Data Analytics

Master Digitale Prozesse und Technologien

Hochschule für Technik Stuttgart

Projektbericht

Analyse von Verkehrsunfällen in den USA 2016-2023

Vorgelegt von:

Antonino Forte
Matrikelnummer: 1005833, Fachsemester: 2

Bruno Vidal dos Santos
Matrikelnummer: 1000935, Fachsemester: 1

Vinko Jelić
Matrikelnummer: 1005836, Fachsemester: 2

Abgabedatum: 09.01.2024

Modul: Data Analytics
Semester: Wintersemester 2023/24
Dozenten: Prof. Dr. Sebastian Speiser, Prof. Dr. Ulrike Pado

Inhaltsverzeichnis

1. Motivation

Für unser Projekt liegt die Motivation in der immer weiter fortschreitenden Entwicklung der Datenanalyse und des maschinellen Lernens im Kontext von „Big Data“. Dabei gerät der Trend zur Entwicklung einer künstlichen Intelligenz durch diese Ansätze in den Hintergrund.

Statt das menschliche Denken zu simulieren, sollen spezifische Lösungen für klar definierte Probleme geschaffen werden. Setzt man Algorithmen des maschinellen Lernens ein, können entsprechende Muster aus historischen Daten erkannt und verarbeitet werden.

In der Regel greifen die Methoden des maschinellen Lernens auf objektive und umfassende Daten („Big Data“). Dadurch werden genauere und effizientere Vorhersagen ermöglicht, da die Algorithmen nach Mustern suchen, die am ehesten zur jeweiligen Situation passen.

Dementsprechend lassen sich drei Schlüsselvoraussetzungen aufstellen:

- Datenverfügbarkeit (z.B. durch IoT-Sensoren)
- Rechen- und Speicherkapazitäten
- Verfügbare Algorithmen und Tools (Python inkl. Bibliotheken)

Unser Projekt verfolgt die Zielsetzung, das Potenzial der Datenanalyse und des maschinellen Lernens auszunutzen und tiefe Einblicke in das komplexe Problemfeld zu wahren. Zum Einsatz kommen das Tool Jupyter Notebook sowie die Programmiersprache Python, unterstützt durch leistungsstarke Bibliotheken, wie „pandas“, „numpy“, „scikit-learn“ und weitere.

2. Problemstellung

Unser Fokus im Projekt liegt auf der prädiktiven Analyse der Schwere („Severity“) von Verkehrsunfällen in den USA. Als Grundlage nutzen wir den Datensatz „US Accidents“ (vgl. US Accidents (2016 - 2023) 2024) von Kaggle. Somit konzentrieren wir uns auf die Entwicklung von Modellen zur Vorhersage der Unfallschwere und dem zeitgleichen Einblick in die Faktoren, die zu den schwerwiegenderen Unfällen führen.

Die Hauptaspekte unserer Problemstellung werden durch folgende Punkte untermauert:

1. Zielvariable – Severity: Die Vorhersage der Unfallschwere dient als zentrales Ziel der Analyse. Die verschiedenen Kategorien der Schwere (eins bis vier) werden betrachtet, um einordnen zu können, welche Faktoren oder Zusammenhänge mit schweren Unfällen in Verbindung stehen.
2. Analyse: Die Analyse der Faktoren auf die Unfallschwere anhand der verfügbaren Daten, wie Witterungsbedingungen, Verkehrsbedingungen, Straßenzustand, geografische Koordinaten und weiterer relevanter Daten.
3. Modell: Mit Hilfe der ausgewählten Tools sollen prädiktive Modelle erstellt werden. Dabei kommen z.B. Methoden zur Klassifikation zum Einsatz, um die Unfallschwere mit Hilfe der ermittelten und relevanten Faktoren vorherzusagen.
4. Interpretation: Die gewonnenen Erkenntnissen aus den Modellen sollen Interpretationen abgeleitet werden. Dabei spielen z.B. mögliche Maßnahmen zur Verbesserung der Verkehrssicherheit oder Infrastruktur eine Rolle. Dazu gehören mögliche Erweiterungen oder Umverteilungen von Ressourcen in Prozessen rund um das Verkehrs- und Rettungsmanagement.

3. Lösungsansatz

Der Lösungsansatz besteht aus folgenden Schritten:

Datenvorbereitung/Datenbereinigung
Die Datenvorbereitung oder Datenbereinigung beinhaltet den Prozess der systematischen Aufbereitung von Datensätzen, um fehlerhafte oder unvollständige Daten zu identifizieren und zu eliminieren. Dies kann die Entfernung von Zeilen mit fehlenden Werten oder unerwünschten Datenpunkten umfassen, um die Qualität und Zuverlässigkeit der Daten für eine anschließende Analyse oder Modellierung zu verbessern.

Datenexploration
Die Datenexploration umfasst eine eingehende Analyse der "US Accidents" Daten, einschließlich statistischer Kennzahlen, Visualisierungen und Identifikation von Mustern, um eine fundierte Grundlage für den weiteren Analyseprozess zu schaffen.

Datentransformation
In diesem Schritt werden notwendige Datentransformationen durchgeführt, um die Daten für die Modelleignung vorzubereiten. Dazu gehört die Behandlung von fehlenden Werten, Normalisierung von numerischen Variablen und Kodierung von kategorialen Variablen.

Feature Engineering und Zeitreihen
Feature Engineering beinhaltet die Auswahl und Schaffung relevanter Merkmale, um die Vorhersagegenauigkeit zu verbessern. Zeitreihenanalyse ermöglicht die Berücksichtigung zeitlicher Aspekte in den Daten, was insbesondere für die Unfallvorhersage von Bedeutung ist.

Clustering
Die Methoden zum Clustering werden angewendet, um homogene Gruppen von Unfällen (Schwere der Unfälle) zu identifizieren und mögliche Muster oder Trends innerhalb dieser Gruppen zu erkennen. Dies erleichtert die differenzierte Betrachtung verschiedener Unfalltypen.

