Im Datenmeer Versunken

Eine Studie des Titanic Datensatzes

Arne Berner

Fachhochschule Kiel

Einleitung

In dieser Hausarbeit möchte ich auf einige Modelle des maschinellen Lernens eingehen und anhand vom Titanic Datensatz einige Stärken und Schwächen der Modelle zeigen. Außerdem möchte ich erkunden, was genau Benchmarking- und Trainings-Datensätze besonders geeignet macht. Dabei werde ich k-Nearest-Neighbor, Support Vector Machine, Decision Tree - bzw. Random Forests - und neuronale Netzwerke nutzen. Anschließend werde ich darauf eingehen, wieso die angewendeten Modelle besonders gut, oder schlecht funktioniert haben. Als Hilfsmittel kamen die Dokumentationen von sklearn, Tensorflow und (im Jupyter Note dokumentiert) Tech Blogs zum Einsatz.

Datensatz

Wie kam es dazu?

Meine Anforderungen an den Datensatz waren, dass der Datensatz gut klassifizierbar ist und ich die vorgestellten Klassifizierer aus der Vorlesung darauf anwenden kann. Ich bin vor einigen Jahren schon auf den Titanic Datensatz gestoßen und er wird in vielen Tutorials online als Einführungsbeispiel genutzt. Deswegen war auch mein erster Gedanke, diesen Datensatz zu nehmen, mit dem so viele vertraut sind. Andere Benchmarking Datensätze habe ich nicht gewählt, weil die bekanntesten - wie Cifar-10 oder fashion-mnist - mit Deep Learning Algorithmen zu den besten Resultaten kommen.

Wichtigste Charakteristika des Datensatz

Ich habe mir den Datensatz aus zwei Titanic Datensätzen von Kaggle zusammengestellt. Dieser Datensatz beinhaltet die Gender Submission und den Testsatz aus einer Challenge und dieser die Trainingsdaten. Obwohl die Daten schon als Datensätze getrennt sind, habe ich sie zusammengefügt, bereinigt und nochmal in Training- und Testdatensatz getrennt. Der Datensatz wurde vom Kaggle-Team kuratiert und besteht aus realen Daten. Die Spalten des Datensätzes sind (nach bereinigen) die folgenden:

Spalte	Erklärung
Cabin	Raumnummer
Survived	0: nicht überlebt, 1: überlebt
Age	Alter in vollen Jahren

Name	Nachname, Titel Vorname
Sex	Geschlecht
Pclass	1: erste Klasse, 2: zweite Klasse, 3: dritte Klasse
SibSp	Anzahl an Geschwistern an Board
Parch	Anzahl an Eltern/Kindern an Board
Embarked	Wo gestartet? 0: Southampton, 1: Cork, 2: Queensland
Fare	Fahrpreis

Die Kategorie, die ich zum Klassifizieren nutze, ist “Survived”.

Wer hat zu dem Thema geforscht?

Die Technische Universität Kocaeli hat zur “7th International Conference on Advanced Technologies” dieses Paper veröffentlicht. Hierbei handelt es sich um einen Benchmarking Vergleich verschiedener Algorithmen. Ihre Forschungsergebnisse haben ergeben, dass einige Kombination per “Voting” (Abstimmung) verschiedener Algorithmen zum besten Ergebnis führt.

Jörg Stolz und Anaid Lindemann haben in diesem Paper für das Methodology European Journal of Research Methods for the Behavioral and Social Sciences über die Möglichkeiten geschrieben, mit dem Datensatz Data Analysis zu unterrichten. Dabei sind sie nicht nur auf den quantitativen Datensatz, den ich genutzt habe, eingegangen, sondern haben auch einen qualitativen Datensatz zur Titanic mit dazu genommen. In dem Paper wird außerdem eine drei stündige Unterrichtseinheit beschrieben, in der “mixed Methods” aus der Datenanalyse genutzt werden können.

