Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Explorative Datenanalyse und datenbasierte Modellierung eines Vorhersagemodells zur Ermittlung der monatlichen Kostenbelastung im RentSharing-Modell

Rick Lüdicke

Bachelorarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 24.06.2020

Hintergrund und Aufgabenstellung

Die AMS Fuhrparkmanagement GmbH bietet Dienstleistungen im Bereich RentSharing an. Ein Teil ihrer Webseite ist ein sogenannter Dienstwagenrechner (DWR), mit welchem sich der Nutzer die monatliche Kostenbelastung für ein oder mehrere Fahrzeuge ausrechnen kann. Momentan muss sich der Nutzer bei der Verwendung des DWRs sein Fahrzeug entweder manuell oder über einen Filter selektieren. Falls der Nutzer jedoch ohne eine ungefähre Vorstellung sein Fahrzeug auswählt, könnte er Gefahr laufen, ein für ihn ungeeignetes Fahrzeug auszuwählen. An dieser Stelle soll ein „Recommendation System" implementiert werden, welches dem Nutzer geeignete Vorschläge anbietet. Ziel der Arbeit ist es, einen Ausgangspunkt für ein solches System zu legen.

Erstellung des Datensatzes

Für den benötigten Datensatz müssen einige Anforderungen festgelegt und implementiert werden, um aussagekräftige Ergebnisse bei den darauf folgenden Schritten zu erlangen. Ein Eintrag in diesem Datensatz besteht aus der vom DWR berechneten mtl. Kostenbelastung und einigen selbst bestimmten Parametern inklusive Fahrzeugdaten, welche für die Berechnung benötigt werden. Bei den selbst bestimmten Werten wird auf eine möglichst hohe Varianz an möglichst realitätsnahen Werten geachtet. Des Weiteren wurden bei der Berechnung nur Fahrzeuge aus drei unterschiedlichen Preisklassen eingesetzt (~30tsd, ~40tsd. und ~50tsd.).

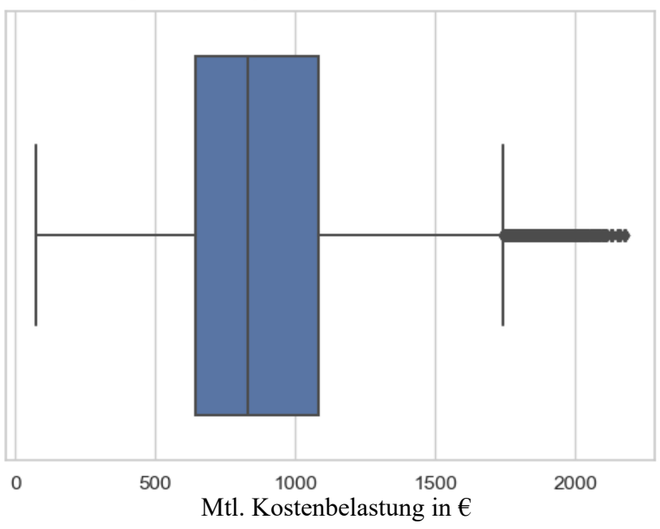

Abb. 1 Mtl. Kostenbelastung in einem Boxplot

Explorative Analyse

Eine derartige Analyse ist darauf ausgelegt, sich einen Gesamtüberblick über die zu untersuchende Datenmenge zu verschaffen. Hierbei wird der Fokus auf die berechnete mtl. Kostenbelastung gelegt, da der andere Teil der Daten vorbestimmt ist. Aus diesem Grund wird auch nur über die mtl. Kostenbelastung eine Fünf-Punkte-Zusammenfassung zuzüglich eines Boxplots (siehe Abb. 1) erstellt. Ziel der Analyse ist es außerdem, relevante Merkmale über ihren Bezug zur Zielgröße für alle weiteren Schritte in diesem Projekt herauszufiltern. Deshalb wurden Korrelationskoeffizienten nach Pearson berechnet und Graphen erstellt. Für die Arbeitsschritte, welche im Rahmen dieser Arbeit getätigt werden, werden jedoch alle Merkmale benötigt. Die Analyse wurde mithilfe der Python Bibliotheken Pandas und Seaborn/matplotlib durchgeführt.

Vorhersagemodell

Um ein geeignetes Setup (siehe Abb. 2), der aus einem Algorithmus für eine Regression besteht, eine Variante, wie mit nicht numerischen Merkmalen umgegangen werden soll, und eine Liste von Hyperparametern für die Algorithmuskonfiguration für ein Vorhersagemodell zu finden, müssen mehrere unterschiedliche getestet und evaluiert werden. Wie das genau vonstatten geht, ist der Abb. 2 zu entnehmen. Wichtig ist hierbei, dass die Performanz auch auf unbekannten Daten getestet und ausgewertet wird. Als Metrik, um die Performanz zu messen, wurde hier der Mean Absolute Error (MAE) genommen.