Klassifikation
Um die Unfallschwere vorherzusagen, werden Klassifikationsalgorithmen benutzt. Dadurch lassen sich Unfälle, basierend auf den identifizierten Einflussfaktoren, zu unterschiedlichen Schweregraden zuordnen.

Dimensionalitätsreduktion

Um die Komplexität des Modells zu reduzieren und die Effizienz zu steigern, werden Methoden zur Dimensionalitätsreduktion eingesetzt. Dadurch werden nur die relevantesten Variablen in die Modelle einbezogen.

Regression

Um quantitative Beziehungen zwischen verschiedenen Variablen und der Unfallschwere zu modellieren, wird die Regression genutzt. Damit erhält man eine präzise Vorhersage der Unfallschwere, welche auf kontinuierlichen Variablen basiert.

Evaluation/Erklärbarkeit

In diesem Abschnitt liegt der Fokus darauf, die erzielten Ergebnisse umfassend zu interpretieren und zu erklären, wobei besonderes Augenmerk auf die Zusammenhänge und Eigenschaften der Daten gelegt wird. Im Rahmen dieser Analyse wird versucht, über die reinen Machine-Learning-Ergebnisse hinauszugehen und tiefere Einsichten und Verständnis in Bezug auf die Charakteristiken der Daten zu gewinnen.

4. Haupterkenntnisse aus den Daten

Datenvorbereitung (Datenbereinigung)

Das gewählte Datenset weist zu Beginn mehr als acht Millionen Unfälle (Zeilen) auf. Damit verbunden sind einige fehlerhafte Daten in verschiedenen Spalten. Wir haben uns vorgenommen, die Daten im Vorhinein aufzubereiten. Primär umfasst die Bereinigung die Entfernung aller Unfälle bzw. Zeilen, die einen „NaN“-Wert haben. Hierzu sollen alle betroffenen Zeilen ausfindig gemacht und eliminiert werden. Nach der Bereinigung bleiben etwas mehr als sieben Millionen Zeilen übrig. Darüber hinaus werden auch, die von uns als unwichtig bewerteten Spalten, gelöscht.

Datenexploration

In diesem Abschnitt werden verschiedene Visualisierungen und Analysen der Verkehrsunfalldaten durchgeführt. Zunächst werden Diagramme erstellt, um Trends in der Gesamtanzahl der Unfälle über die Jahre sowie die Top 10 Bundesstaaten mit den meisten Unfällen zu identifizieren. Ein Scatter-Plot auf einer Weltkarte zeigt die geografische Verteilung der Unfälle, wobei ein Zoom auf Nordamerika aufzeigt, dass die meisten Unfälle an der West- und Ostküste stattfinden.

