Credit Score Classification

Kyle Jow, Jimmy Nguyen, Kyle Pickle, Jacob Tuttle, Kris Wong

December 2023

Introduction and background

A common problem in the finance and banking industry is assessing the risk involved with lending money to clients. To better assess a borrower's creditworthiness, lenders formulate and assign "credit scores" to their clients, based on preconceived data like income, spending, and credit history. This score can then be used by a bank as an indicator of whether a client will default or otherwise miss a payment on a given loan. The rise of cloud-based computing in the 21st century has made it more advantageous and easier than ever for banks to share a standardized database of credit scores and financial information on their clients. As a result, credit scoring is being used more and more frequently to determine "worthiness" for almost anything – qualifying for mortgages, insurance, and even more nuanced decisions like cell phone plans and determining your employability.

With this increase in credit scoring has come an increased interest in financial literacy among clients as to how to understand and improve one's score. As such, credit standards like FICO and VantageScore have begun giving their clients a way to view an estimation of their credit score, often provided through the bank and credit card services that utilize them. These credit score estimates are predicted in a way that is simple for the client to understand, often broken down into categories like "poor," "standard," and "good" and presented alongside graphs displaying factors like age bracket and income to put them into perspective. This demand for straightforward and transparent credit scoring by governments and clients alike has interestingly compelled banks to consider less intricate or "black box" machine learning models, creating a delicate balance between accuracy and simplicity.

Credit scoring is considered to be one of first and most common instances of machine learning used in the field of economics. The financial sector is historically always one of the first sectors to adopt technological advancements, and with so much data now digitized by banks and numerous complex variables that drive one's credit score, it has become a token example for machine learning and how to not only predict creditworthiness but constantly tune the equations and hyperparameters used to determine it alongside a fluctuating, real-world economy.

Literature review

One of the first fields machine learning techniques were tested in was economics, particularly credit scoring. Researchers have utilized a variety of machine learning algorithms to predict credit scores and risk. Common machine learning algorithms used include logistic regression, support vector machines (SVM), and decision trees. This section contains relevant work by researchers.

Dumitrescu et. al. used a particular model for credit score classification with an improved logistic regression model that has non-linear decision tree effects. They created the penalised logistic tree regression, which predicted credit score more accurately than the benchmark logistic model commonly used in industry. Additionally, they argue that this model preserves the interpretability of logistic regression, an aspect that makes logistic regression popular in industry. Vidovic et.al. compare the performance of various models for predicting the probability of default.

Dataset description and exploratory data analysis

The dataset contains 27 features. The features we will be focusing on in order to classify a person's credit score, which is categorical, into either "good", "standard", or "poor" credit will be features that had the most correlation with credit score according to the heat map generated from this data and features used in real-life credit score assessing. Although the Credit_Utilization_Ratio has little correlation to credit score, this feature is used to assess credit in real life, so we kept it. (Figure 1).

