recommend the best product to people. Of course, it is necessary for financial institutions like public banks to monitor people’s trade records. Also, analyzing former data by machine learning models, it is largely easier to recognize fin-crime and other malice financial behaviors. As we all know, if the recognition work is assigned to a normal banker or expert, it will take multiple time and resource to analyze, sometimes wrong analysis are caused by lacking of data or misunderstanding of statistic result. In conclusion, utilizing machine learning models in finance not only save time and resource but also could provide more accurate analysis of financial questions. Thus, things like protecting individual security, giving investment advice, making risk evaluation and protecting finance safety can be done by machine learning, which is thus very important to people.

In this topic, while the models to evaluate credit scores attracts me most, it makes me confused as well. It is amazing to use hundreds of, even thousands of data to train a model, and through trainings the model people are able to judge the test condition in a high accuracy. Data processing is of great important, because useless and disturbing data sometimes exist and need to be processed or removed. If a un-monitored training model is adopted, it means model study all the data by itself without human reminding what kind of data is positive and others are not. So the formula and frames for the model is the core technology. Thus, I am interested in how the model’s self training works and whether different training methods leads to different results.

**Main steps:**

Before using the data to train models test them, it is necessary to preprocess the dataset. As we all know, each customer’s account data probably differ from each other, and we cannot guarantee all of the customers who register to get a credit card own the complete account data. For example, students may lack their income records, and someone might have no delinquency records, while other people have comparably complete data. So in order to make up for those missing data, we are able to try out several methods, such as K-means clustering, Regression, Model Completer and so on. The main principle of these methods is finding out some new data to fulfill the missing data position. While directly grab data from other customers’ account or adopt random data, it is more reasonable to calculate the new data using the methods above.

There are always abnormal data as well. After using methods like detecting whether the data is out of three times standard deviation, abnormal data are found out. Then it’s time to utilize the similar methods for missing data and calculate new data to replace the abnormal ones.

Then it is necessary to figure out data features and lower the data dimension. There are two mainly used methods. One is PCA(Primary Component Analysis), and the other is LDA(Linear Discriminant Analysis), one apparent difference between them is whether or not we could monitor the process. Primary component analysis uses mathematics calculation and gain the data matrix’s eigenvalue, lowering the matrix dimension.

Last but not least, in the preprocessing process, filtering the features is also crucial. How can we select the useful features while discarding comparably useless ones? It is impractical to put all the features into training models. Thus, it is practical to applying judging standards, such as variance, relativity and so on. For example, when using variance standard, the feature deviate from other features most probably satisfy the selecting standards.

In addition, wrapper methods and embedded methods can also be used.

After preprocessing the data and select out the features, machine learning programmers are able to process the data and train the model. According to my teammates research, models as logistic regression, support vector machine(SVM), ensemble learning, ANN(artificial neutral network) are practical schemes. After viewing them on the internet, I preliminarily know about their concepts. For ensemble learning, the main spirit is combining several machine learning learners by certain combination strategies to get a sturdy learner. For support vector machine, it reflect real space spots into other spaces and aims at drawing a line to distinguish different types of spots.

In conclusion, to try different models and decide the best learner is the main task of the project. The best learner is expected to distinguish whether the customer could return the money in time.

**Initial list of papers and open source projects to study**

* <https://medium.com/henry-jia/how-to-score-your-credit-1c08dd73e2ed#:~:text=%20Credit%20Scoring%20with%20Machine%20Learning%20%201,the%20feature%20selection%2C%20I%20replace%20the...%20More%20>
* [Credit Scoring Project using Machine Learning | Risk Modelling | Logistic Regression | ML Education - Bing video](https://www.bing.com/videos/search?q=credit+scoring+machine+learning&docid=608013343469363882&mid=D6FE2236FA05A917D28ED6FE2236FA05A917D28E&view=detail&FORM=VIRE)
* <https://www.zest.ai/resources/winning-internal-support-for-ai-with-todd-peeples-of-all-in-credit-union>
* <https://towardsdatascience.com/machine-learning-in-finance-why-what-how-d524a2357b56>
* <https://algorithmxlab.com/blog/barclays-pushing-into-artificial-intelligence-to-manage-risk/>
* *Book: Machine Learning in Finance:*

<https://link-springer-com.ezproxy.bu.edu/book/10.1007/978-3-030-41068-1>

* *Paper: Financial credit risk prediction in internet finance driven by machine learning:*

<https://link.springer.com/article/10.1007/s00521-018-3963-6#citeas>

* *Article: Application and Comparison of Multiple Machine Learning Models in Finance*

[Application and Comparison of Multiple Machine Learning Models in Finance (hindawi.com)](https://www.hindawi.com/journals/sp/2022/9613554/)

* *Paper:Consumer credit-risk models via machine-learning algorithms*

[Consumer credit-risk models via machine-learning algorithms - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0378426610002372?via%3Dihub)

* *Competition champion’s open project:*

[jxzly/Kaggle-American-Express-Default-Prediction-1st-solution (github.com)](https://github.com/jxzly/Kaggle-American-Express-Default-Prediction-1st-solution)

[American Express - Default Prediction | Kaggle](https://www.kaggle.com/competitions/amex-default-prediction/discussion/348111)

* *Video on credit score using ML:*

[Credit Scoring Project using Machine Learning | Risk Modelling | Logistic Regression | ML Education - Bing video](https://www.bing.com/videos/search?q=credit+scoring+machine+learning&docid=608013343469363882&mid=D6FE2236FA05A917D28ED6FE2236FA05A917D28E&view=detail&FORM=VIRE)