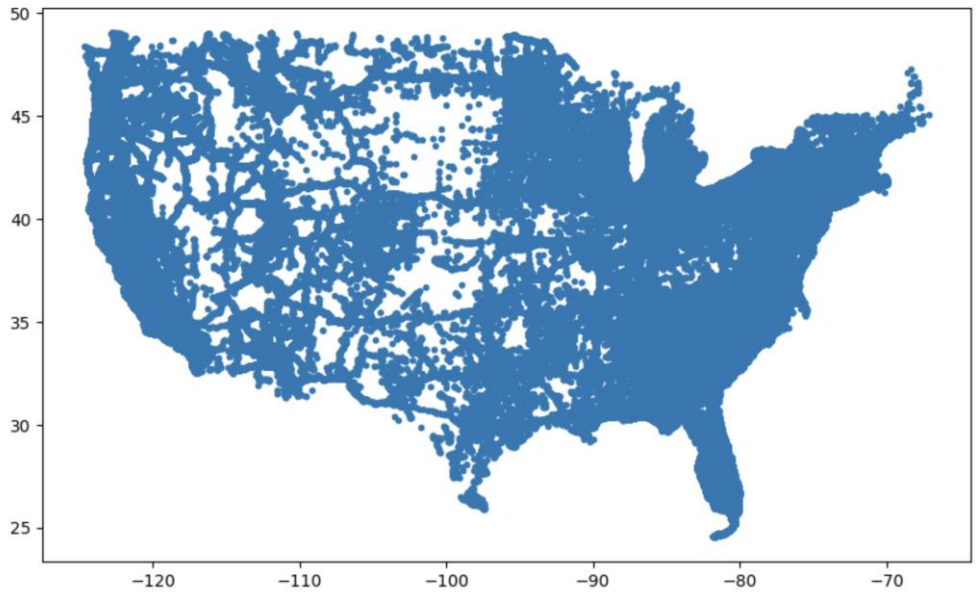

Abbildung 1Verteilung der Unfälle auf der Karte

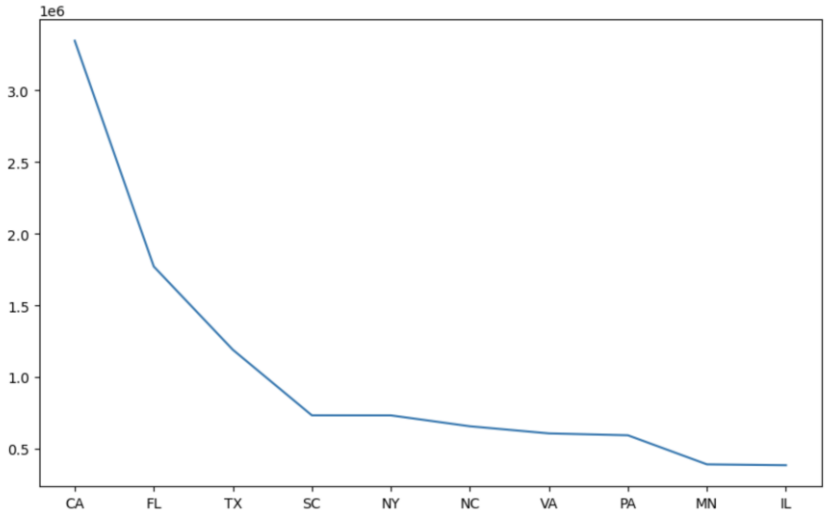

Abbildung 2 Staaten mit den häufigsten Unfälle

Weiterhin werden Histogramme für die Gesamtanzahl der Unfälle pro Jahr, Monat, Wochentag und Stunde erstellt. Es folgt eine statistische Zusammenfassung der Wetterbedingungen, einschließlich Temperatur, Luftfeuchtigkeit, Druck, Sichtweite und Windgeschwindigkeit. Die Daten werden auch nach Bundesstaaten gruppiert, um Durchschnittswerte für Schweregrad und Wetterbedingungen zu erhalten. Die Daten werden schließlich in einer CSV-Datei gespeichert.

Die Daten kommen in unterschiedlichen Datentypen vor (int, float, bool, obj). Dabei handelt es sich oft um Messwerte oder Beschreibungen. Die Daten beschreiben Features, die Verkehrsunfälle dokumentieren. Der dabei am wichtigsten erscheinende Faktor hier ist die Schwere des Unfalls (Severity). Weiterhin können die Daten in Kategorien unterteilt werden, wie Wetter, Straßeninformationen, zeitliche Faktoren (Sonnenuntergang etc.) und räumliche Faktoren. Die Daten könnten für die Vorhersage der Schwere eines Unfalles dienen. Dadurch könnte erreicht werden, dass beispielsweise Notdienste besser und schneller auf Unfälle reagieren können. Weiterhin könnten man die Analysen durchführen, um Unfälle und deren Schwere auf räumliche Faktoren und Infrastruktur zu untersuchen. Dadurch können Hotspots identifiziert werden, die zu Maßnahmen wie Umbauten an einer Straße führen könnten.

Datentransformation

Hier werden zusätzliche Features zu den Unfalldaten hinzugefügt. Wir haben dafür eine Konvertierungstabelle von Bundesstaaten zu Regionen eingebunden und in unsere Daten integriert. Damit sind drei neue Spalten hinzugekommen. Im Anschluss wird eine Pivot-

Tabelle erstellt, die die Anzahl der Unfälle nach der Schwere und Wetterbedingungen aufschlüsselt.
Eine grafische Darstellung in Form eines Kreisdiagramms verdeutlicht die Anteile der Unfälle nach verschiedenen Wetterbedingungen. Die Analyse zeigt, dass die meisten Unfälle bei heiterem Wetter („Fair“) auftreten.

Feature Engineering und Zeitreihen

Die Daten wurden nochmal genauer betrachtet, um neue und für unseren Fall besser geeignete Features zu erstellen. Dafür haben wir die Spalte Start_Time und End_Time in datetimes umgewandelt. Darauffolgend haben wir neue Spalten mit verschieden Gruppierungen erzeugt. Diese Spalten wurden Year, Month, day, hour genannt. Durch diese neuen Features konnten wir spezifischere Visualisierungen erstellen und Aussagen treffen. Dadurch konnten wir aus den Daten erkennen, wann die meisten Unfällen passieren. Abschließend konnten wir hier eine Zeitreihenanalyse durchführen. Die Visualisierungen werden im folgenden Kapitel dargestellt (Abbildung 6/Abbildung 7).

Clustering

Es wird eine Clusteringanalyse auf den Unfalldaten durchgeführt, um mögliche Muster oder Trends in den Daten herauszufinden. Zunächst werden alle verfügbaren Features für die Analyse verwendet.

Wir haben festgestellt, dass die Unfallschwere (Severity) nicht dazu beiträgt, klare Cluster in den Daten zu erkennen. Die erste Visualisierung der Cluster war daher intuitiv nicht aussagekräftig. Die Erhöhung der Anzahl der Cluster brachte keine Verbesserung, wie im zweiten Scatterplot ersichtlich. Durch die Änderung der Featureauswahl konnten jedoch gut differenzierbare Cluster identifiziert werden. Insbesondere eignet sich die Verwendung von drei Clustern im Vergleich zu einer höheren Anzahl (15).

Die Analyse der Zentroide und Cluster in der Visualisierung zeigt, dass die Cluster unterschiedliche klimatische und geografische Bedingungen repräsentieren. Diese Bedingungen korrelieren mit der Häufigkeit von Unfällen. Zum Beispiel zeigt Cluster 1 Unfälle in kälteren Temperaturen und höherer Luftfeuchtigkeit, während Cluster 2 Unfälle in wärmeren Gebieten mit hoher Luftfeuchtigkeit repräsentiert.

