h_da
HOCHSCHULE DARMSTADT
UNIVERSITY OF APPLIED SCIENCES

Above the Clouds: SAP Success by Usage Forecasting

Cedric Schmith

Supervisors: Prof. Dr. Timo Schürg, Prof. Dr. Christoph Becker

Introduction

In the era of cloud computing, accurate consumption forecasting is critical for optimizing resource allocation and driving data-driven decision-making[1]. This research focuses on developing a robust model to forecast customer consumption for SAP's Business Technology Platform (BTP) Core KPI, specifically targeting subscription-based customers.

Leveraging machine learning techniques such as XGBoost[2], Prophet[3], and linear regression models, the research aims to predict the euro-based consumption values at a customer level. The forecast enables SAP's sales workforce to identify trends, prevent contract non-renewals, and explore upsell opportunities, supporting a more proactive approach to customer relationship management.

This research not only addresses the forecasting challenge but also investigates key consumption-driving features, aiming to enhance both prediction accuracy and interpretability.

Methodology

This study follows the CRISP-DM[4] methodology, a structured approach to data science projects, encompassing six phases: **Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment**.

Data Sources: Customer-level data for SAP BTP Core consumption was sourced from HANA databases, accessed through REST APIs, and loaded into Databricks via JDBC. The dataset includes 31 months of data with over 1,100,000 entries.

Data Preprocessing:

- Cleaning and aggregation of time-series data.
- Handling categorical features using target encoding and enabling categorical handling in XGBoost.
- Splitting data into training and testing sets based on time (pre-2024 for training, post-2024 for testing).
- Feature engineering, derived features, time-related features, customer data.

Tools and Techniques:

- **Databricks:** Data processing and analysis.
- **HANA:** Data storage and retrieval.
- **XGBoost, Prophet, and Linear Regression:** Modeling approaches.
- TimeSeriesSplit and Hyperopt for cross-validation and hyperparameter tuning.

By systematically following CRISP-DM, the study ensured a robust and replicable process for forecasting customer consumption and identifying key consumption drivers.

Modeling Approaches

This study implemented four modeling approaches to forecast customer consumption for SAP BTP Core services, applied across forecasting horizons of 1, 2, and 3 months using two distinct datasets:

1. Naive Forecasting: Naive forecasting was used as a baseline model, where the most recent observation was carried forward as the prediction. This simple approach provided a benchmark to evaluate the performance of more advanced models.

2. XGBoost: XGBoost, a gradient boosting algorithm, was employed as the primary model due to its ability to handle large datasets and capture complex non-linear relationships. Forecasts were generated for 1-, 2-, and 3-month horizons. Categorical variables were processed using `enable_categorical=True`, and hyperparameter optimization was performed using Hyperopt.

3. Prophet: The Prophet model was applied as a univariate time-series forecasting approach, with separate models trained for each customer. It effectively modeled individual trends and seasonality for varying forecasting horizons.

4. Linear Regression: Linear regression was used as a simple model, offering interpretability and serving as a benchmark. Models were trained for generalized predictions across all customers as well as for individual customer-level forecasts.

Datasets and Features:

- **Dataset 1:** A complete dataset where missing months were populated to provide a continuous historical view for all customers.
- **Dataset 2:** A subset of customers that maintained at least one active contract throughout the entire historical period, enabling focused analysis.

The features were consistent across datasets but were tailored differently for each model to leverage their unique strengths.

Model Comparison: All models were evaluated using metrics such as Mean Abosulte Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and R^2.

Results

Key Findings:

- **Naive Forecasting:** The naive model performed best, achieving a Mean Absolute Error (MAE) of €138. This success was attributed to the strong autocorrelation in the time series data, combined with a high number of zero values, making it effective for short-term forecasting.
- **XGBoost:** XGBoost demonstrated solid performance with an MAE of €298. While it was outperformed by the naive model, it provided valuable insights into feature importances and non-linear relationships, showing potential for more complex forecasting with larger datasets and extended horizons.
- **Linear Regression:** The global model showed an MAE of €390, while customer-specific models achieved a better MAE of €275. Although simpler, Linear Regression provided interpretability and consistency but lagged behind XGBoost in predictive accuracy.
- **Prophet:** Prophet performed poorly with the given dataset, largely due to the high imbalance (many zero values) and small dataset size, which hindered the extraction of meaningful trends and seasonality.

The graphic compares the performance of all models across the datasets and forecasting horizons, highlighting the superiority of the naive forecast for this particular dataset, while also showcasing the potential of XGBoost for more complex forecasting tasks.

Conclusion: While simpler models like the naive forecast yielded the best results, more advanced models such as XGBoost could benefit from larger datasets and more refined feature engineering. Future work should focus on hybrid models and improving data granularity to enhance forecasting accuracy for SAP's strategic goals.

