**Title:** Comparison of Machine Learning Algorithms for Predicting Bank Loan Default

**Authors:** Rishi Dave and Vikki Fiore

**Abstract:**

Like many industries, finance has begun using artificial intelligence and automation technologies to create greater efficiency. Loan companies apply machine learning techniques to approve or deny applicants. This paper touches upon previous research as to which algorithms provide best results for automating the loan application process. Additionally, the benefits and limitations of using artificial intelligence in this manner are discussed. We then experimented on our own and applied various algorithms to a loan dataset. The best results came from using the SVM algorithm.

**Keywords:** classification, machine learning, finance, automation, loans

**Introduction:**

Issuing credit is oftentimes a complex process. Market demands and the clients’ individual circumstances must be taken into account in order to decide if a loan should be approved or denied. This process requires calculating multiple possible scenarios, which over the years have increased in number. It is for this reason that, like many other industries, many in the field of finance have begun turning to automated systems to expedite the loan process in the most efficient possible manner.

Currently, many startups are already exploring more efficient alternatives to established financial services, with larger corporations soon to follow.

**Literature Review:**

The process of approving or denying a loan of a certain amount for a certain client is a complex one. In most cases, lenders turn to established methods for determining creditworthiness. These are often based on a credit system that Simeon Kostadinov refers to as the Five C’s: character, capacity, capital, collateral, and conditions. Character simply refers to the borrower’s credit history and their record of repaying past loans. Capacity compares the client’s debt to income ratio, commonly referred to as DTI. Payments needing to be made on a continuing basis are referred to as debt, and the client’s DTI ratio is used to determine whether enough income will remain for the client to pay the new loan after paying for their existing recurring debts. The next C, capital, is simply the amount of money a borrower puts forward. For instance, a mortgage will require a certain percentage as a downpayment. Collateral refers to an asset that the borrower provides in order to secure the loan. In the previous example of a mortgage, the home itself would serve as collateral. The final C refers to conditions and includes all remaining variables specific to the situation and client. Everything from the conditions of the loan (such as interest rates) to conditions outside of borrowers control (such as the state of the current economy) are included in this category (Kostadinov, S., 2019).

Given the amount of variables and the degree of uncertainty that goes into approving and issuing a loan, automating the process via artificial intelligence and machine learning technologies seems a logical evolution in the field of finance. In his 2017 paper, Bagherpour explores the process of predicting mortgage loan default via machine learning. This paper discusses the process of applying and comparing various machine learning algorithms in order to predict mortgage loan default. A large dataset with over 20 million loan observations was split into test/train categories and used in this research. Non-parametric, nonlinear were found to be, “substantially better than the traditional logit model” (Bagherpour, A., 2017). Additionally, machine learning algorithms allowed for much greater identification of the predictive power of specific variables. The results of this paper found that loan age is the most important predictor of loan default. The research also found this to be true both before and after the 2008 financial crisis (Bagherpour, A., 2017).

Additional research includes that of Addo, Guegan, and Hassani. In their 2018 paper, these researchers compared the performance of 3 machine learning and 3 deep learning algorithms in order to determine whether a company would default on its bank loan or not. The 3 machine learning algorithms used were [random forest](https://towardsdatascience.com/understanding-random-forest-58381e0602d2), [gradient boosting](https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab), and elastic net, which is an extension of logistic regression. The researchers observed that the gradient boosting algorithm outperformed the other two machine learning models as well as all three of the deep learning models.

In another 2018 paper on the applications of machine learning to finance, Sun and Vasarhelyi went further into depth regarding deep learning applications. These researchers applied deep learning neural networks in order to predict credit card deliquesces. They used a 5-layer neural network, which was evaluated on a dataset obtained from a large bank in Brazil, containing 711,397 records of credit card owners, of which 0.92 were delinquent. This deep learning model outperformed existing machine learning algorithms such as [decision trees](https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052), [naive Bayes](https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c) and [logistic regression](https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc) on both on the [F1 score](https://en.wikipedia.org/wiki/F1_score) and the overall accuracy. The researchers did note some drawbacks to deep learning models, though. One main drawback is the fact that governments most often require lenders to present solid and clearly explainable arguments as to why their credit decision should be considered fair and unbiased. Deep neural networks, however, act as black-box models, with much of the calculations occuring “behind the scenes”. Therefore, data scientists find it hard to explain the reasons behind a given output (Sun, T., & Vasarhelyi, M., 2018).