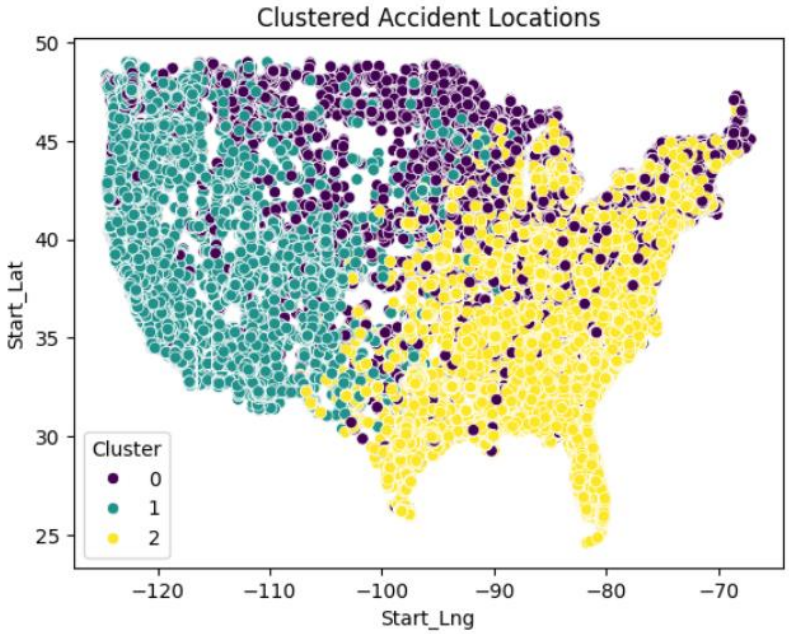

Abbildung 3 Clustering mit drei clustern

Des Weiteren könnte das Clustering verwendet werden, um Ausreißer, insbesondere bei den numerischen Werten (Temperature, Humidity), zu identifizieren. Eine mögliche Vorgehensweise wäre, den Abstand zwischen den Datenpunkten und dem Zentroid zu berechnen (als absoluten Wert). Die maximalen und minimalen Temperaturen könnten aus externen Quellen wie dem Internet abgelesen werden und Datenpunkte außerhalb dieser Grenzen könnten eliminiert werden.

Klassifikation

Aus den Daten hat sich ergeben, dass man auf Hinsicht des Verkehrsflusses gut nach der Unfallschwere „Severity“ klassifizieren kann. Die Severity unterteilt sich in vier Kategorien:

- 1 (leichter Einfluss auf den Verkehr),
- 2 (mittelschwerer Einfluss auf den Verkehr),
- 3 (schwerer Einfluss auf den Verkehr),
- 4 (sehr schwerer Einfluss auf den Verkehr).

Daraufhin haben wir alle Features in Erwägung gezogen, die hinsichtlich mit dem Feature „Severity“ kohärieren. Wir haben alle Features verwendet, um möglichst ein genaues Ergebnis für die Severity zu bestimmen. Es wurde der LGBM-Klassifizier verwendet aufgrund des leistungsstarken Algorithmus und den vielen Datensätzen die wir verwenden. Es konnten folgende Ergebnisse erzielt werden:

```
Accuracy=  {0.9271515806431485}
              precision    recall  f1-score   support

           1       0.77      0.64      0.70     13074
           2       0.94      0.97      0.96   1133919
           3       0.86      0.78      0.82    227753
           4       0.94      0.47      0.63     35566

    accuracy                           0.93   1410312
   macro avg       0.88      0.72      0.77   1410312
weighted avg       0.93      0.93      0.92   1410312
```

Abbildung 4 Klassifikationsreport mit LGMB-Klassifizierer

- Severity 1: Moderate Precision und Recall, guter F1-Score.
- Severity 2: Hohe Precision, Recall und F1-Score,
- Severity 3: Moderate Precision und Recall, guter F1-Score.
- Severity 4: Hohe Precision, niedriger Recall, moderater F1-Score. (möglicherweise hat das Modell Schwierigkeiten, diese Klasse zu erkennen.)
- Gesamtgenauigkeit beträgt 93%

Dimensionalitätsreduktion

Nachdem wir den Klassifikationsreport generiert haben, versuchten wir mit der Dimensionalitätsreduktion, mögliche unbrauchbaren Features zu erkennen. Um diese zu erkennen, wurden Trainings und Testdaten in Entwicklung jeweils in einem 50/50 Split geteilt und stratifiziert, um einen Klassifikationsreport zu erstellen. Anhand des vorherigen Klassifikationsreports konnten wir den IST-Zustand erkennen. Im aktuellen Zustand hat das Modell eine Genauigkeit von 93%.

Nun wird die VarianceThreshold Methode verwendet, um mögliche Varianzen in den Features zu erkennen und eine möglichst genauere Vorhersage zu treffen. Der Schwellwert wurde erstmals auf 0.1 gesetzt und ausgeführt. Daraufhin wurden zwei Featuers erkannt: „Turning-Loop“ und „Timezone“, die für den darauffolgenden Klassifikationsreport ausgeschlossen wurden. Anschließend wurden beide Reports miteinander verglichen, jedoch konnte keine bessere und keine schlechtere Genauigkeit erzielt werden.

Ebenso haben sich die Precision, Recall und F1-Score in keiner Klasse verändert. Die einzige Verbesserung lag in der Durchlaufzeit aufgrund der Kürzung von Features.

Regression

Im Rahmen der Regression haben wir die Modellierung und Prädikation ausgetestet und uns am Ende für die Prädikation entschieden, weil wir uns als Ziel genommen haben die Severity vorherzusagen. Wie bereits in der Klassifikation benötigen wir die Trainingsdaten für das Training des Modells, diese werden dann auf ungesehene Testdaten geprüft.

Unsere Daten wurden gegen den Zielwert (Severity) geplottet und anschließend die Daten in Trainings-Entwicklung-Test aufgeteilt. Ein Feature (Humidity) konnte aufgrund eines negativen Koeffizienten als negativer Einfluss auf die Severity identifiziert werden. Wir erzielten ein RMSE-Wert von 0,2286, welches zusammenfasst, dass wir durchschnittlich um ca. 0,2286 Einheiten von den tatsächlichen Werten abweichen.

Daraufhin haben wir den nichtlinearen Anteil in den Features trainiert und auf den Entwicklungsdaten evaluiert. Schließlich konnten wir eine leichte Verbesserung sehen, denn der RMSE-Wert betrug: 0,2281.

Evaluation und Interpretation

In der Evaluierung haben wir unser Modell mit der Frequenzbaseline (Klassifikation) und der Mittelwertbaseline (Regression) verglichen. Zuerst haben wir die Frequenzbaseline untersucht und konnten die häufigste Klasse, in den Fall die Klasse „2“, identifizieren. Der F1-Score für die konstante Vorhersage der Klasse 2 beträgt: 0,7167. Dies weist daraufhin, dass wir eine zufriedenstellende Balance zwischen Präzision und Recall haben.