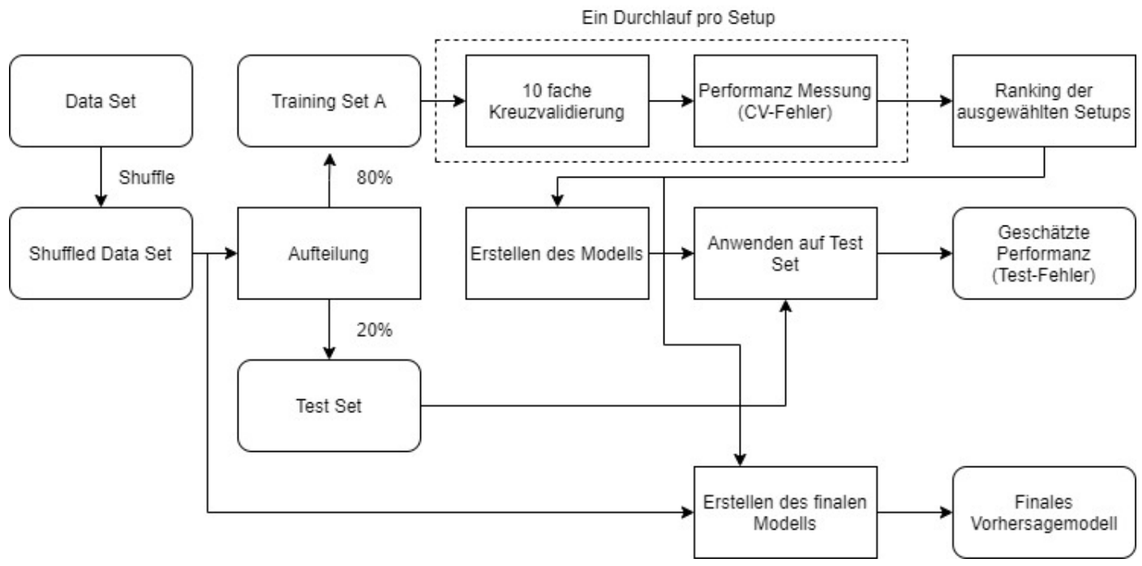

Abb. 2 Evaluationsprozess bei der Erstellung des Vorhersagemodells nach [1]

Aus der Evaluation, mit insgesamt 42 Setups, ergab sich, dass Setup mit einem Decision Tree Regressor (weitere getestete Reg. waren Random Forest, K Nearest Neighbor und Extreme Gradient Boosting) ohne kategorische Merkmale und ohne eine spezifische Algorithmuskonfiguration am performantesten mit einem Test-Fehler (MAE) von 1.302. Allgemein schnitten Setups ohne kategorische Merkmale am besten ab. Evaluiert wurde ebenfalls mit der Python Programmiersprache und einiger Bibliotheken (Scikit-learn, Pandas und Xgboost).

Fazit und Ausblick

Mithilfe der Beobachtungen aus der explorativen Analyse und dem finalen Vorhersagemodell kann mit der Entwicklung des Recommendation Systems begonnen werden. Eine mögliche Implementation könnte so aussehen, dass mithilfe des Vorhersagemodells dem Nutzer einige Fahrzeuge, aus einer großen Menge von Fahrzeugen, vor der eigentlichen Berechnung des DWRs vorgerechnet bzw. geschätzt werden und diese, falls der geschätzte Wert nicht zu stark vom berechneten abweicht, im Anschluss daran dann angezeigt werden. Um jedoch eine höhere Präzision bei der Bestimmung der Werte des Vorhersagemodells zu erlangen, müssen noch weitere Tests durchgeführt und noch mehr Setups auf weiteren Datensätzen evaluiert werden, da der erstellte Datensatz dieser Arbeit nicht repräsentativ für die Gesamtmenge an Fahrzeugen ist.

Quellen

[1] Ingo Boersch, Uwe Füssel, Christoph Gresch, Christoph Großmann, Benjamin Hoffmann: Data mining in resistance spot welding. Springer-Verlag, 2016

Betreuung Dipl. -Inform. Ingo Boersch • Technische Hochschule Brandenburg • Prof. Dr. Susanne Busse • Technische Hochschule Brandenburg