**Approach & Implementation:**

There are many people who are approved for loans that eventually have to get charged off. This means that the loan providers will give up on trying to receive the money. Our goal was to create a classification model able to predict whether or not a candidate will default on their loan or not based on various factors similar to those used in existing research. We set out to experiment with different classification algorithms in order to see which would yield the best results in this situation.

Jupyter Notebook and Python were utilized for cleaning data and preprocessing data, as well as data exploration to understand some of the independent variables (Ex: Credit Score).

**Dataset & Preprocessing:**

The dataset used was obtained from an online repository[[1]](#footnote-0) and contained the following variables:

*Numerical Variables:*

Current Loan Amount, Credit Score, Annual Income, Years in current job, Monthly Debt, Years of Credit History, Months since last delinquent, Number of Open Accounts, Current Credit Balance, Maximum Open Credit, and Bankruptcies.

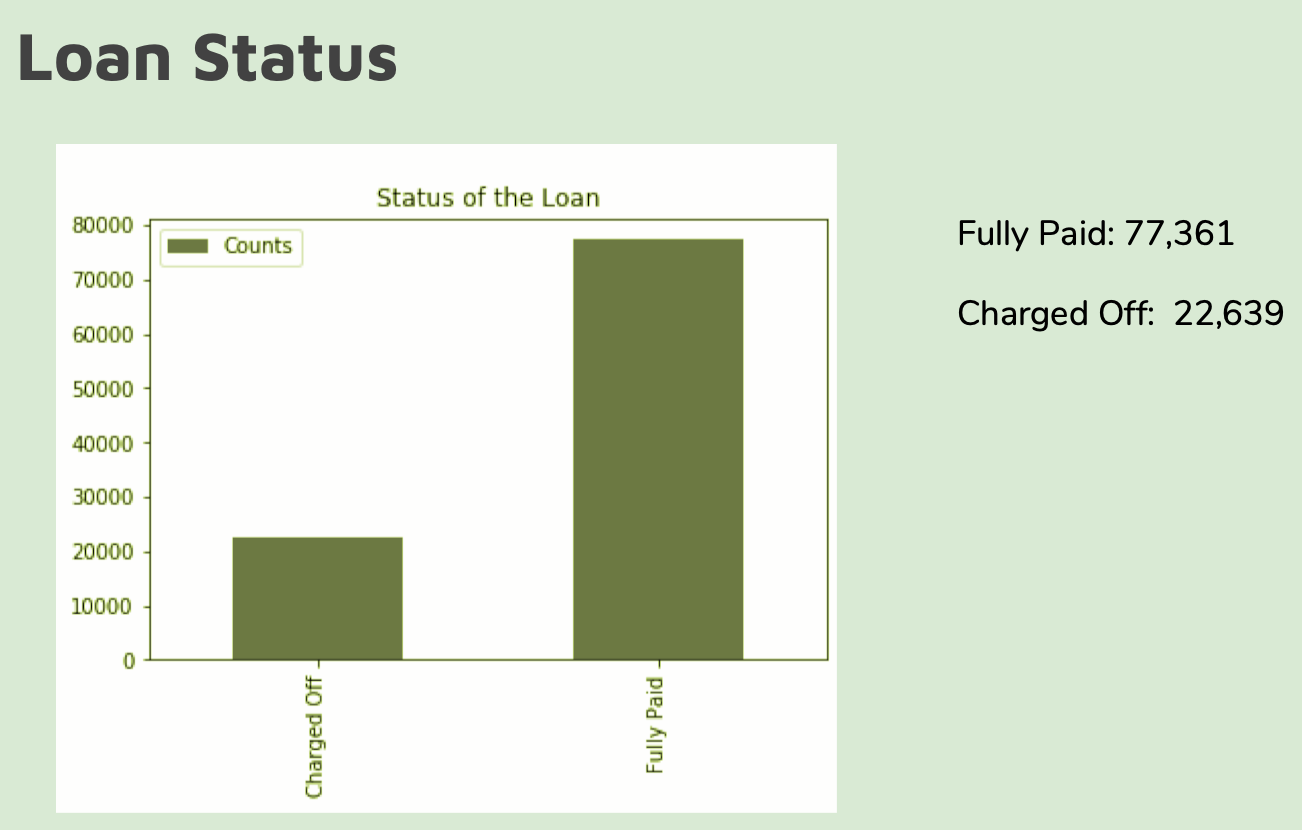
*Categorical Variables:*

Loan ID, Customer ID, Loan Status, Term, Home Ownership, Purpose, Number of Credit Problems, and Tax Liens.

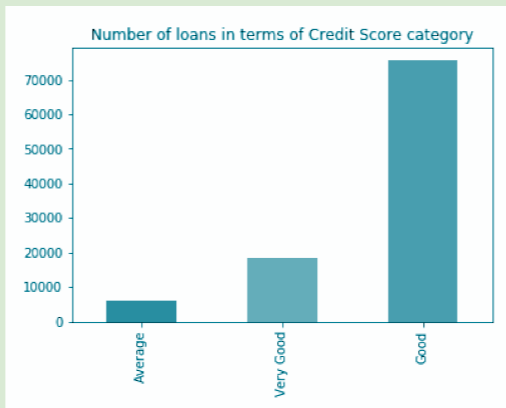
The dataset also contained 100,000 rows and 19 columns. The most important variable that we focused on was the categorical variable of Loan Status. From this variable we created two classes of “Fully Paid” and “Charged Off” for our model.

The dataset was then preprocessed by removing the Loan ID and Customer ID, removing out of range values from credit score (Ex: Credit Score of 1000), removing missing values from Monthly Debt and Annual Income, replacing missing values with average, and assigning bins for Credit Scores as follows:

* + Less than 550 = Poor
  + 550 to 680 = Average
  + 680 to 750 = Good
  + 750 to 800 = Very Good

Next, the dataset was divided into training and testing so that we could train and run different classification models.

*Figure 1: Loan Status*



