

Moderne Ansätze des maschinellen Lernens zur Vorhersage von Cash-Flows in Versicherungsverträgen

Autor: Can Berkil

Referent: Prof. Dr. Thümmel

Koreferent: Prof. Dr. Schürg

Hochschule Darmstadt, University of Applied Sciences, Fachbereich Mathematik und Naturwissenschaften & Informatik

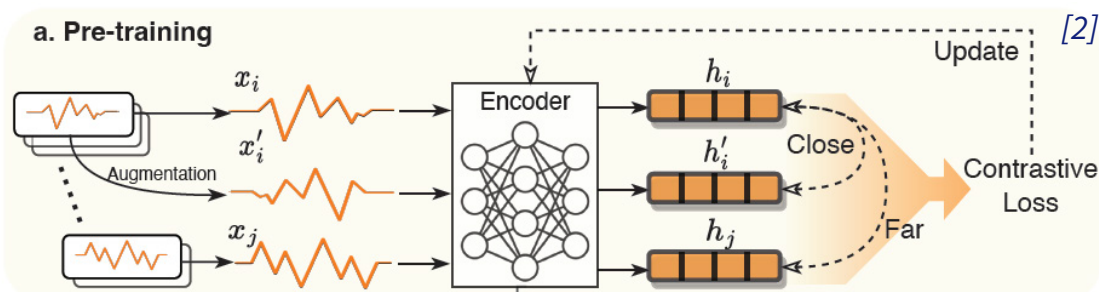
Motivation und Hintergrund

Die präzise Vorhersage von Cashflows stellt eine zentrale Herausforderung für Unternehmen dar, insbesondere im Kontext von Lebensarbeitszeitkonten (LAZ), da Einzahlungsverhalten und Muster stark variieren. Klassische Prognoseverfahren stoßen hierbei an ihre Grenzen, da sie oft nicht in der Lage sind, komplexe und unregelmäßige Muster präzise zu erfassen.

Um dieses Problem zu adressieren, wird in dieser Arbeit Self-Supervised Learning (SSL) mithilfe von TS2Vec untersucht. Ziel ist es, durch automatisch generierte Repräsentationen verborgene Muster in den Zeitreihen zu extrahieren und die Vorhersagequalität zu verbessern.

Besonders innovativ ist die Erweiterung von TS2Vec um eine Frequenzdimension sowie die Einführung neuer Augmentationsmethoden mit unterschiedlichen Intensitätsstufen, um die Robustheit der Modelle zu erhöhen. Zudem wird überprüft, ob ein Finetuning und ein Hyperparameter-Tuning die Modellperformance weiter steigern können.

Contrastive Learning



TS2Vec nutzt Contrastive Learning, um aussagekräftige Repräsentationen von Zeitreihen zu erzeugen. Dabei werden ähnliche Segmente näher zusammengebracht, während unähnliche weiter auseinander gehalten werden. Dies geschieht durch die Erstellung positiver Paare, bestehend aus einer Originalzeitreihe und ihrer augmentierten Version, sowie negativer Paare aus nicht verwandten Zeitreihen. Die Optimierung erfolgt über die Contrastive Loss-Funktion, die sicherstellt, dass die Repräsentationen relevante Muster erfassen und für Vorhersageaufgaben nutzbar sind.

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(h_i, h'_i)/\tau)}{\exp(\text{sim}(h_i, h'_i)/\tau) + \sum_{j \neq i} \exp(\text{sim}(h_i, h_j)/\tau)}$$

h_i : Repräsentation der originalen Zeitreihe.

h'_i : Repräsentation der augmentierten Zeitreihe.

h_j : Repräsentation negativer Paare, die nicht mit h_i verwandt sind.

τ : Temperaturparameter zur Skalierung.

$\text{sim}()$: Ähnlichkeitsfunktion (z. B. Cosine Similarity).

Referenzen

[1] Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong und Bixiong Xu. TS2Vec: Towards Universal Representation of Time Series. 2022. arXiv: 2106.10466 [cs.LG]. url: <https://arxiv.org/abs/2106.10466>.

[2] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis und Marinka Zitnik. Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency. 2022. arXiv: 2206.08496 [cs.LG]. url: <https://arxiv.org/abs/2206.08496>.

