Lending Club Credit Default Prediction: Analysis with Machine Learning / Deep Learning Models

Bowen Zhang
bz2461@columbia.edu
Columbia University
MRKT B9653: Artificial Intelligence
George A. Lentzas
2023/05/05

Abstract

The research paper investigates the application of deep learning and machine learning models for predicting loan default in LendingClub, a leading peer-to-peer lending platform in the United States. A dataset containing loan features and borrower characteristics are collected and preprocessed for model training and evaluation. Two machine learning models and a feed-forward neural network model are designed to compare their performance in predicting loan default. The results demonstrate that the deep learning model outperforms the baseline machine learning models, achieving higher f1 and accuracy scores.

1. Background

LendingClub is a leading online marketplace lending platform that connects borrowers with investors. The platform has facilitated over $60 billion in loans since its inception in 2007, making it one of the largest peer-to-peer lending platforms in the United States. The lending process on LendingClub involves borrowers submitting loan applications, which are evaluated based on various criteria, including credit score, income, employment status, etc. Investors can then choose to invest in a portion of the loan and earn interest based on the borrower's repayment.

Despite its success, the lending industry is facing the challenge of managing default risks. Loan default occurs when a borrower fails to make timely payments on their loan, resulting in losses for investors. Predicting loan default is crucial for lenders to mitigate risk and make informed lending decisions. Traditional credit risks models, such as logistic regression and decision trees, have been widely used in the lending industry for predicting loan default. However, these models have limitations in capturing complex non-linear relationships and high-dimensional data.

Recently, deep learning models have shown great promise in various applications, including image recognition, speech recognition, and natural language processing. In the finance industry, deep learning models have been used for credit scoring, default detection, and risk management. The advantages of deep learning models include their ability to learn complex patterns and relationships in large-scale data and their ability to automatically extract relevant features without manual feature engineering.

Given the advantages of deep learning models and the importance of predicting loan default, this research paper aims to investigate the effectiveness of deep learning models for predicting loan default for LendingClub. Specifically, this paper proposes to design and compare various machine learning and deep learning models, namely Decision Tree, Random Forest, and feed-forward neural network. The findings of this study can provide insights into the application of machine learning and deep learning models in the lending industry and contribute to the development of more accurate and effective credit risk models.

2. Dataset Overview

The data were collected from the LendingClub's website that contains loan information from 2007 to 2017. It is a popular dataset that has been widely used and analyzed by many data scientists or researchers in the finance industry to find appropriate models to help LendingClub prevent loan default. The dataset contains 400 thousand rows and 28 columns, with demographic features including employment title, home ownership, annual income, and loan features such as loan amount, interest rate, issue date, verification status. The target class is loan status, which indicates whether that loan has been defaulted or not (0: Charged Off, 1: Fully Paid).

0	loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
1	term	The number of payments on the loan. Values are in months and can be either 36 or 60.
2	int_rate	Interest Rate on the loan
3	installment	The monthly payment owed by the borrower if the loan originates.
4	grade	LC assigned loan grade
5	sub_grade	LC assigned loan subgrade
6	emp_title	The job title supplied by the Borrower when applying for the loan.*
7	emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
8	home_ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
9	annual_inc	The self-reported annual income provided by the borrower during registration.
10	verification_status	Indicates if income was verified by LC, not verified, or if the income source was verified
11	issue_d	The month which the loan was funded
12	loan_status	Current status of the loan
13	purpose	A category provided by the borrower for the loan request.
14	title	The loan title provided by the borrower
15	zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
16	addr_state	The state provided by the borrower in the loan application
17	dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
18	earliest_cr_line	The month the borrower's earliest reported credit line was opened
19	open_acc	The number of open credit lines in the borrower's credit file.
20	pub_rec	Number of derogatory public records
21	revol_bal	Total credit revolving balance
22	revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
23	total_acc	The total number of credit lines currently in the borrower's credit file
24	initial_list_status	The initial listing status of the loan. Possible values are – W, F
25	application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
26	mort_acc	Number of mortgage accounts.
27	pub_rec_bankruptcies	Number of public record bankruptcies