Als zweites haben wir die Mittelwertbaseline (Regression) ausgegeben und konnten festellen, dass die Mittelwertbaseline einen etwas höheren RMSE-Wert (0,48) beträgt im Vergleich zur Modellvorhersage (0.4516). Daraufhin wurde der GaussianNB Klassifizierer erstellt, trainiert und verwendet, um die Vorhersagen auf den Testdaten zu machen. Wir konnten eine Gesamtgenauigkeit von 18% und ein Weigthed Avergage von 9% erzielen.

Im Anschluss haben wir eine Lernkurve erstellt, um das Modell auf Over- oder Underfitting zu prüfen. Dafür benutzten wir den LGBM-Klassifizierer und lassen uns die Lernkurve in 300.000 tausend Daten pro Schritte generieren. Als Ergebnis konnte eine Stagnierung gegen 0,925 der Trainingsdaten und Entwicklungsdaten festgestellt werden. Demzufolge ist weder ein Over- noch ein Underfitting der Fall.

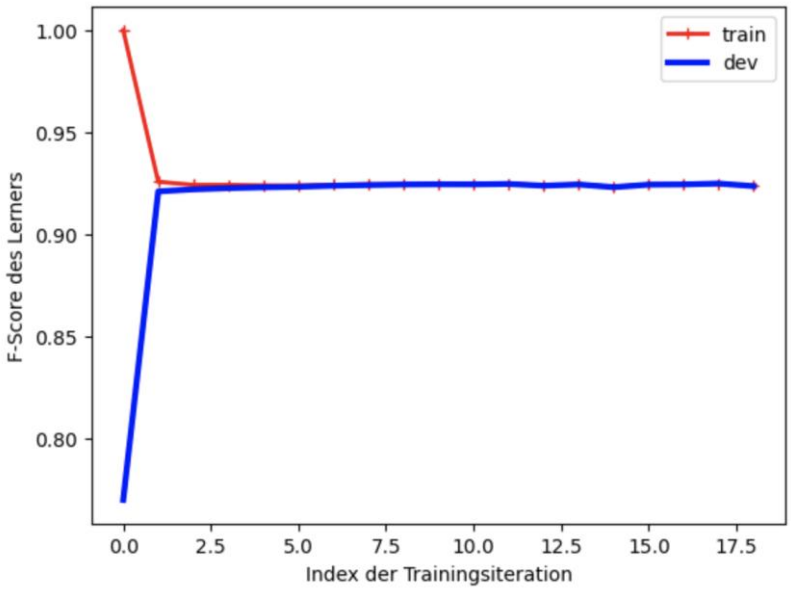

Abbildung 5 Over- oder Underfitting Überprüfung

Als letztes haben wir noch mit LIME die Interpretation durchgeführt und haben eine Wahrscheinlichkeitsverteilung über unsere vier Severitys ausgegeben. Daraufhin haben wir den Explainer (LimeTabularExplainer) generiert. Wir haben einen zufälligen Index (= 705153) genommen und die Vorhersage stimmte mit der tatsächlichen Klasse überein. Es wurden zwei Labels mit den entsprechenden Features und den Evidenzwerten vorgeschlagen. Je größer der Evidenzwert, desto ausschlaggebender ist das Feature für die Klasseneinteilung.

5. Erkenntnisse aus der Umsetzung inklusive Schwierigkeiten

Die Umsetzung unseres Projekts begann mit der Einführung und der Datenexploration. Hier haben wir uns einen ersten konkreten Überblick über unsere Daten verschafft. Durch unsere Programmierung konnten wir schon einige Zusammenhänge in den Daten erkennen. Einerseits konnten wir unsere Features durch genauere Beobachtungen in Gruppen wie Wetterbedingungen, Straßeninformationen, Geoinformationen und zeitliche Faktoren unterteilen. Im Bereich der Datenexploration haben wir zu diesen Gruppen einige Visualisierungen durchgeführt, um im Voraus Muster erkennen zu können. Vor allem konnten wir durch die Geoinformationen erkennen, dass der Großteil unserer Unfälle an der West- und Ost-Küste stattfinden. Wir haben weiterhin durch die zeitlichen Faktoren in unserer Datei erkannt, dass die meisten Unfälle im Dezember 2022 registriert wurden. Die meisten Unfälle geschahen um 08:00 Uhr und 17:00 Uhr, was aus diesem Grund erklärt werden kann, dass um diese Zeit viele Pendler unterwegs sind.