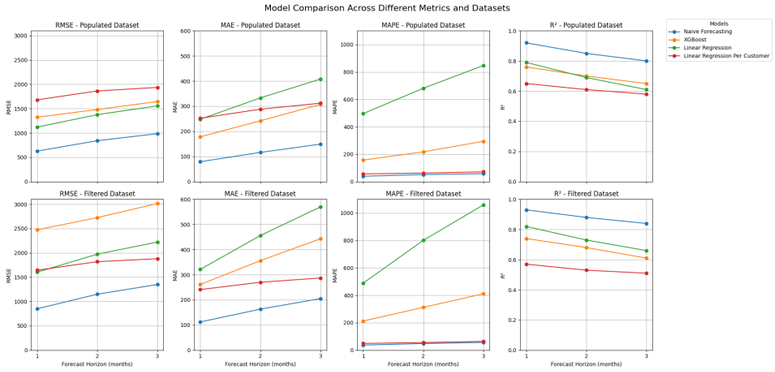

Figure 1. Model Comparison across different Metrics and Datasets

Feature Analysis

In this study, feature importance analysis was conducted using the XGBoost model to identify the most influential variables in predicting customer consumption. The feature importance scores are derived based on the model's internal decision-making process, highlighting which features contribute the most to the prediction.

Key Insights:

- The feature importance graphs reveal the relative contribution of each feature in driving the consumption forecasts.
- Key features identified include *customer-specific variables*, such as contract duration and usage patterns, as well as time-related features like month and seasonality indicators.
- Some categorical variables, like *COUNTRY* emerged as highly significant, reflecting the geographic and market-specific factors that influence consumption behaviour.
- Derived features, such as time-lagged consumption values, also showed substantial importance, indicating the predictive power of historical consumption trends.

These findings underscore the value of advanced feature engineering and the ability of XGBoost to capture complex relationships within the data. The insights provided by the feature importance graphs guide the refinement of future models and highlight areas for further exploration, such as the inclusion of additional temporal or customer segmentation features to improve prediction accuracy.

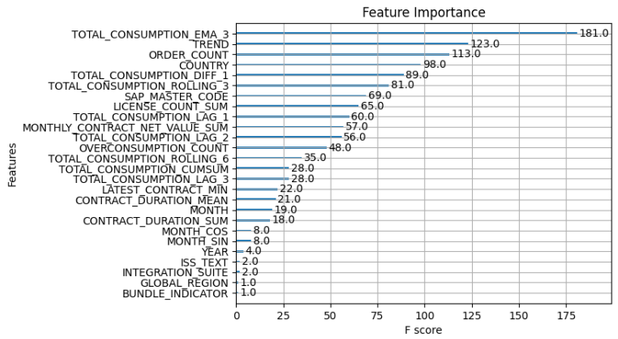

Figure 2. Feature Importance for Horizon 1 using the Populated Dataset.

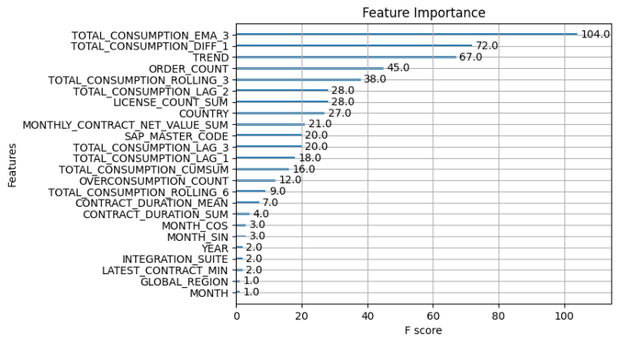

Figure 3. Feature Importance for Horizon 1 using the Filtered Dataset.

Conclusion and Future Work

This study highlights the challenges and opportunities in forecasting SAP's Business Technology Platform (BTP) Core KPI consumption. Naive Forecasting excelled due to strong autocorrelation, while XGBoost identified key features and showed potential for long-term predictions with larger datasets. Linear Regression offered interpretability but struggled with complexity, and Prophet underperformed due to data sparsity.

Key Takeaways:

- Simpler models like Naive Forecasting are effective for short-term predictions.
- XGBoost holds promise for capturing trends and driving insights with more data.
- Feature engineering and hybrid approaches can further enhance accuracy.

Future Work: Expanding datasets, refining feature engineering, and exploring hybrid and real-time modeling approaches will support SAP's strategic goals, including boosting customer engagement, optimizing consumption patterns, and enhancing contract management.

References

1. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... Zaharia, M. (2010). *A view of cloud computing*. Communications of the ACM, 53(4), 50-58.

2. Chen, T., Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

3. Taylor, S., Letham, B. (2017). Prophet: forecasting at scale. PeerJ Preprints, 5, e3190. https://doi.org/10.7287/peerj.preprints.3190v1

4. Shearer, C. (2000). *The CRISP-DM model: The new blueprint for data mining*. Journal of Data Warehousing, 5(4), 13-22.