Eine weitere Suche auf Researchgate zeigt, dass dieser Datensatz hauptsächlich genutzt wird, um verschiedene Algorithmen miteinander zu vergleichen. Ich habe leider keine Forschungsergebnisse auf das tatsächliche Ereignis bezogen finden können.

Zielsetzung:

Ich möchte mit einer hohen Wahrscheinlichkeit voraussagen können, ob ein Eintrag aus dem Testset eine Überlebenschance hat. Dafür möchte ich eine Genauigkeit von mindestens 85% erreichen. Außerdem möchte ich verschiedene Algorithmen auf diesem Datensatz miteinander vergleichen.

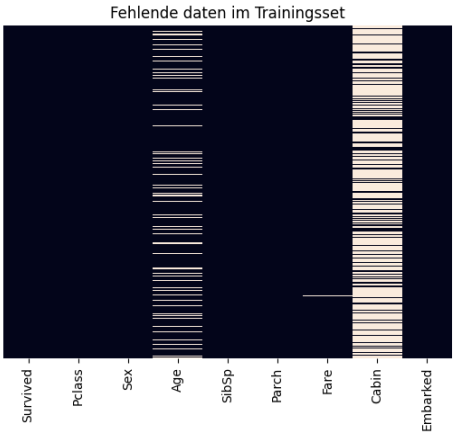
Methoden

Vorgehen

1. Daten importieren und bereinigen
2. Explorative Data Analysis (EDA) - Exploratives Daten Analysieren
3. Maschinelles Lernen Modelle anwenden und optimieren
4. Die Ergebnisse auswerten

Datenaufbereitung und Visualisierung

Es fällt schnell auf, dass es einige Nullwerte gibt. Eine kurze Suche im Notebook hat ergeben, dass “Cabin” und “Age” betroffen sind (die weißen Striche geben Nullwerte an).

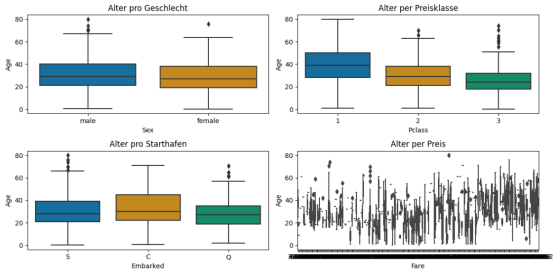


Außerdem gab es 190 Duplikate in den Zeilen, die wahrscheinlich beim Kombinieren der Datensätze entstanden sind. Solche Duplikate können einfach entfernt werden. Nullwerte brauchen jedoch etwas mehr Zuwendung.

Es gibt mehrere Möglichkeiten.

1. Die Werte können gelöscht werden, wenn sie so unvollständig sind, dass sich nichts aus ihnen ableiten lässt.
2. Die Werte werden in “Vorhanden: 1” und “Nicht Vorhanden: 0” umgewandelt.
3. Die Werte werden durch Dummy Werte ersetzt, welche sich oft am Mittelwert orientieren oder eine andere Bedingung für ihren Wert bekommen.

Für das fehlende “Fare” habe ich mich entschieden, die erste Möglichkeit zu nehmen und die gesamte Zeile zu löschen. Für Cabin hätte ich die gesamte Spalte löschen können, habe mich jedoch für die zweite Möglichkeit entschieden. Bei Age habe ich mir Kategorien gesucht, die einen großen Einfluss auf das Alter haben und habe dort den Mittelwert genommen.



(l.o: Alter per Geschlecht, r.o: Alter per Preisklasse, l.u: Alter per Heimathafen, r.u: Alter per Fahrpreis)

Das Geschlecht scheint kaum einen Einfluss auf den Mittelwert zu haben. Der Starthafen hat zu starke Ausschläge und der Fahrkartenpreis hat zu viele Kategorien. Doch die Preisklasse scheint sehr unterschiedlich pro Kategorie zu sein. Deswegen habe ich mich für diesen Mittelwert entschieden.

