Machine Learning Elementarschaden-Hochrechnung

Studiengang: BSc in Informatik
Vertiefung: Data Engineering
Betreuerin: Prof. Dr. Mascha Kurpicz-Briki
Experte: Peter Matti

Meteorologische Extremereignisse wie Hagel, Hochwasser oder Stürme richten jährlich Millionenschäden an. Die Arbeit thematisiert eine automatisierte ML-Pipeline in Databricks, die Schadensbeschreibungen per NLP-Modul klassifiziert und historische Schadensdaten sowie Meteo-Schweiz-Radardaten kombiniert. Boosted Trees und Generalisierte Lineare Modelle erreichen deutlich niedrigere Vorhersagefehler unter Berücksichtigung von Meteo-Daten und Schadensbeschreibungen.

Ausgangslage

Die Versicherungsbranche steht der Häufigkeit und Intensität von Hagel-, Sturm- und Überschwemmungsschäden gegenüber. Manuelle oder regelbasierte Verfahren zur Ermittlung künftiger Schadenskosten sind zeitaufwendig, nicht skalierbar und oft ungenau. Gleichzeitig wächst das Datenvolumen durch digitale Schadensmeldungen rasant. Z. B. liegen Freitextbeschreibungen der Schadensmeldungen, historische Schadenstatistiken und hochaufgelöste Radardaten aus dem MeteoSchweiz-Netzwerk isoliert vor.

Ziel des Projekts

Ziel ist die Entwicklung einer skalierbaren End-to-End-Machine Learning (ML)-Pipeline auf Databricks:

1. – Natural Language Processing (NLP)-Vorverarbeitung: Die Schadensbeschreibungen werden durch Tokenisierung und Embeddings automatisiert in Schadenskategorien überführt, basierend auf einem KNN- oder LLM-Modell.
2. – Data-Lake-Integration: Historische Schadensfälle und MeteoSchweiz-Radardaten werden in einem zentralen Data Lake zusammengeführt.
3. – Feature Engineering & Modellierung: Extrahierte Textkategorien und meteorologische Parameter dienen als Features für Boosted-Trees-Modelle und Generalisierte Lineare Modelle (GLM).
4. – Reproduzierbarkeit und Deployment: MLflow übernimmt Modell-Versionierung und automatisiertes Deployment.

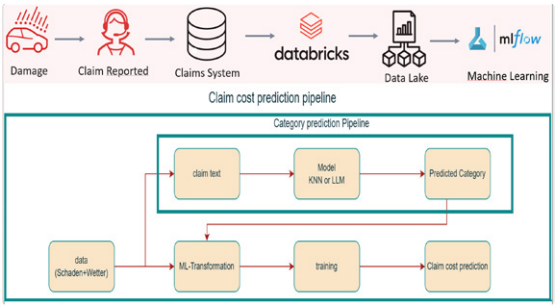
Ergebnisse

Die implementierte Pipeline wurde erfolgreich in Databricks bereitgestellt und ermöglicht insgesamt deutlich präzisere Kostenprognosen als das bisherige Baseline-Modell. Die automatisierte Textklassifikation liefert zuverlässige Schadenskategorien und die Kombination mit Wetterdaten führt zu robusteren Vorhersagen. Reproduzierbare Trainings- und Inferenz-Workflows über MLflow und Databricks-Jobs sichern Skalierbarkeit und Stabilität.
Eine präzise Vorhersage der Schadenskosten ermöglicht es einer Versicherungsgesellschaft, die Schadensabwicklungen zu verkürzen, um somit die Kundenzufriedenheit zu steigern.

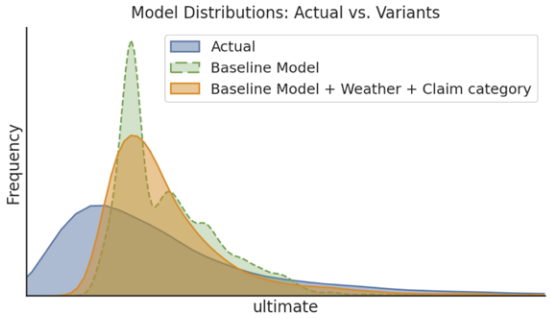
Ausblick

- Erweiterung auf weitere Schadenstypen: Neben Hagel werden künftig auch Sturm- und Überschwemmungsschäden modelliert.
- NLP-Vertiefung: Direkte Einbindung von Text-Embeddings aus Schadensbeschreibungen in Kostenprognosemodelle, etwa via Transformer-Features
- Produktlinien-Integration: Rollout auf Gebäude und weitere Versicherungssparten
- Echtzeit-Monitoring: Aufbau eines interaktiven Dashboards zur Visualisierung von Ist- und Prognosewerten

Matin Mahmoudzadeh
matin_mhz@outlook.com



End-to-End-Prozess von Schadenmeldung bis Kostenprognose mit KNN oder LLM



Die echte Verteilung der Daten (blau) im Vergleich zu den Vorhersagen von Modell 1 (grün) und Modell 2 (orange)