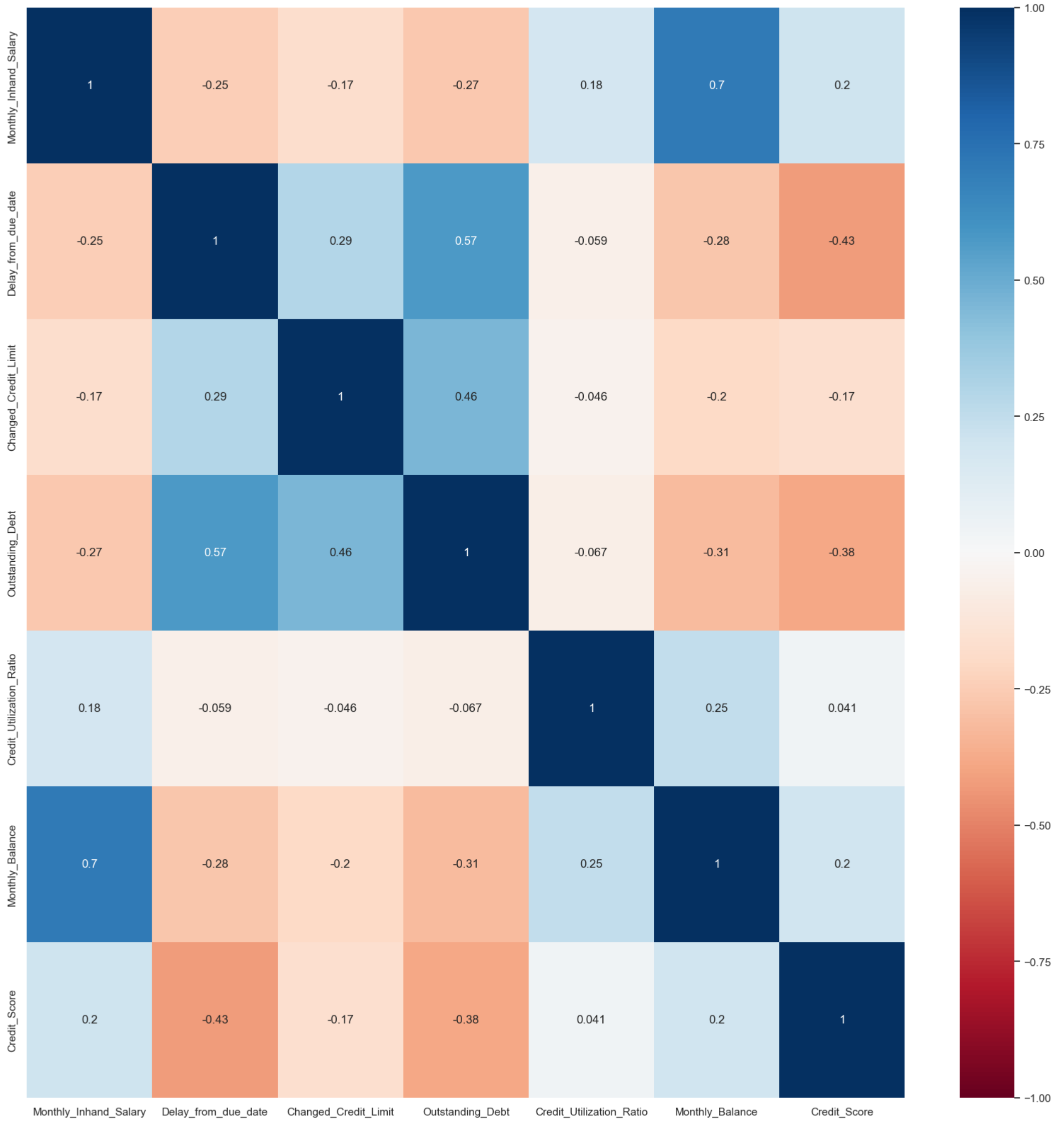


Figure 1: Heat map correlation.

Proposed methodology

From the literature review and the fact that this problem involves multiclassification, we decided to make and compare two SVM models with a linear and RBF kernel and one multinomial logistic regression model. Additionally, we used Streamlit to host our model online.

Experimental results

Below are the classification reports for each of the models:

	precision	recall	f1-score	support
1	0.7	0.48	0.57	2774
2	0.6	0.88	0.72	5064
3	0.56	0.02	0.05	1565
accuracy			0.62	9403
macro avg	0.62	0.46	0.44	9403
weighted avg	0.62	0.62	0.56	9403

Table 1: Classification report for SVM with RBF kernel

	precision	recall	f1-score	support
1	0.62	0.41	0.5	2774
2	0.58	0.87	0.7	5064
3	0	0	0	1565
accuracy			0.59	9403
macro avg	0.4	0.43	0.4	9403
weighted avg	0.5	0.59	0.52	9403

