Summary

In my analysis of the merged dataset, which combines client_df and price_df, I discovered that there are more users in price_df than in client_df. To ensure data consistency, we decided to remove the data of those users in price_df that were not present in client_df.

My main focus in this analysis was to understand customer churn, which is reflected in the "churn" column of client_df. I observed that approximately 90.3% of the observations correspond to customers who did not churn, indicating a significant class imbalance in the dataset. Handling this imbalance is crucial for obtaining reliable insights and making better decisions in predicting churn.
Approximately 10% of customers have churned

Regarding the feature types in client_df, I found that it contains different kinds of information. Some features are words (strings), such as "id," while others are numbers. For example, "date_modif_prod" and "date_renewal" have 2,129 and 386 unique values, respectively. The discrete features, like "forecast_discount_energy," "nb_prod_act," "num_years_antig," and "churn," have 12, 10, 13, and 2 unique values, respectively. Additionally, I noticed various continuous features in client_df, such as "cons_12m," "cons_gas_12m," "cons_last_month," and "forecast_cons_12m," each with different unique values.

Similarly, in price_df, I found a mix of string, discrete, and continuous features. "id" is a string feature with 16,096 unique values. The discrete feature "price_date" has 12 unique values, while the continuous features, like "price_off_peak_var," "price_peak_var," "price_mid_peak_var," "price_off_peak_fix," "price_peak_fix," and "price_mid_peak_fix," have varying numbers of unique values.

My analysis also involved examining the correlation between features in client_df and price_df. I identified high positive correlations between several pairs of features in both datasets, such as "margin_gross_pow_ele" and "margin_net_pow_ele," "cons_last_month" and "cons_12m," among others. Similarly, I found high negative correlations between certain feature pairs, like "forecast_price_energy_off_peak" and "forecast_meter_rent_12m."

Client data is highly skewed with some multimodal features and must be treated before modeling. There are outliers present in the data and these must be treated before modeling. Hence, Feature Engineering is required before modeling to increase predictive power of model.

Furthermore, I observed that the "Churned" and "Not Churned" observations in the dataset significantly overlap. This indicates that a linearly separable classifier may not be appropriate for modeling the data. To improve predictions, I might need to explore more advanced techniques, such as non-linear classifiers, to better capture the complex relationships between features and the target variable.

Suggestions

1. Competitor price data - perhaps a client is more likely to churn if a competitor has a good offer available?
2. Average Utilities prices across the country - if PowerCo's prices are way above or below the country average, will a client be likely to churn?
3. Client feedback - a track record of any complaints, calls or feedback provided by the client to PowerCo might reveal if a client is likely to churn