ABSTRACT

This thesis investigates the feasibility of forecasting customer consumption for SAP's Business Technology Platform (BTP) Core using historical consumption data and multiple models involving multivariate and univariate approaches. The primary objective is to develop predictive models that accurately forecast monthly cloud consumption, with the ultimate goal to provide accurate consumption forecasts to optimize SAP's customer engagement strategies and contract renewals.

To achieve this goal, various models were applied to two prepared, cleaned and enriched customer datasets which consisted of SAP's customer consumption data based on a 31 month time frame. The forecasting horizon was set to 3 months. The models applied included naive forecasting, XGBoost, Linear Regression (both as a 'global' model and on a per-customer basis), and Prophet. Prophet was used to train individual models for each customer to capture time-dependent trends and seasonality. The datasets were aggregated, then enriched with additional features, such as derived time features and customer-specific data. Model performance was evaluated using several metrics, with a primary focus on Mean Absolute Error (MAE).

The naive forecasting model outperformed all other models applied to the datasets based on the performance metrics. It achieved a MAE of €138, benefiting from strong autocorrelation in the time series data for short-term forecasting in combination with many zero values. Compared to XGBoost's MAE of €298 and Linear regression with €390 for the 'global' model and €275 MAE for the averaged customer-based models. Prophet significantly underperformed relative to the other models, likely due to the highly skewed dataset with many zero values. This imbalance made it difficult for Prophet to extract meaningful trends and seasonality from such a small dataset, making it less suitable for accurate monthly consumption forecasting.

The strong autocorrelation in the first lags of the datasets, combined with the high occurrence of zeros and constant time series, favored simpler models like the naive forecast. While the naive model showed more promising results, XGBoost could generate more insights into feature importances and potential for capturing non-linear relationships, more advanced models like this one could benefit from larger datasets and extended forecast horizons while focusing on feature engineering.

This research highlights the challenges and opportunities in cloud consumption forecasting, emphasizing the need for ongoing refinement of model selection and feature engineering. Future work should explore hybrid modelling approaches and greater data granularity to enhance forecasting precision and support SAP's strategic goals.

ZUSAMMENFASSUNG

In dieser Arbeit wird die Machbarkeit der Vorhersage des Euro Cloudverbrauchs von SAP Business Technology Platform (BTP) Kunden anhand historischer Verbrauchsdaten und mehrerer Modelle mit multivariaten und univariaten Ansätzen untersucht. Ziel ist die Entwicklung von Modellen, die den monatlichen Cloud-Verbrauch prognostizieren, um SAPs Kundenbindungsstrategien zu optimieren. Um dieses Ziel zu erreichen, wurden verschiedene Modelle auf zwei aufbereitete Datensätze angewandt, welche monatliche Nutzungsdaten von SAP-Kunden über einen Zeitraum von 31 Monaten enthielten. Der Prognosehorizont wurde auf 3 Monate festgelegt. Zu den angewandten Modellen gehörten ein naives Vorhersagemodell, XGBoost, lineare Regression (sowohl als 'globales' Modell als auch auf Einzelkundenbasis) und Prophet. Die Datensätze wurden aggregiert, dann mit zusätzlichen Merkmalen wie abgeleiteten Zeitmerkmalen und kundenspezifischen Daten angereichert und verarbeitet. Die Leistung des Modells wurde anhand verschiedener Metriken bewertet, wobei der Schwerpunkt auf dem mittleren absoluten Fehler (MAE) lag. Das naive Prognosemodell lieferte in Bezug auf die Leistungsmetriken die besten Ergebnisse und übertraf alle anderen angewandten Modelle. Es erreichte einen MAE von €138 und profitierte von der starken Autokorrelation in den Zeitreihendaten für kurzfristige Vorhersagen und der hohen Anzahl von Nullen. Verglichen mit dem MAE von XGBoost von €298 und der linearen Regression mit €390 für das 'globale' Modell und €275 MAE für die gemittelten kundenbasierten Modelle. Prophet schnitt im Vergleich zu den anderen Modellen deutlich schlechter ab, was wahrscheinlich auf den stark verzerrten Datensatz mit einer großen Anzahl von Nullwerten zurückzuführen ist. Aufgrund dieser Unausgewogenheit war es für Prophet schwierig, aussagekräftige Trends und saisonale Schwankungen aus einem so kleinen Datensatz zu extrahieren, was es für eine genaue monatliche Verbrauchsprognose weniger geeignet machte.

Die starke Autokorrelation in den ersten Lags der Datensätze in Verbindung mit dem hohen Vorkommen von Nullen und konstanten Zeitreihen begünstigte einfachere Modelle wie die naive Prognose. Während das naive Modell vielversprechendere Ergebnisse zeigte, konnte XGBoost mehr Erkenntnisse über die Bedeutung von Merkmalen und das Potenzial zur Erfassung nichtlinearer Beziehungen liefern. Diese Forschung zeigt die Herausforderungen und Möglichkeiten bei der Vorhersage des Cloud-Verbrauchs auf und unterstreicht die Notwendigkeit einer sorgfältigen Modellauswahl und des Feature Engineerings. Zukünftige Arbeiten sollten hybride Modellierungsansätze, granularere Daten und weitere Methoden beleuchten, um die Prognosegenauigkeit zu verbessern und die strategischen Ziele von SAP zu unterstützen.