Customer Churn Analysis Report for PowerCo

---

1 Introduction

PowerCo, a leading utility provider for gas and electricity, serves small and medium-sized enterprises in a highly competitive energy market. In recent years, the industry has seen rapid changes, offering customers numerous options to switch providers. This has intensified competition, and PowerCo is now facing a significant challenge: customer churn.

Customer churn occurs when clients leave PowerCo to seek better pricing, improved services, or other benefits offered by competitors. The impact of churn is critical, directly affecting revenue and market share. Recognizing the urgency of this issue, PowerCo has engaged our team to analyze factors influencing churn and provide actionable insights to mitigate it.

Among the potential drivers of churn, price sensitivity has emerged as a key area of interest. Specifically, PowerCo suspects that changes in off-peak prices between December and January could significantly influence customer retention decisions. This report aims to determine the extent to which price sensitivity impacts churn and assess other contributing factors.

---

2 Problem Statement

The objective of this analysis is to evaluate whether price sensitivity is the most significant factor driving customer churn. If not, we aim to:

1. Quantify the influence of price sensitivity relative to other factors.
2. Investigate other drivers of churn, such as consumption patterns, contract details, and customer demographics.

To address these goals, we:

- Conducted Exploratory Data Analysis (EDA) to understand key trends.
- Developed meaningful features that capture price and customer behavior.
- Built predictive models to identify factors most associated with churn.

---

3 Exploratory Data Analysis (EDA)

3.1 Data Exploration

1. **Missing and Duplicate Values**:
    - Missing values: None detected across the dataset.

    - Duplicate entries: Identified and removed to ensure data consistency.
2. **Categorical Variables**:
    - Variables like channel_sales, origin_up, and churn were reviewed for consistency.
    - No anomalies detected in categorical labels.
3. **Numerical Variables**:
    - Outliers observed in features like net_margin, pow_max, and imp_cons.
    - Significant skewness in features such as forecast_cons_12m and imp_cons.

3.2 Observations and Insights

Distribution of Churn Variable

- **Observation:**
    - Significant class imbalance:
        - Non-Churn (0): 90.28% of clients.
        - Churn (1): 9.72% of clients.
    - Class imbalance may lead to biased machine learning model performance.
- **Insights:**
    - Models might favor the majority class (non-churn), reducing accuracy for churn predictions.
    - Churn rate (9.72%) highlights a critical area for customer retention strategies.

Sales Channels Distribution

- **Observation:**
    - Dominant sales channels: Channels 4 and 0 account for over 70% of clients.
    - Underperforming channels: Channels 6, 1, and 3 contribute a negligible proportion.
- **Insights:**
    - High-performing channels warrant further investment.
    - Underperforming channels may require optimization or targeted improvement initiatives.

Acquisition Campaigns Distribution

- **Observation:**
    - Campaign 4 is the most effective (48.59% of clients).

    - Campaigns 2 and 3 collectively contribute 50.95% of clients.
    - Minimal impact from campaigns 0, 5, and 1 (<0.5% of clients).
- **Insights:**
    - Focus on successful campaigns to maximize client acquisition.
    - Investigate and optimize the performance of less impactful campaigns.

Churn by Sales Channels

- **Observation:**
    - Channels 1, 3, and 6: Perfect retention.
    - Channel 4: Highest churn rate (12.1%).
- **Insights:**
    - Strategies from high-retention channels could inform improvements in low-retention channels.
    - Addressing issues in Channel 4 may significantly reduce churn.

Contract Start Years

- **Observation:**
    - Peak years: 2010 and 2012 (50% of contracts).
    - Early years (2003-2008): High retention and low churn.
    - Post-2009 cohorts show increased churn rates.
- **Insights:**
    - Increased churn in later cohorts may stem from intensified competition or market changes.
    - Early cohorts demonstrate the success of past retention strategies worth revisiting.

Client Antiquity

- **Observation:**
    - Concentration around 4-6 years of tenure (~75.77% of clients).
    - High churn for clients in their first two years.
- **Insights:**
    - Early onboarding and engagement are critical for retention.

- Loyal, long-tenure clients represent stable revenue streams and can act as brand advocates.

Gas Service Subscription

- **Observation:**
  - Electricity-only clients dominate (81.85%).
  - Gas service subscribers have slightly lower churn rates (8.2% vs. 10.1%).
- **Insights:**
  - Cross-selling electricity-only clients presents significant opportunities.
  - Gas service may enhance customer stickiness, albeit marginally.

Forecasted Discount Energy

- **Observation:**
  - 96.49% of clients do not receive discounts.
  - Minimal discounts suggest selective pricing strategies.
- **Insights:**
  - Selective discounting could be optimized to target high-risk churn segments.
  - Experimenting with broader discount campaigns may improve retention metrics.

Active Products and Services

- **Observation:**
  - Single-product clients dominate (78.26%).
  - Clients with multiple products (21.74%) present upselling opportunities.
- **Insights:**
  - Promoting multi-product adoption can drive revenue growth.
  - Bundled offerings may enhance retention rates.