[3] Ziyu Liu, Azadeh Alavi, Minyi Li und Xiang Zhang. Self-Supervised Learning for Time Series: Contrastive or Generative? 2024. arXiv: 2403.09809 [cs.LG]. url: <https://arxiv.org/abs/2403.09809>

[4] Guillermo Iglesias "Data Augmentation techniques in time series domain: a survey and taxonomy". In: Neural Computing and Applications 35,14 (März 2023), 10123–10145. issn: 1433-3058. doi:10.1007/s00521-023-08459-3. url: <http://dx.doi.org/10.1007/s00521-023-08459-3>

Daten

Pretraining-Daten (TS2Vec):

- Zeitreihen aus einer anderen Domäne (z. B. Energieverbrauchsdaten).
- Genutzt, um allgemeine Muster in Zeitreihen zu lernen, bevor das Modell auf spezifische Cashflow-Daten angewendet wird.
- Dient als erste Evaluationsbasis für die neuen Augmentationsmethoden.

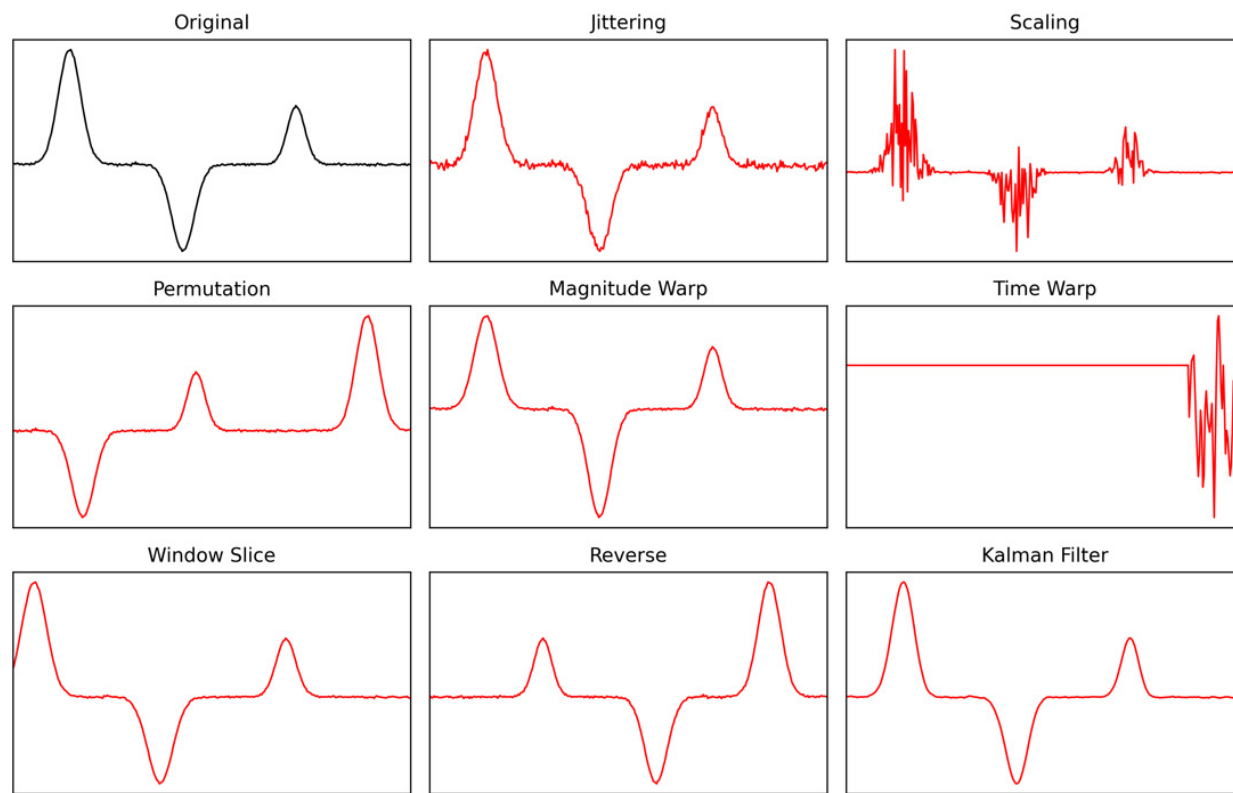
Generische Testdaten:

- Simulierte Sinuskurven mit und ohne Ausreißer.
- Verwendet zur Evaluierung der Modellrobustheit gegenüber künstlich generierten Mustern.
- Ziel: Sicherstellen, dass die erzeugten Repräsentationen allgemeingültig sind.