Explorative Datenanalyse (EDA)

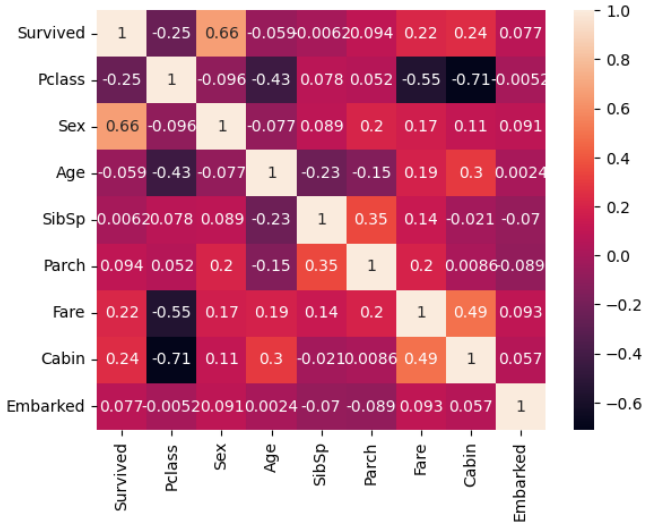
Beim erstmaligen Untersuchen der Daten ist es oft wichtig, mir eine Übersicht zu verschaffen.

Ich weiß, dass ich die Spalte “Survived” zum Klassifizieren benutzen möchte. Dabei haben etwas mehr als ⅓ der Passagiere überlebt. Aus der Verteilung kann ich schließen, dass die Genauigkeit später als Metrik sinnvoll eingesetzt werden kann. Wäre eine Kategorie nur sehr wenig vertreten, würde ein Algorithmus, der sich immer für die Seite der Mehrheit entscheidet, sehr gut abschneiden. Doch er würde nicht gut klassifizieren. In solchen Fällen ist es sinnvoll, eine ROC Kurve zu nutzen, auf die ich hier verzichten werde.

Als erster Schritt lässt sich durch einen Pairplot häufig eine Tendenz erkennen:

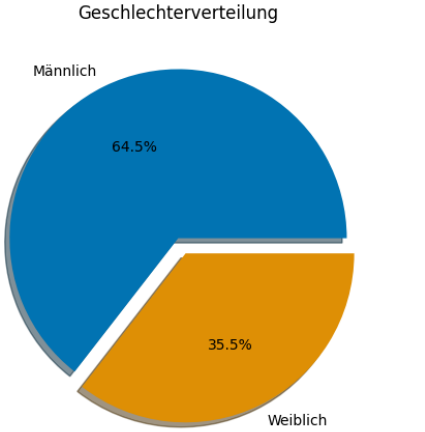
- Menschen in der ersten Klasse überleben deutlich häufiger als in der dritten Klasse und damit zusammenhängend auch Menschen, die mehr für das Ticket bezahlen
- jüngere Leute haben auch eine höhere Überlebenschance
- Menschen mit 1-3 Geschwistern auf der Titanic überlebten auch häufiger
- genauso Menschen mit ein bis zwei Kinder

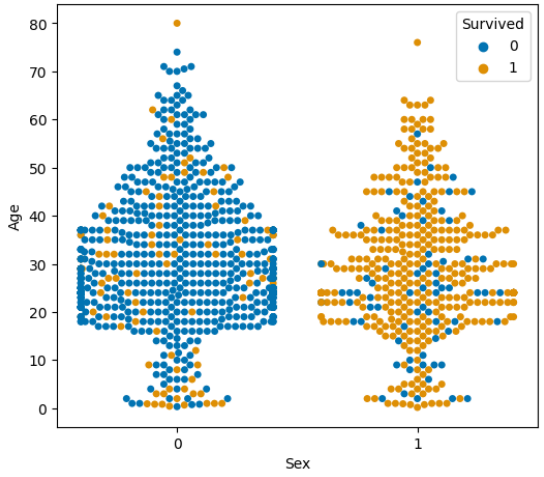
Eine leichte Methode, um diese Thesen zu überprüfen ist, sich die Korrelationsmatrix dazu anzuschauen.