*Figure 2: Loans by credit score*

**Classification Models & Results:**

We set out with a goal to compare models using a confusion matrix and accuracy in order to choose the model with the best predictive value.

The three main models used were as follows:

*Logistic Regression:*

* Accuracy = 77.2%
* Precision = 77.1%
* Recall = 99.9%

*K-Nearest Neighbor:*

* Accuracy = 74.9%
* Precision = 77.6%
* Recall = 94.5%

*SVM:*

* Accuracy = 77.2%
* Precision = 77.2%
* Recall = 100%

**Conclusions & Future Directions:**

The highest Accuracy and Recall were obtained by using SVM, the highest Precision by using KNN, and the highest. No model seemed to be significantly better than the rest, and all models have very high recall rates. This would indicate that the model is overfitting our data and could be improved by using a distinct testing set and by using n-fold cross validation.

Clearly, the issue of whether to deny or approve a loan for a given client is a very relevant problem faced by many credit-giving firms, all of which are constantly looking for ways to minimize risk and maximize return. Via machine learning algorithms, we are able to predict whether or not a loan candidate will default on their with at least a 75% accuracy, and we will be able prevent a charge-off ¾ times. This model can be used by a bank to better vet their loan applicants and guarantee more Fully Paid loans.

**References:**

Addo, P. M., Guegan, D., & Hassani, B. (2018). Credit risk analysis using machine and deep

learning models. *Risks*, *6*(2), 38.

Bagherpour, A. (2017). Predicting mortgage loan default with machine learning methods.

*University of California/Riverside*.

Kostadinov, Simeon. “The Future of Lending Money Is Deep Learning.” *Medium*, Towards Data

Science, 30 July 2019,

towardsdatascience.com/the-future-of-lending-money-is-deep-learning-61a9e21cf179.

Sun, T., & Vasarhelyi, M. A. (2018). Predicting credit card delinquencies: An application of deep

neural networks. *Intelligent Systems in Accounting, Finance and Management*, *25*(4),

174-189.

1. <https://www.kaggle.com/zaurbegiev/my-dataset> [↑](#footnote-ref-0)