Cashflow-Daten aus Lebensarbeitszeitkonten:

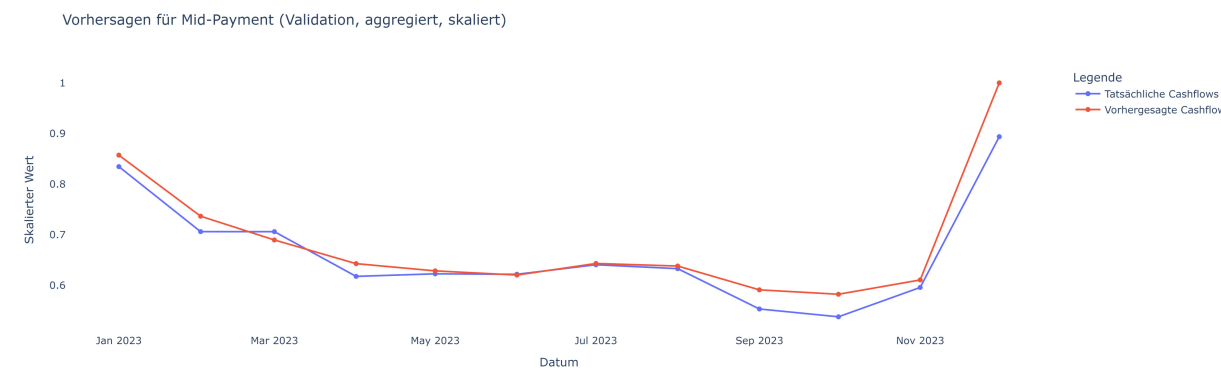
- Reale Einzahlungsdaten von Unternehmen in Zeitreihenform.
- Gruppierung in Low-Payment, Mid-Payment und High-Payment Kunden zur besseren Modellanpassung.
- Charakterisiert durch starke Schwankungen, saisonale Muster und unregelmäßige Einzahlungen.

Methoden und Erweiterungen



In das Contrastive Learning des TS2Vec-Ansatzes wurden verschiedene Augmentationsverfahren mit unterschiedlichen Intensitätsstufen integriert. Ziel dieser Erweiterung ist es, zu untersuchen, welche Augmentationsvarianten und -stärken die Pretext-Task anspruchsvoller gestalten und dadurch die Generalisierungsfähigkeit des Modells verbessern. Durch gezielte Verzerrungen in den Zeitreihen wird das Modell dazu angeregt, robustere Repräsentationen zu lernen, die auch unter veränderten Bedingungen zuverlässige Muster extrahieren. Dabei wurde sowohl die Zeitdimension als auch die Frequenzdimension betrachtet, um die Auswirkungen verschiedener Transformationsarten zu analysieren. Besonders die komplexeren Augmentationen, wie Time-Warping oder Permutation, haben sich als effektive Methoden erwiesen, um das Modell gezielt herauszufordern und so die Qualität der gelernten Repräsentationen zu steigern.

Ergebnisse und Herausforderungen



Model	Group	Repr_Mean_MAE	Cashflow_Mean_MAE
jittering_schwer tuned	High-Payment	0,25	104320,2
add_frequency_schwer tuned	Mid-Payment	0,08	4047,8
add_frequency_schwer tuned	Low-Payment	0,13	536,3

Die Analyse zeigt, dass die **Frequenzdimension in Kombination mit starken Augmentationen bereits vor dem Finetuning signifikant bessere Repräsentationen erzeugt** als das originale TS2Vec-Modell. Besonders **komplexe Transformationen wie Scaling-Frequency oder Add-Frequency haben die Qualität der Repräsentationen erheblich verbessert**.

Das **Finetuning hat die Modellleistung nochmals deutlich gesteigert**, indem es die Repräsentationen besser an die Cashflow-Daten angepasst hat. Im Gegensatz dazu hat das **Hyperparameter-Tuning keine signifikante Verbesserung gebracht**.

Trotz der Erfolge bestehen weiterhin Herausforderungen:

- Low-Payment-Gruppe** → Viele Nullwerte und unregelmäßige Einzahlungen erschweren die Modellierung, da Zahler und Nicht-Zahler nicht klar getrennt werden können.
- High-Payment-Gruppe** → Extreme Schwankungen und hohe Peaks führen dazu, dass das Modell plötzliche Einbrüche nicht präzise genug erfasst..

Ausblick

Neuen Die Ergebnisse zeigen, dass **komplexe Augmentationen und die Frequenzdimension** die Repräsentationen deutlich verbessern konnten. Dennoch bestehen Herausforderungen bei der Trennung von Zahlern und Nicht-Zahlern sowie der Erkennung plötzlicher Rückgänge.

Mögliche Optimierungen:

- Dynamische Gruppierung** statt fixer Kategorien zur besseren Anpassung an Datenverteilungen.
 - Gruppenspezifisches Finetuning** für präzisere Vorhersagen.
 - Management-Integration** zur frühzeitigen Identifikation kritischer Muster.
- Diese Ansätze könnten die Modellleistung weiter verbessern und stabilere Cashflow-Prognosen ermöglichen.Text hinzufügen