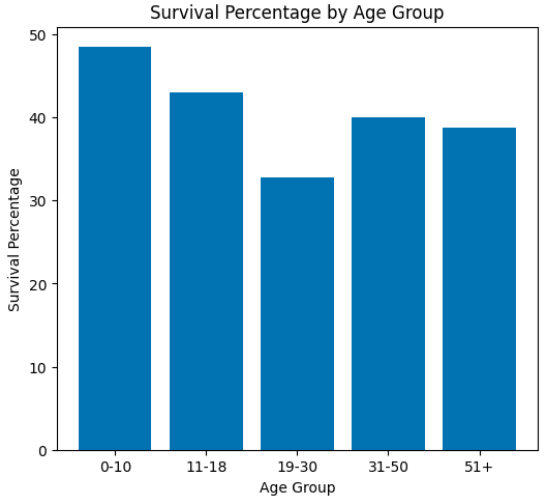


Diese zeigt eine starke Korrelation zwischen Preisklasse,Fahrpreis und Überleben. Doch eine nicht vorhandene Korrelation sagt nicht aus, dass die Kategorien nicht voneinander abhängen. Eine Anhäufung, die einer Glockenkurve ähnelt, wird eine niedrige Korrelation vorweisen, trotzdem können wir ein Muster erkennen. Ähnliches ist bei der Kategorie SibSp und Age passiert.

Exemplarisch gehe ich auf den Zusammenhang zwischen Alter, Geschlecht und Überleben ein. Mehr Analysen lassen sich im Notebook finden.
Das folgende Diagramm verdeutlicht, dass es einen Zusammenhang gibt:



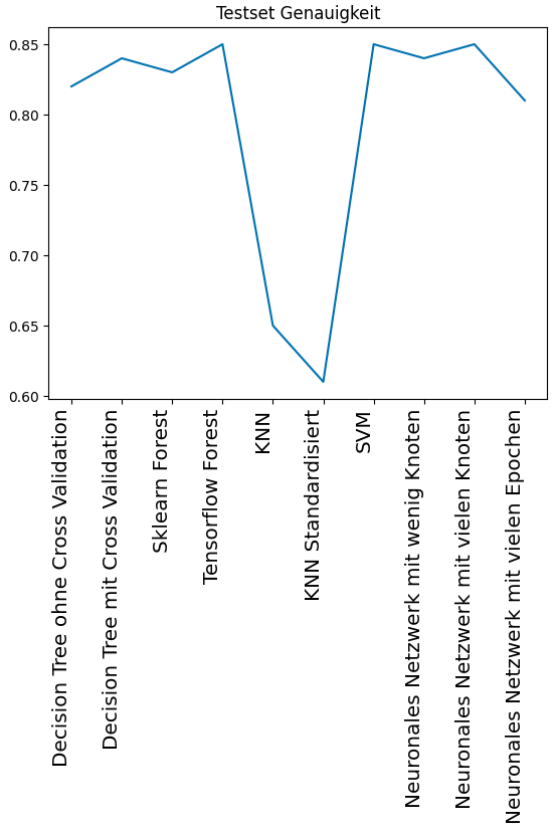




Nicht nur hatten Kinder eine höhere Überlebensrate, sondern vor allem Frauen haben häufiger überlebt als Männer. Die Vermutung liegt nahe, dass die Rettungsboote nach der Regel “Frauen und Kinder zuerst” befüllt wurden.

Anwenden der Algorithmen und Optimieren

Ich habe Decision Tree (mit und ohne Cross Validation), Random Forest (Sklearn und Tensorflow), Support Vector Machine, k-Nearest-Neighbor (standardisiert und nicht standardisiert) und Fully Connected Neural Networks (mit vielen und wenigen Knoten und vielen Epochen) Algorithmen genutzt.