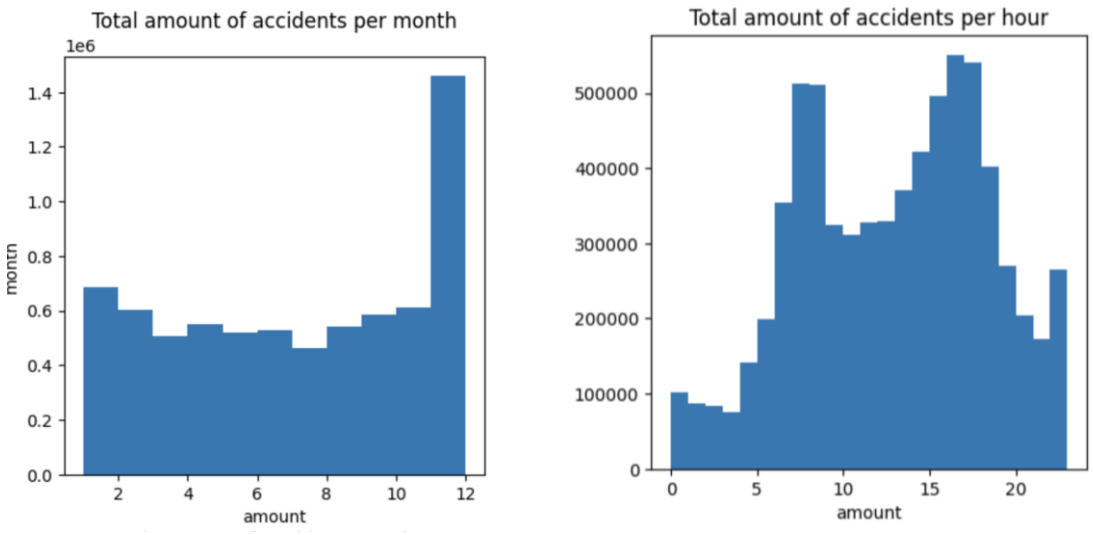

Abbildung 6 Anzahl der Unfälle je Monat und zur welchen Uhrzeit

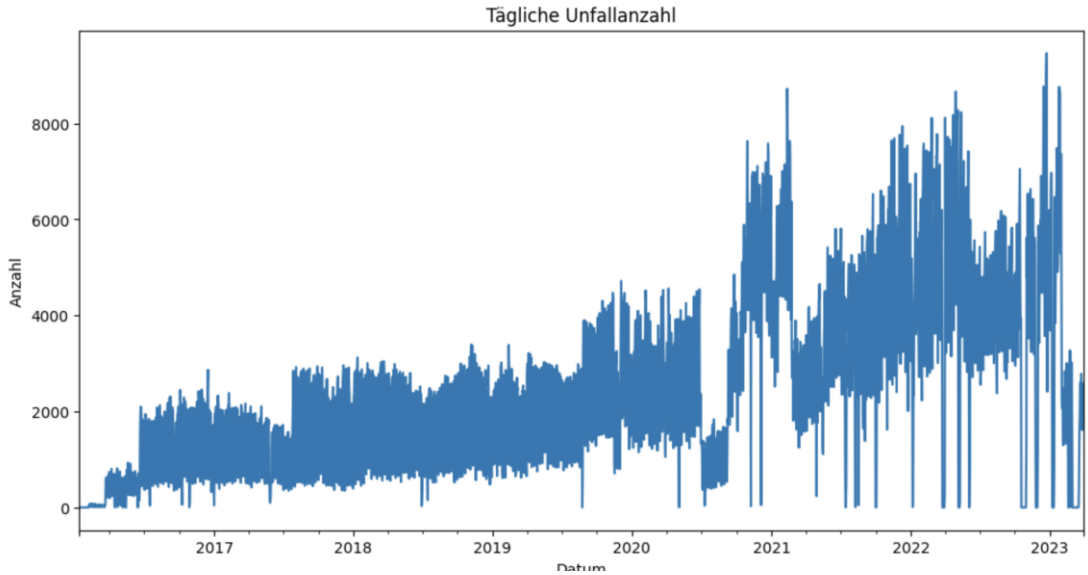

Abbildung 7 Tägliche Unfallszahlen von 2016 bis 2023

Eine Problematik, die uns hier Schwierigkeiten bereitet hat, war die große Datenmenge. Durch die große Datenmenge dauerten die Operationen in unseren Notebooks entsprechend lange.

Aufbauend dazu folgte die Datentransformation, das Feature Engineering und die Zeitreihenumwandlung. In diesem Kapitel haben wir die erlernten Fähigkeiten aus dem Skript angewandt, um den Datensatz auf das Modelling vorzubereiten. Hier haben wir erkannt, dass unserer Datensatz nach dem Abschließen der Mindestanforderung noch nicht bereit für das Modeling war. Wir fügten dementsprechend eine Datei hinzu die vor den Mindestanforderungen den Datensatz bereinigen soll. Dies hat zu diesem Zeitpunkt sehr gut gepasst, da wir hier bereits genug Erfahrung mit dem Datensatz gesammelt haben, um zu wissen was uns genau fehlt. Ein Aspekt, den wir hier betrachten mussten, waren die „NaN“-Werte. Die Schwierigkeit hier war es zu entscheiden, ob wir diese Werte künstlich befüllen sollten, oder ob wir diese komplett aus unserem Datensatz streichen sollten. Unsere Entscheidung fiel auf die komplette Streichung dieser Daten, da uns nach dieser Aktion noch über sieben Millionen Unfalldaten zur Verfügung standen.

Durch das Clustering konnten wir erkennen, welche Features für unsere Zentroide relevant sind. Wir sind hier stark davon ausgegangen, dass die Unfallschwere – Severity einen Einfluss auf die Cluster haben wird. Dies war allerdings nicht der Fall. In unserem Fall konnten wir die Cluster so anordnen, dass die Punkte eine Karte der USA zeigen. Die Cluster visualisieren hier zum Großteil nach den Wetterbedingungen. Die Schwierigkeit

hierbei lag sowohl an der großen Datenmenge als auch an der Auswahl der relevanten Features.

Bei der Klassifikation war unser Ziel die Klassifizierung nach der Unfallschwere, beziehungsweise diese vorherzusagen. Hier traten Schwierigkeiten bei der Feature-Auswahl und dem Algorithmus „Support Vector Machines“ auf. Bei beiden hat die große Datenmenge dazu geführt, dass die Algorithmen lange für die Berechnung unserer Werte gebraucht haben. Die Ergebnisse, die berechnet wurden, waren im schlechten Bereich, daher haben wir versucht einen anderen Algorithmus zu finden, der für unsere Anforderungen besser geeignet ist. Insgesamt haben wir drei verschiedene Algorithmen getestet. Der „Logistic Regression“-Algorithmus hat ähnlich wie die „Support Vector Machines“ ein schlechtes Ergebnis bei langer Berechnungsdauer geliefert. Der „LGMB“-Klassifizierer hat für unseren Anwendungsfall gut funktioniert und lieferte uns ein annehmbares Ergebnis.