Figure 1. Dataset Overview

	loan_amnt	int_rate	installment	annual_inc	dti	open_acc	pub_rec	revol_bal	revol_util	total_acc	mort_acc	pub_rec_bankruptcies
count	396030.000000	396030.000000	396030.000000	3.960300e+05	396030.000000	396030.000000	396030.000000	3.960300e+05	395754.000000	396030.000000	358235.000000	395495.000000
mean	14113.888089	13.639400	431.849698	7.420318e+04	17.379514	11.311153	0.178191	1.584454e+04	53.791749	25.414744	1.813991	0.121648
std	8357.441341	4.472157	250.727790	6.163762e+04	18.019092	5.137649	0.530671	2.059184e+04	24.452193	11.886991	2.147930	0.356174
min	500.000000	5.320000	16.080000	0.000000e+00	0.000000	0.000000	0.000000	0.000000e+00	0.000000	2.000000	0.000000	0.000000
25%	8000.000000	10.490000	250.330000	4.500000e+04	11.280000	8.000000	0.000000	6.025000e+03	35.800000	17.000000	0.000000	0.000000
50%	12000.000000	13.330000	375.430000	6.400000e+04	16.910000	10.000000	0.000000	1.118100e+04	54.800000	24.000000	1.000000	0.000000
75%	20000.000000	16.490000	567.300000	9.000000e+04	22.980000	14.000000	0.000000	1.962000e+04	72.900000	32.000000	3.000000	0.000000
max	40000.000000	30.990000	1533.810000	8.706582e+06	9999.000000	90.000000	86.000000	1.743266e+06	892.300000	151.000000	34.000000	8.000000

Figure 2. Summary Statistics

3. Exploratory Data Analysis

I first checked the distribution of the target class, as visualized in the chart below. In the dataset, there are 19.6% of the clients defaulting, versus 80.4% of those who do not default, indicating a class imbalance.

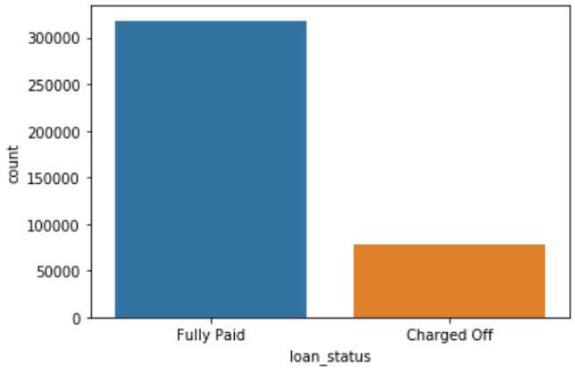

Figure 3. Class Distribution

Then, I checked the correlation among the features, excluding the ID column. From the correlation heatmap below, I noticed that the loan amount almost has perfect correlation with installment.

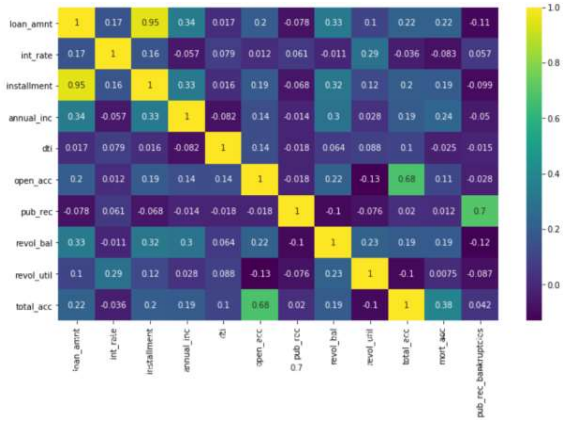

Figure 4. Correlation Heatmap

I also explored the relationship between loan grades and default status. Clearly, from the charts, lower grade loans are more likely to default than higher grade loans.