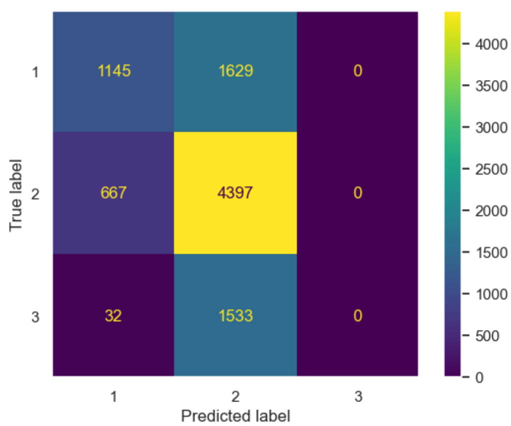
Table 2: Classification report for SVM with linear kernel

	precision	recall	f1-score	support
1	0.63	0.42	0.5	4317
2	0.59	0.82	0.69	8022
3	0.47	0.17	0.25	2632
accuracy			0.59	14971
macro avg	0.56	0.47	0.48	14971
weighted avg	0.58	0.59	0.56	14971

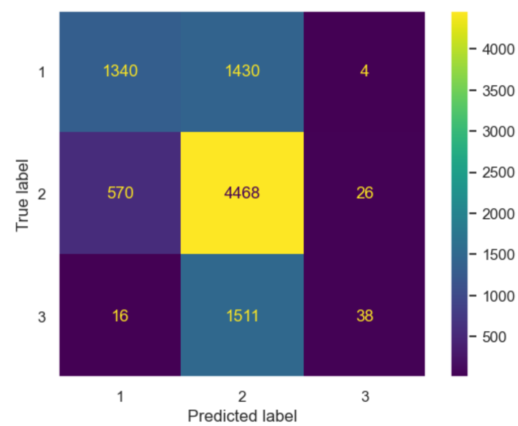
Table 3: Classification report for logistic model with outliers

	precision	recall	f1-score	support
1	0.6	0.37	0.45	3483
2	0.61	0.81	0.69	7642
3	0.5	0.29	0.36	2719
accuracy			0.59	13844
macro avg	0.57	0.49	0.5	13844
weighted avg	0.58	0.59	0.57	13844

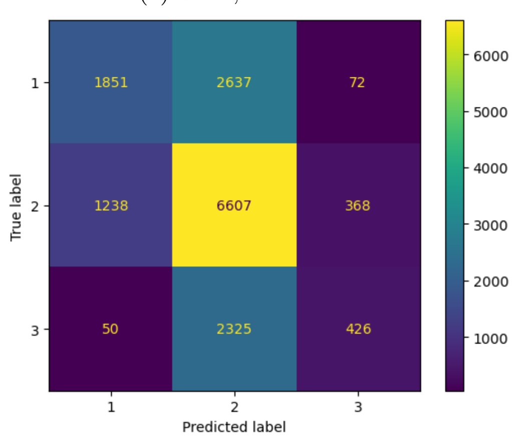
Table 4: Classification report for logistic model with no outliers



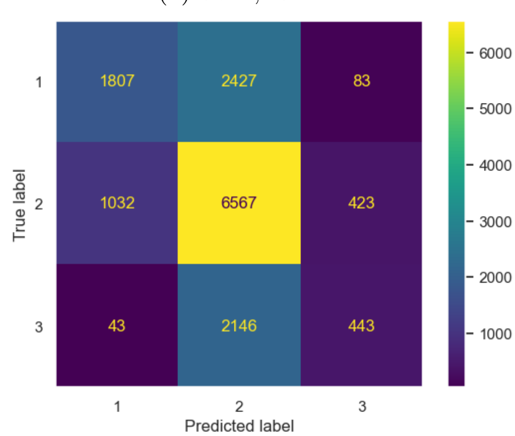
(a) SVM, linear kernel



(b) SVM, RBF kernel



(c) Logistic with no outliers



(d) Logistic with outliers

Figure 2: Confusion matrices for the models

Conclusion and discussion

References

Dataset
https://www.kaggle.com/datasets/parisrohan/credit-score-classification/data

Literature Review
Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178-1192. doi:10.1016/j.ejor.2021.06.053

Yue, Lei, and Luka Vidovic. "Machine Learning and Credit Risk Modelling." *Machine Learning and Credit Risk Modelling | S&P Global Market Intelligence*, S&P Global Market Intelligence, 30 Nov. 2020, www.spglobal.com/marketintelligence/en/news-insights/blog/machine-learning -and-credit-risk-modelling.