Die Dimensionalitätsreduktion ergab, dass nur ein Bruchteil unserer Features für unsere Vorhersage nicht relevant sind. Eine Schwierigkeit, die wir hier hatten, waren die nicht-numerischen Werte. Dies konnten wir umgehen, indem wir durch Encoding diese Werte umgewandelt haben.

Bei der Regression wollten wir zuerst eine Modellierung versuchen, um zu erkennen, welche Features in unserem Fall stark zusammenhängen und relevant für die Regression sind. Allerdings traten hier Schwierigkeiten bei der Verwendung der Daten auf, die nicht numerisch waren. Hier war es relevant nochmals die Daten vom Datentype „boolean“ und „object“ zu „encoden“. Für diesen Prozess haben wir den „LabelEncoder“ verwendet, der sich für unseren Fall sehr gut geeignet hat. Eine große Schwierigkeit, die für hierbei entstanden ist, war die Verschlechterung des R-Squared-Faktors nach dem Encoding der nicht numerischen Daten bei der Verwendung aller Features. Eine Möglichkeit wäre es gewesen die Features zu benutzen die nicht encoded wurden, allerdings entsprach das nicht unseren Vorgehensplan. Daher haben wir uns entschieden, ein prädikatives Modell zu trainieren. Hier berechneten wir den RMSE-Wert. Dieser Wert gab uns die Erkenntnis, dass unser Model sehr genau vorhersagen kann, wie Unfälle aufgrund der zahlreichen Features kategorisiert werden.

Aus der Evaluation und Interpretation konnten wir feststellen, dass unsere Daten unausgeglichen waren. Durch die Berechnung der Frequenzbaseline, der Mittelwertbasline

und des Trainings unseres Naive-Bayes-Klassifikator konnten wir nochmals erkennen, dass für unsere Berechnungen die Klasse 2 (Severity-level 2) am relevantesten und am stärksten in unseren Daten vertreten ist. Durch die Lernkurven konnten wir erkennen, dass unser Modell weder zu Over- noch zu Underfitting neigt. Mittels Lime hat sich nochmals bestätigt, dass unser Datensatz unausgeglichen ist. Wir konnten in den Visualisierungen erkennen, dass die meisten Daten zur Severity 2 kategorisiert wurden. Um diese Problematik anzugehen, gibt es mehrere Strategien, die wir uns überlegt haben. Man könnte einerseits Daten mit Extremen Eckpunkten künstlich hinzufügen. Dies würde allerdings zu großem Rauschen führen, da die Daten nicht realitätsgetreu wären. Die wahrscheinlich bessere Strategie, wäre es mehr Daten zu finden, um die Anzahl der Severitys auszugleichen.

6. Fazit und Ausblick

Unser Data Analytics Projekt war eine herausfordernde Erfahrung. Trotz der Schwierigkeiten, die wir bei der Auswahl des Algorithmus und der Datenmenge hatten, konnten wir einen Einblick in die Daten und in das Thema gewinnen. Durch unsere Algorithmen konnten wir zufriedenstellende Ergebnisse erzielen. Die visuelle Darstellung war besonders erfreulich. Die Schwierigkeiten, die wir aufgrund unserer Daten hatten, nehmen wir als Lernerfahrung mit, die wir in zukünftigen Projekten besser umsetzen möchten.

Unseren Hauptanwendungsfall, die Vorhersage der Unfallschwere basierend auf verschiedenen Faktoren, wurden durch unsere Modelle mit annehmbarer Genauigkeit umgesetzt.

Für zukünftige Projekte mit diesem Datensatz gäbe es Bereiche, die verbessert werden könnten. Da wir mit einem sehr großen Datensatz gearbeitet haben, der zudem noch unausgeglichen war, wäre es eine Möglichkeit mehr Zeit in die Vorverarbeitung der Daten zu investieren. Dadurch könnte man bessere Ergebnisse im Bereich des Modellings erhalten. Weiterhin wäre es interessant gewesen, wenn uns noch mehr Datenkategorien zur Verfügung gestanden hätten. Wir hätten dadurch weitere Kategorien, wie Fahrzeugklasse oder weitere Personenbezogene Daten miteinbeziehen können, um unsere Modelle noch weiter spezifischer zu trainieren. Abschließend wäre es interessant noch mehr Algorithmen auszuprobieren und zu vergleichen, ob es einen besseren gibt, der zu unseren Anforderungen passen würde.

Literaturverzeichnis

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

https://www.kaggle.com/datasets/omer2040/usa-states-to-region

Anhang

Link zum GitHub Repository:

https://github.com/31foan1dpt/DataAnalytics

Vinko Jelic:

Während der Projektarbeit beschäftigte ich mich überwiegend mit den ersten Analysen der Datenbereinigung, Datenaufbereitung und der Modellierung. Hauptsächlich habe ich mit den Mindestanforderungen 1 bis 4 gearbeitet, sowie dem Clustering-Algorithmus. Durch die verschiedenen Herausforderungen habe ich zahlreiche Erkenntnisse gewonnen. Die Datenbereinigung erwies sich als eine der bedeutendsten Phasen, damit der Datensatz auf einen verlässlichen Stand gebracht wird. Dabei kam es zu unterschiedlichen Schwierigkeiten, wie z.B. ungültige („NaN“) Werte oder nicht realistische Werte (Temperaturen), sowie fehlende Daten. Der Umgang mit den Herausforderungen hat den Umgang mit realen Datensätzen verbessert und ein erweitertes Verständnis für die Bedeutung von sauberen Daten vermittelt. Die Umsetzung der verschiedenen Analyse-Abschnitten war sehr lehrreich. Das Selektieren von relevanten Features oder die Anwendung von Algorithmen, sowie die Visualisierung der Daten haben mir geholfen, ein besseres Gefühl für den Inhalt des Datensatzes zu erlangen. Das Identifizieren von Trends, zeitlichen Mustern oder der Schweren von Unfällen haben mir einen tieferen Einblick in die Datendynamik verschafft. Verständnisprobleme traten beispielsweise beim Clustering auf, da das Finden von geeigneten Features und der Anzahl dieser herausfordernd war. Auch mit geografischen Daten liefert das Clustering nicht immer unbedingt aussagekräftige Anschauungen. Ich habe gelernt, dass der Weg von der Datenbereinigung zum maschinellen Lernen nicht nur technische Komponenten beinhaltet, sondern auch ein Verständnis der Daten selbst erfordert. Die Arbeit mit verschiedenen Tools, Programmiersprachen und Bibliotheken wie Jupyter Notebook, Python, Pandas, NumPy, Scikit-learn etc. haben meine Programmierkenntnisse erweitert und gezeigt, dass diese Tools zur effektiven Datenanalyse und zum maschinellen Lernen effizient eingesetzt werden können. Das Projekt war eine zum Teil herausfordernde, aber definitiv lohnende Erfahrung, die das Verständnis von Big Data und dem maschinellen Lernen entscheidend erweitert hat.