---

4 Hypothesis Testing

4.1 Objective

To determine whether price sensitivity, particularly changes in off-peak prices between December and January, significantly influences churn.

4.2 Methodology

- **Null Hypothesis (H0)**: There is no significant relationship between price and churn.
- **Alternative Hypothesis (H1)**: There is a significant relationship between price and churn.
- Bootstrap hypothesis testing was conducted on variables such as price_off_peak_var and price_off_peak_fix.

4.3 Findings

- P-values for all price variables were above 0.05, indicating a failure to reject the null hypothesis.
- Price sensitivity alone does not significantly predict churn.
- Other factors, such as consumption behavior and customer engagement, may play a more critical role.

---

5 Feature Engineering

1. **Handling Missing Values**:
   - Converted missing values to np.nan for consistency.
   - Missing values handled via imputation or removal, depending on the feature's importance.
2. **Processing Dates**:
   - Converted date features to datetime objects, enabling time-based calculations.
   - Extracted components like tenure (in years) to assess customer loyalty.
3. **Price Features**:
   - Created total price features (price_period1, price_period2) to capture comprehensive cost insights.
   - Developed price difference features (price_diff12, price_diff23) to measure volatility.
4. **Consumption Patterns**:
   - Derived deviations in last month's consumption from the 12-month average to identify irregular patterns.
   - Calculated the ratio of forecasted to actual consumption to understand usage trends.
5. **Categorical Encoding**:

- Label encoded variables like channel_sales and has_gas for machine learning compatibility.

---

6 Model Building and Evaluation

6.1 Data Splitting

- Split into training (80%) and testing (20%) sets, ensuring class distribution was representative.

6.2 Model Performance

1. Random Forest

- **Performance**:
  - Precision (Class 1): 0.97
  - Recall (Class 1): 0.97
  - F1-Score (Class 1): 0.97
  - Overall accuracy: 99%
- **Insights**:
  - The model effectively balances precision and recall, making it ideal for detecting churn.

2. Logistic Regression

- **Performance**:
  - Precision (Class 1): 0.25
  - Recall (Class 1): 0.01
  - F1-Score (Class 1): 0.01
  - Overall accuracy: 90%
- **Insights**:
  - Struggles with minority class prediction due to class imbalance.

3. XGBoost

- **Performance**:
  - Precision (Class 1): 0.99
  - Recall (Class 1): 0.09
  - F1-Score (Class 1): 0.16

- Overall accuracy: 91%

- **Insights**:
  - **Retention (Class 0):** All models are effective at identifying retained customers.
  - **Churn (Class 1):** Random Forest emerges as the best-performing model for detecting churn, as it balances precision and recall effectively.
  - **Challenges with Imbalance:** Logistic Regression and XGBoost struggle with the minority class due to the dataset's imbalance.
  - **Suggested Improvements:**
    - Employ **balancing techniques** such as SMOTE (Synthetic Minority Over-sampling Technique) or class weighting to improve the models' performance on churn prediction.

6.3 Feature Importance

- Consumption patterns emerged as the strongest predictor of churn.
- Price-related features were less influential, suggesting limited price sensitivity among customers.

---

7 Insights and Recommendations

7.1 Key Findings

1. **Price Sensitivity**:
   - Changes in off-peak prices are not significant predictors of churn.
   - Other factors, such as consumption behavior and contract details, are more influential.
2. **Consumption Patterns**:
   - Deviations in usage trends strongly correlate with churn.
   - Customers exhibiting irregular or decreasing consumption patterns are at higher risk.
3. **Class Imbalance**:
   - A significant imbalance in churn vs. non-churn cases affects model performance.
   - Random Forest proved to be the most reliable model due to its ability to handle imbalanced data.

7.2 Actionable Recommendations

1. **Engage High-Risk Customers**:
   - Monitor consumption patterns to identify potential churn indicators.
   - Proactively reach out to at-risk customers with tailored offers or support.
2. **Enhance Customer Retention**:
   - Focus on early onboarding strategies to reduce churn in the first two years.
   - Leverage insights from high-retention cohorts (e.g., pre-2009 contracts) to inform strategies.
3. **Optimize Pricing and Campaigns**:
   - Experiment with selective discounts targeting high-risk segments.
   - Invest in successful campaigns (e.g., Campaign 4) while reevaluating underperforming ones.
4. **Improve Model Performance**:
   - Address class imbalance using techniques like SMOTE or cost-sensitive learning.
   - Explore additional qualitative factors, such as customer feedback and service quality metrics.

---

8 Conclusion

The analysis reveals that while price sensitivity has some impact, it is not the primary driver of customer churn for PowerCo. Instead, consumption patterns and customer tenure play a more critical role. By leveraging these insights and implementing targeted retention strategies, PowerCo can reduce churn and maintain its competitive edge in the energy market.

Future work should incorporate qualitative factors, address class imbalance, and refine predictive models to ensure sustained improvement in customer retention.