Data Science Project Overview: Customer Prediction Using Time Series Analysis

Goals

The primary goal of this project was to predict the number of customers for a business using time series data. Accurately forecasting customer numbers would enable the company to optimize resource allocation, manage inventory, and design targeted marketing campaigns. A secondary goal was to analyze the impact of various external factors, such as weather conditions and online engagement metrics, on customer behavior over time.

Data Sources

Data was collected from multiple sources to build a comprehensive dataset for the analysis:

1. **CRM Dataset**: Provided historical data on customer numbers, including daily customer transactions.
2. **Meteo Data (Web Scraping)**: Weather data, such as temperature and precipitation, was obtained by scraping meteorological websites using BeautifulSoup.
3. **SQL Queries on GCP**: Additional customer data, including historical, was extracted from the company's databases hosted on Google Cloud Platform (GCP).
4. **Google Analytics API**: Captured online engagement metrics, including the number of website views and other user interaction data.

Process Steps

1. **Data Collection**: We aggregated data from the CRM system, weather sources, GCP, and Google Analytics into a centralized database (a DataMart in GCP). This involved automating the extraction of data via web scraping, API requests, and SQL queries.
2. **Data Cleaning**: This phase involved handling missing data, correcting anomalies, and aligning the temporal data from various sources. Missing weather data was imputed using interpolation techniques, while missing customer data was forward-filled.
3. **Exploratory Data Analysis (EDA)**: We performed EDA to identify trends, seasonality, and potential outliers in the data. Techniques used included line plots, seasonal decomposition, and autocorrelation plots to examine the relationships between the number of customers and other variables over time.
4. **Feature Engineering**: Several new features were engineered to enhance the predictive power of the models. For example, we calculated rolling averages of customer numbers, lagged variables for temperature and precipitation, and interaction terms between online engagement and weather conditions.
5. **Model Selection**: We evaluated several time series forecasting models, including SARIMA (Seasonal Autoregressive Integrated Moving Average), LSTM (Long Short-Term Memory networks), and Facebook Prophet. The objective was to find the model that best captured the temporal dynamics and seasonality in the data.

6. **Model Training and Evaluation**: The models were trained on a portion of the historical data and evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). We performed cross-validation to assess the models' generalizability.
7. **Model Tuning**: Hyperparameter tuning was performed for each model to optimize its performance. For SARIMA, we tuned parameters such as p, d, q, and seasonal components. For LSTM, we adjusted the number of layers, neurons, and dropout rates. Prophet's seasonalities and changepoints were fine-tuned to capture underlying patterns more accurately.
8. **Interpretation of Results**: We analyzed the results to understand the factors influencing customer numbers. Techniques such as feature importance analysis in LSTM and component analysis in Prophet were employed to interpret the models.

Techniques Used

- **SARIMA and Prophet**: These models were chosen for their ability to handle seasonality and trends in time series data. SARIMA provided a statistical approach, while Prophet allowed for easy inclusion of holidays and external regressors.
- **LSTM Networks**: LSTM, a deep learning model, was selected for its ability to capture long-term dependencies in the time series data, particularly useful in recognizing complex patterns that might be missed by traditional models.

Challenges and Unexpected Outcomes

One significant challenge was handling the diverse nature of the data sources, each with its own format and temporal resolution. For instance, aligning daily customer data with hourly weather data required careful preprocessing.

Another challenge was the interpretability of the LSTM model compared to SARIMA and Prophet. While LSTM performed well in terms of predictive accuracy, explaining its decisions to stakeholders was more difficult. SARIMA and Prophet, on the other hand, provided more intuitive insights into seasonal trends and external factors like weather.

Final Conclusions

The project successfully developed models to predict the number of customers, with the best-performing model achieving a MAPE of approximately 10%, which was deemed satisfactory for operational use. The results highlighted the significant impact of weather conditions and online engagement on customer numbers. These insights could inform future marketing and operational strategies, such as increasing advertising spend during periods of low customer traffic or adjusting staffing levels based on weather forecasts.

Learnings

1. **Integration of Diverse Data Sources**: Combining data from different sources required careful preprocessing and alignment but was crucial for building a robust predictive model.
2. **Model Selection for Time Series**: Different time series models have distinct strengths.
3. **Tools**: GIT for versioning, Asana for project management, jupyter notebook, python, …