Bruno Vidal dos Santos:

Die Projektarbeit hat mir einen spannenden Einblick in die Welt der Datenanalyse verschafft. Da ich vorher noch keine Erfahrung mit Jupyter Notebooks und Pandas hatte, musste ich mich zunächst in diese Themen einarbeiten. Die Einführung ist aufgrund der Anweisungen und Tutorials gut gelungen. Während der Projektarbeit lag mein Fokus zunächst darauf, die Daten zu verstehen und vorzubereiten, was die Beschaffung, Datenaufbereitung und -bereinigung einschloss.

Die Datentransformation zeigte mir wie ich Daten sortiere und gruppiere, um wichtige Zusammenhänge zu erkennen. Im Bereich des Modellierens beschäftigte ich mich mit der Dimensionalitätsreduktion, was anfangs aufgrund verschiedener Datentypen im Gesamtmodell herausfordernd war. Dennoch gelang es mir dank eines Encoders, die nicht-numerischen Werte erfolgreich umzuwandeln, was zu verbesserten Ergebnissen führte.

Obwohl anfänglich Probleme mit dem SVM-Klassifizierer auftraten, gelang es uns erfolgreich, diesen durch den leistungsstärkeren LGBM-Klassifizierer zu ersetzen. Dadurch wurden deutlich bessere und schnellere Ergebnisse erzielt. Abschließend widmete ich mich der Evaluation und Interpretation der Ergebnisse. Trotz der Schwierigkeiten und Frustration aufgrund unseres umfangreichen Datensatzes bin ich zufrieden mit dem erreichten Ergebnis und der Teamarbeit, da wir uns stets gegenseitig unterstützten.

Trotz anfänglichen Problemen mit GitHub und der Konfiguration war das Github-Repositorie eine super Idee um als Team gemeinsam am Code zu arbeiten und immer auf dem aktuellen Stand zu sein.

Abschließende kann ich sagen, dass das Generieren von Test- und Entwicklungsdaten, um daraufhin Vorhersagen auf Grundlage der Datenmerkmale zu treffen, hat mich sehr fasziniert. Es hat mich überrascht zu sehen, wie man mithilfe einzelner Dateninformationen bestimmte Dinge prognostizieren kann. Ebenso hat mir die Datenaufbereitung verdeutlicht, wie umfangreich die Bereinigung sein kann, um möglichst „richtige“ und „genaue“ Daten zu erhalten.

In zukünftigen Projekten im Bereich Data Analytics möchte ich gerne erneut diese Aufgabe übernehmen und die gewonnene Erfahrung vertiefen. Möglicherweise mit Vertiefung in Richtung Neuronale Netze und maschinelles Lernen.

Antonino Forte:

Die Projektarbeit gab mir einen interessanten Einblick in die Herausforderungen des Data Analytics. Innerhalb unseres Teams war ich für verschiedene Aufgaben verantwortlich. Im Bereich des Data Understandings und Data Preparation habe ich mich um die Datenexploration gekümmert, bei der ich die Daten Visuell darstellen durfte. Durch die visuelle Darstellung konnte ich ein besseres Verständnis für die Daten gewinnen, die uns zur Verfügung standen. Hier wurde mir klar, wie wichtig es ist, die Daten richtig zu interpretieren.

Im Bereich des Modellings war ich unter anderem verantwortlich für die Klassifikation und die Regression. Hier hatten wir die Schwierigkeit einen geeigneten Algorithmus für unseren Datensatz zu finden. Nach einigen Versuchen konnte ich allerdings einen geeigneten Algorithmus finden, der unseren Anforderungen entsprach und uns ein geeignetes Ergebnis lieferte. Der „LGBM“-Klassifizierer diente als Grundlagen für das gesamte Modelling-kapitel und war daher ein wichtiger Meilenstein für uns.

Ein Problem bestand in unseren Daten im Allgemeinen. Wir haben uns vorgenommen die Unfallschwere – „Severity“ vorherzusagen. Dies erfolgte allerdings nur mäßig, da unser Datensatz unausgeglichen verteilt war. Dies führte zu Frustration, da die Ergebnisse dadurch litten.

Neben den konkreten Projektaufgaben kümmerte ich mich auch um die Struktur außerhalb des Projekts. Die Erstellung eines Github-Repositories war für mich ein wichtiger Punkt, um Code einfach und schnell austauschen zu können.

Zusammenfassend kann gesagt werden, dass ich trotz der Herausforderungen und Schwierigkeiten sehr zufrieden mit unseren Ergebnissen bin. Ich konnte viel zum Umgang mit einem großen Datensatz lernen und wie man durch maschinelles Lernen Modelle auf bestimmte Anforderungen trainieren kann.

In meinem nächsten Projekt im Kontext von Data Analytics nehme ich mir vor mehr im Bereich des Modellings zu versuchen. Als Ziel setze ich mir hier ein besseres Verständnis für mehrere Algorithmen zu erlernen.