Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Performance-Optimierung beim maschinellen Lernen am Beispiel der Bonitätsprüfung von Bankkunden

Bhirawa Satrio Nugroho

Bachelorarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 11.02.2021

Motivation und Aufgabenstellung

Die Kreditwürdigkeitsprüfung ist ein wichtiger Schritt, der von Kreditvergabestellen durchgeführt wird und der darüber entscheiden kann, ob das Bankinstitut potenziellen Kreditnehmern einen Kredit gewährt oder nicht. Diese Prüfung hat einen großen Einfluss auf Agenturen, insbesondere im Finanzsektor. Um finanzielle Probleme zu vermeiden, die aufgrund von Risiken bei der Kreditvergabe auftreten, wird eine Methode benötigt, die die Kreditwürdigkeitsprüfung unterstützt, indem die statistische Leistung eines Kreditscoring-Modells erhöht wird. Mit Hilfe von maschinellen Lernmodellen können Zeit, Aufwand und Kosten für die Durchführung statistischer Analysen, die auf Big Data angewendet werden, reduziert werden. Aus diesem Grund werden in dieser Arbeit Algorithmen des maschinellen Lernens, namentlich von Logistic Regression, K-Nearest Neighbors und Support Vector Machine, verglichen. Ferner werden Experimente durchgeführt, die die Leistung dieser Modelle verbessern können.

Erstellung des Datensatzes

Bei den verwendeten Daten handelt es sich um Daten aus Open Data LMU [1] mit Beobachtungen von 1000 Personen, die Kredite aufgenommen haben. 30% davon können den Kredit nicht zurückzählen. Der Datensatz enthält 21 Features bzw. Merkmale. In der Abbildung 1 kann die Merkmale (oder Spalten) gesehen werden, die im Datensatz enthalten sind. Die Merkmale werden in zwei unterteilt, nämlich in kontinuierliche und kategoriale Merkmale. Fast alle kategorialen Merkmale sind ordinal. Diese Merkmale enthalten die finanzielle Situation, den Charakter und den Status der Person, die den Kredit aufnimmt.

kredit	laufkont	laufzeit	moral	verw	hoehe	sparkont	beszeit	rate	famges
buerge	wohnzeit	verm	alter	weitkred	wohn	bishkred	beruf	pers	telef
gastarb									

Abb. 1: Liste aller Merkmale

Experimente und Vergleich der Modelle

Beim Experimentieren wurden verschiedene Methoden zur Verbesserung der Leistung verwendet. Einer davon ist durch die Verwendung von GridSearchCV. Um eine Kombination von Parametern (Hyperparametern) zu finden, hat Sklearn eine Klasse GridSearchCV, die die gewünschten Hyperparameter mit Hilfe von Kreuzvalidierung testen kann. So kann das Modell die besten Parameter gut generalisiert auswählen. Die verwendete Kreuzvalidierung ist die stratifizierte Kreuzvalidierung. Die Verteilung des Trainings-Sets beträgt 70 % und des Test-Sets 30 % von Datensatz, und der kontinuierliche Wert wurde skaliert. Jedes Modell erhält eine Bewertung basierend auf der verwendeten Matrix, wie z. B. Recall, Precision, F1-Score, ROC AUC-Score und Accuracy. Wegen der unausgewogenen Klasse ist davon auszugehen, dass die Verwendung von ROC AUC als Matrix für die weitere Auswertung besser ist, als die Verwendung der Accuracy [2].

Ergebnisse

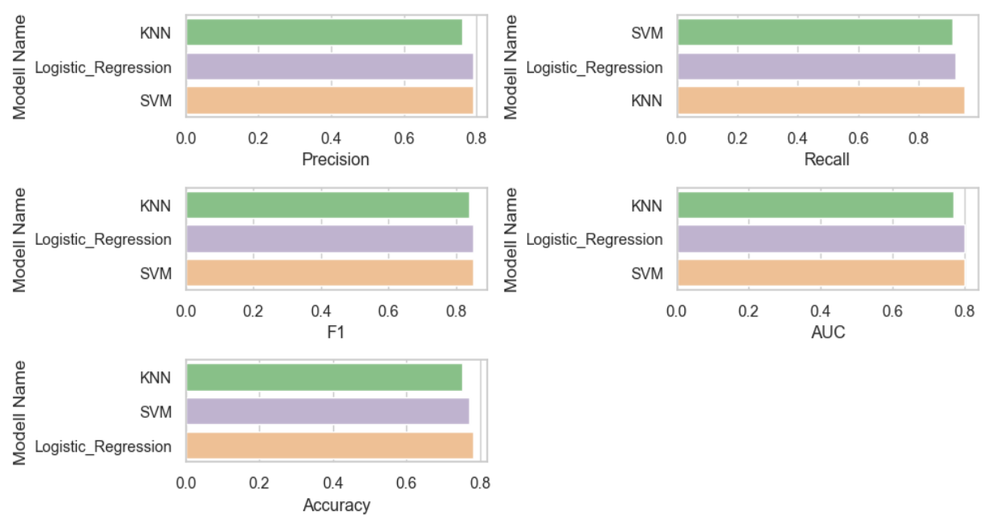

Abb. 2: Vergleich der erstellten Modelle mit den besten Parametern, die von GridSearchCV erhalten wurden.

Das Endergebnis dieses Experimentes ist in der Abbildung 2 zu sehen, dass basierend auf Experimente die logistische Regression im Vergleich zu anderen Modellen einige Vorteile hat. Dies zeigte sich, nachdem das Abstimmen von besten Parametern, die von GridSearchCV erhalten.

Fazit und Ausblick

Letztendlich hängen die Kriterien eines guten Modells zur Bestimmung der Kreditwürdigkeit vom Geschäftsmodell des Finanzinstituts ab. Selbstverständlich ist es das Ziel eines jeden Finanzinstituts, den Gewinn zu maximieren und den zu zahlenden Preis zu reduzieren, wenn es einen Verlust erleidet. In dieser Arbeit wird empfohlen, weitere Forschung zu betreiben, indem Merkmale angewendet werde, um die Komplexität eines Modells zu reduzieren. Weitere Empfehlungen sind eine bessere Verteilung der Trainingsets und Testsets, die Verwendung von Kreuzvalidierungsarten, die bessere Schätzungen für weniger große Datensätze haben. Und Verwendung einer Pipeline, um bei der Ausführung von Workflows für maschinelles Lernen noch ordentlicher zu sein. Dadurch können Fehler bei der Datenverarbeitung, Modellierung und Evaluierung reduziert werden.

Quellen

[1] Kreditscoring zur Klassifikation von Kreditnehmern. 2010. Open Data LMU. doi:10.5282/ubm/data.23
[2] David James. 2018. *Introduction to Machine Learning with Python: A Guide for Beginners in Data Science (1st. ed.).*

Betreuung Dipl.-Inform. Ingo Boersch • Technische Hochschule Brandenburg • Prof. Dr.-Ing. Jochen Heinsohn • Technische Hochschule Brandenburg