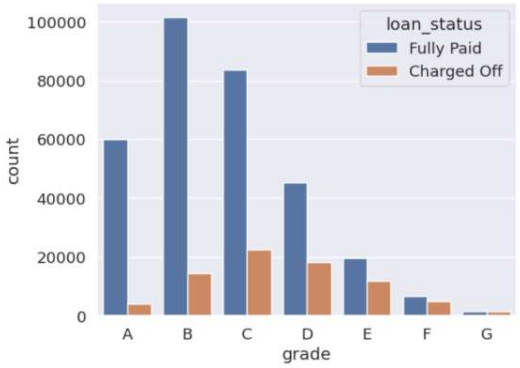

Figure 5. Loan Status vs Grade

Lastly, I examined the correlation between loan status and other variables, and ranked them in ascending order. As the graph shown, the interest rate has the most negative correlation while mort_acc has the most positive correlation. This infers that bonds with higher interest rates are more likely to default than the ones with low interest rates. On the other hand, borrowers with more mortgage accounts are less likely to default.

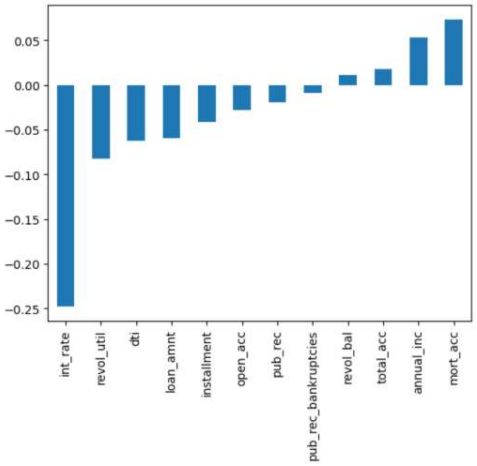

Figure 6. Loan Repaid Correlation

4. Preprocessing

Data preprocessing is a critical step in any data analysis project, which involves transforming, removing, and cleaning of the original dataset to make it more robust for further analysis.

I first explored all columns that have missing values

in the dataset, and decided to drop the ones with too many missing fields. For the fields with few missing values, I either dropped the rows or applied imputation methods to fill in the dataset. I also came up with a methodology to fill in missing values for "mort_acc" by taking the average of the total numbers of accounts, which is a highly correlated column.

loan_amnt	0
term	0
int_rate	0
installment	0
grade	0
sub_grade	0
emp_title	22927
emp_length	18301
home_ownership	0
annual_inc	0
verification_status	0
issue_d	0
loan_status	0
purpose	0
title	1755
dti	0
earliest_cr_line	0
open_acc	0
pub_rec	0
revol_bal	0
revol_util	276
total_acc	0
initial_list_status	0
application_type	0
mort_acc	37795
pub_rec_bankruptcies	535
address	0
loan_repaid	0

Figure 7. Missing Values

Then, I transformed the categorical variables into dummy variables using One-Hot-Encoding. This allows me to use these non-numeric fields as predictors. I also normalized the dataset with StandardScaler to ensure that all predictors are within the same range.

5. Classification Metrics

The goal of the project is to predict whether the loan status given all other information (0: default, 1: fully paid). For metrics to evaluate model performance, I chose accuracy and auc score. Since the dataset is highly imbalanced, I also be examined the f1 score, which is a harmonic mean of recall and precision. The metrics formulas are defined as below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{F1 Score} = \frac{2TP}{2TP+FP+FN}$$

6. Model Building

In this project, I tested out a couple of different machine learning models (Decision Tree and Random Forest) and compared their performance with a feed-forward neural network model using metrics such as accuracy and f1 score. Decision Tree and Random Forest are two machine learning models that have been widely applied in the field of default predictions, so I used them as baseline models to see the performance.

All models were trained with Google Colab T4 GPU, as I could leverage its parallel processing capacity to handle complicated deep learning tasks.

6.1. Decision Tree Classifier

Decision Tree Classifier makes prediction by recursively splitting the data based on the features that provide the most information gain. To get the best set of hyperparameters, I conducted cross-validation with GridSearchCV. The resulting confusion matrix is shown in Figure 8.