Auffällig war, dass die Decision Trees je nachdem, ob Cross Validation genutzt wurde, unterschiedlich abgeschnitten haben. Ohne Voreinstellung hat der Decision Tree für alle Daten ein Blatt erzeugt und hat somit das Trainingsset auswendig gelernt (siehe Note). Auf dem Trainingsset kam es so zu

beachtlichen 98% Genauigkeit, auf dem Testset hingegen nur auf 82%. Damit war der Klassifizierer overfitted. Cross Validation hat hier Abhilfe verschafft und kam nach kurzer Zeit zu einem der besten Ergebnisse mit 84%. Dabei scheint es fast offensichtlich zu sein, dass der RandomForest mit seinen 83% sehr Nahe an dem Ergebnis ist, da der RandomForest verschiedene Decision Trees für sich nutzt. Der Unterschied zu verschiedenen Hyperparametern ist nicht allzu groß. Viel spannender ist, dass die beiden RandomForests von Tensorflow und Sklearn ähnliche Genauigkeit, aber verschiedene Sensitivität und Spezifität aufweisen.

Die gleiche Genauigkeit wie Random Forest hatte SVM und somit das beste Ergebnis. Verwunderlich war nur, dass eine standardisierte Version des SVM sehr schlecht abgeschnitten hat. Außerdem hat es mich gewundert, dass obwohl SVM Radial Basis als Funktion für den Kernel nutzt und somit KNN sehr ähnelt, KNN dermaßen viel schlechter abgeschnitten hat. Auch hier hat standardisieren zu einem schlechteren Ergebnis geführt, weswegen ich vermute, dass der Standardisierungsprozess Fehlerhaft war. Nicht standardisiert hat KNN zwar besser abgeschnitten, als standardisiert, war aber hoffnungslos overfitted und ließ sich mit den gewählten Hyper Parametern für die Cross Validation auch nicht auf Kurs bringen.

Für KNN hatte ich mir kein gutes Ergebnis erhofft, weil es meistens für Unsupervised Learning genutzt wird, aber sie waren schlechter als erwartet. Die Sensitivitäten und Spezifitäten waren bei allen Algorithmen ähnlich, trotzdem sollte bei der Wahl des Algorithmus geschaut werden, für den jeweiligen Datensatz stimmt.

Diskussion

Einordnung der Ergebnisse:

Zum Großteil war die Genauigkeit zwar in der Nähe von 85%, aber wenn hier von der Maschine entschieden werden sollte, ob eine Rettungsaktion sinnvoll ist, würde ich mich auf das Ergebnis verlassen. Mich überrascht es nicht, dass der Decision Tree so gut abgeschnitten hat, da bei der EDA schon klar wurde, dass sich meistens das Überleben an recht einfache Regeln knüpfen lies. Leider konnten die neuronalen Netzwerke ihre Stärke, Regeln in etwas Abstraktem zu finden, nicht gut ausspielen. Die Regeln waren einfach genug für simplere Methoden. Hier zeigt sich, dass alte Algorithmen je nach Datensatz nicht unbedingt schlechter sind.

Würde ich diesen Datensatz weiterempfehlen?

Ich würde den Datensatz definitiv zum Lernen von Data Analysis empfehlen, vor allem - wie im obigen Paper geschrieben - mit qualitativen Daten gemischt. Als Benchmark für neuere Algorithmen ist er weniger zu gebrauchen, weil er lückenhaft ist und zu wenig Daten besitzt.

Offene Fragen?

Wie realitätsnah ist das Bearbeiten eines solchen Datensatzes? Die Hauptarbeit beim Machine Learning steckt eigentlich in der Datenbereinigung. Auch eine gut dokumentierte CSV Datei als Datensatz ist relativ selten vorhanden. Die eigentliche Arbeit der Datenanalyse steckt meiner Meinung nach im Beschaffen und Aufbereiten der Daten. Ist es dann richtig gelernt, wenn es so weit von der Realität entfernt ist? Ich hoffe, in meiner weiteren Karriere diese Fragen beantworten zu können.

Fazit

Es ist wichtig, den richtigen Algorithmus für den richtigen Datensatz zu finden. Abschließend ist zu sagen, dass einfache Datensätze gut zum Lernen, aber nicht unbedingt zum Benchmarken sind, da sie in ihrer Struktur zu weit von der Realität entfernt sind.