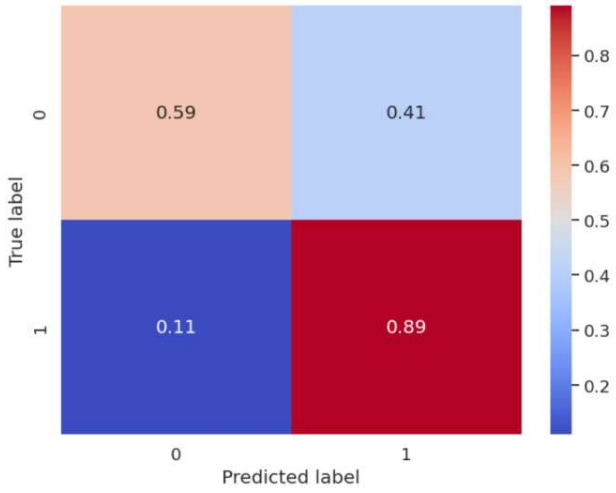

Figure 8. Confusion Matrix of Decision Tree Classifier

6.2 Random Forest Classifier

To improve from the baseline model, I also tried to fit a Random Forest Classifier. It is an ensemble method that builds a set of decision trees on random subset of the training data and features and averages

the results to produce a final prediction. As a result, random forest usually yields more robust and accurate models.

I have also tuned hyperparameters using GridSearch with cross validation and the resulting confusion matrix is shown in Figure 9.

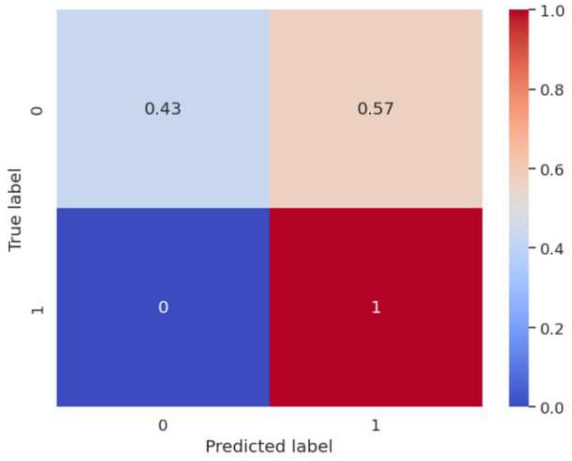

Figure 9. Confusion Matrix of Random Forest Classifier

6.3 Artificial Neural Network

The neural network model used in this study is a feed-forward neural network with 3 levels of hidden layers. The input layer of the network consists of features from the dataset, and the output layer consists of a single node that represents the predicted probability of loan default. This output is then transformed into a binary result by the sigmoid activation function. Furthermore, the activation functions used in the hidden layers is the rectified linear unit (ReLU).

To optimize the model, I used the Adam optimizer with a learning rate of 0.001. I also used the binary cross-entropy loss function to evaluate the model's accuracy and AUC scores.

To mitigate the problem of overfitting, I dropped 10% of the neurons after each hidden layer, and implemented an early stopping mechanism with patience of 10 to monitor the model performance. I have also passed in a batch normalization layer after dropout to ensure the neural network training is faster and more stable.

I have tried to implement different activation functions in the hidden layers, such as LeakyReLU, Tanh, and Sigmoid, but none of them could beat the performance of ReLU.

After running the model, EarlyStopping has interrupted the training process at the 11th iteration. The resulting confusion matrix is shown in Figure 10 and the graph that demonstrates training and validation accuracy is shown in Figure 11.

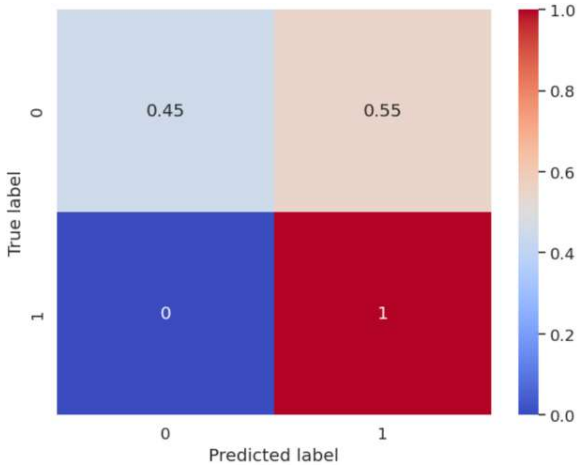

Figure 10. Confusion Matrix of ANN

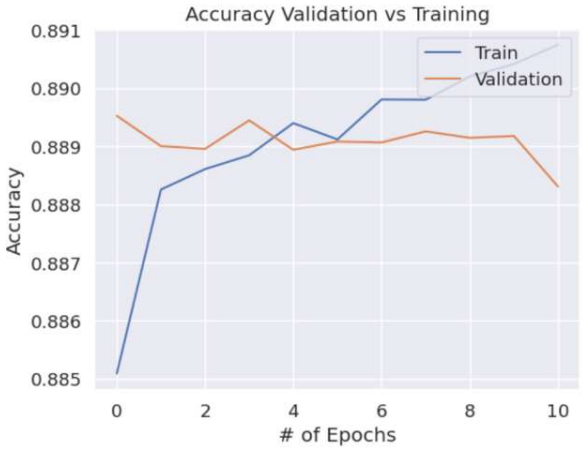

Figure 11. Train vs Validation Accuracy

7. Result Discussion

To compare the performance among the three models, I made a table to demonstrate their accuracy, f1, and AUC scores (Figure 12).

As shown below, the neural network model has the best performance for both accuracy and f1, and has slightly lower AUC score than the baseline decision tree model. The accuracy and f1 scores for random forest are close to the neural network model, but it has a poorer AUC performance.

Since the dataset is imbalanced, f1 score might be the best metric to reflect the model's real performance. Therefore, I would prefer to use ANN in loan default prediction for LendingClub.

The result has shown that the deep learning model would generally have better performance than machine learning models. There could be three reasons based on this result. First, the additional hidden layers can capture the non-linear relationships between the input features and the target variables, which may not be well captured by the traditional machine learning models. This can be especially useful in loan default prediction as the relationships between variables may be complex and non-linear. Second, the dataset in real life could be noisy, and deep learning models are often more robust to the noisy data as they can learn to filter out irrelevant information and focus on the most important features. Last but not least, the LendingClub dataset contains 400 thousand rows and 28 columns, which is too large and computational-heavy for machine learning models. On the other hand, deep learning models can be trained on large datasets with many features and leverage the power of GPU to boost computational speed.

Model	Accuracy	F1 score	AUC
Decision Tree	0.829	0.893	0.738
Random Forest	0.888	0.935	0.713
ANN	0.889	0.935	0.722

Figure 12. Model Comparison

8. Conclusion

In conclusion, the research paper aimed to develop and evaluate machine learning and deep learning models for predicting loan default. I compared the performance of Decision Tree Classifier, Random Forest Classifier, and feed-forward neural network models, and found that the neural network model achieved the best overall performance, with an accuracy of 0.89, f1 score of 0.94, and AUC score of 0.72 on the test set.

This result suggests that deep learning models can be effective tools for predicting loan default, especially when dealing with large dataset.

Overall, this research highlights the potential of machine learning and deep learning approaches for improving loan default prediction, which could have important implications for lenders, borrowers, and policymakers. Future research could explore the use of other types of models, such as gradient boosting, or recurrent neural networks, as well as the incorporation of additional data sources, such as alternative credit card or social media data, to further improve loan default prediction.

Appendix

Previous Projects:

1. Yelp Review Prediction, IEOR 4525 Machine Learning for FE & OR, predict star ratings based on user reviews using models including DistilBERT, Naïve Bayes, and XGBoost

2. Capital Quotient Prediction, IEOR 4524 Capstone, iterative fills in each missing values in a survey dataset using Linear Regression and sum up the result to estimate projected CQ score

Reference

1. LendingClub Dataset: https://www.kaggle.com/datasets/wordsforthewise/lending-club

2. Liang, Junjie. "Predicting borrowers' chance of defaulting on credit loans."

3. Tsai, Kevin, Sivagami Ramiah, and Sudhanshu Singh. "Peer Lending Risk Predictor