Predicting Churn Risk for PowerCO

Hilda Nderitu

MACHINE LEARNING I CAPSTONE PROJECT

15th February 2025

Introduction

- **PowerCo**
  - Major gas & electricity utility company that supplies to small & medium sized enterprises
- **BCG X**
  - Consulting company hired to advise **PowerCo** on how to retain their customers
  - Data from **PowerCo** will be analysed, model developed & used to predict churn risk

Project Overview

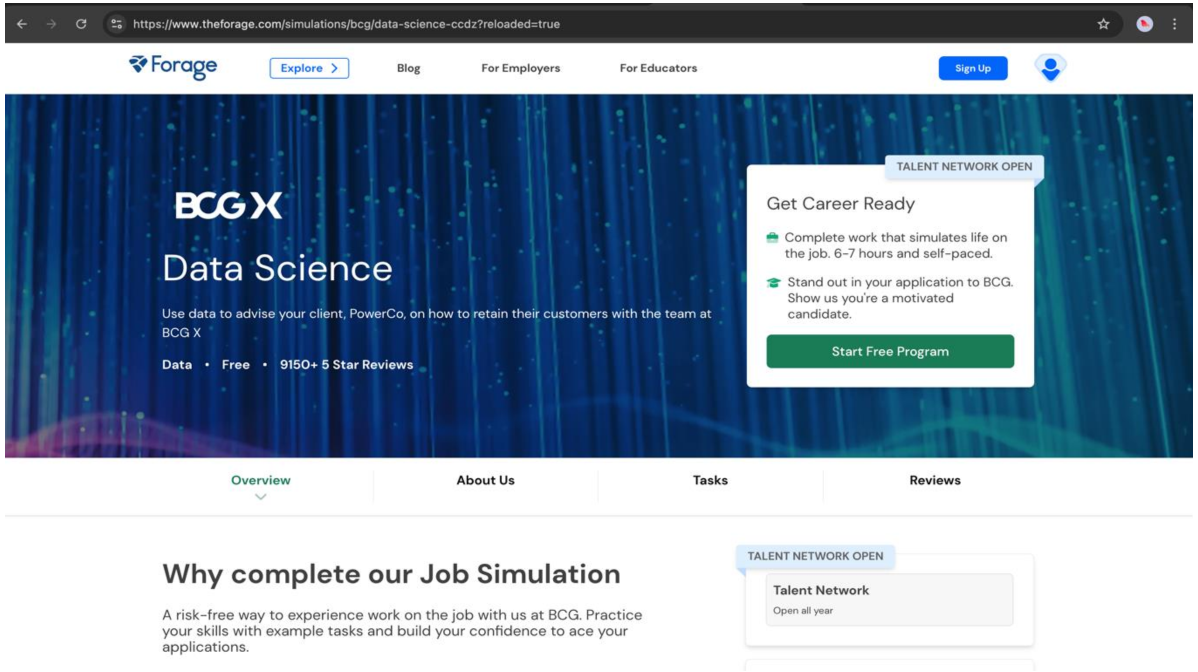
- Determine business problem
- Generate hypothesis
- Import datasets
- Conduct exploratory data analysis
- Conduct data pre-processing & feature engineering
- Develop & evaluate predictive model
- Generate insights & recommendations

Business Problem

- A lot of change in the energy market in recent years
- Increasing availability of more energy options than ever for customers to choose from
- PowerCo concerned about their customers leaving for better offers from other energy providers, which is now a big problem
- Thus, need to determine reasons for customers churning
- Hypothesis: churn is driven by customers' sensitivity to price

Dataset

- Dataset derived from a job simulation exercise provided by BCG X on the forage website https://www.theforage.com/simulations/bcg/data-science-ccdz
- 3 csv files provided



Dataset - ii

csv file	rows	columns	Data type	Use
client_data	14,606	• 25 features • 1 target variable	• 17 numerical columns • 8 categorical columns • Target variable – numerical column, boolean values	• Provided for EDA
price_data	193,002	• 8 features	• 2 categorical columns • 6 numerical columns	• Provided for EDA
clean_data_after_eda .csv (18 features added to initial 25, data on variation of prices yearly & 6 monthly)	14,606	• 43 features • 1 target variable	• 35 numerical columns • 8 categorical columns • Target variable – numerical column, boolean values	• Provided for feature engineering & data pre-processing

Datasets - iii

```
client_df.info()
✓ 0.0s
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14606 entries, 0 to 14605
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              14606 non-null  object
 1   channel_sales                   14606 non-null  object
 2   cons_12m                        14606 non-null  int64
 3   cons_gas_12m                    14606 non-null  int64
 4   cons_last_month                 14606 non-null  int64
 5   date_activ                      14606 non-null  object
 6   date_end                        14606 non-null  object
 7   date_modif_prod                 14606 non-null  object
 8   date_renewal                    14606 non-null  object
 9   forecast_cons_12m               14606 non-null  float64
 10  forecast_cons_year              14606 non-null  int64
 11  forecast_discount_energy        14606 non-null  float64
 12  forecast_meter_rent_12m         14606 non-null  float64
 13  forecast_price_energy_off_peak  14606 non-null  float64
 14  forecast_price_energy_peak      14606 non-null  float64
 15  forecast_price_pow_off_peak     14606 non-null  float64
 16  has_gas                         14606 non-null  object
 17  imp_cons                        14606 non-null  float64
 18  margin_gross_pow_ele            14606 non-null  float64
 19  margin_net_pow_ele              14606 non-null  float64
...
 24  pow_max                         14606 non-null  float64
 25  churn                           14606 non-null  int64
dtypes: float64(11), int64(7), object(8)
memory usage: 2.9+ MB
```

```
price_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 193002 entries, 0 to 193001
Data columns (total 8 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   id                  193002 non-null  object
 1   price_date          193002 non-null  object
 2   price_off_peak_var  193002 non-null  float64
 3   price_peak_var      193002 non-null  float64
 4   price_mid_peak_var  193002 non-null  float64
 5   price_off_peak_fix  193002 non-null  float64
 6   price_peak_fix      193002 non-null  float64
 7   price_mid_peak_fix  193002 non-null  float64
dtypes: float64(6), object(2)
memory usage: 11.8+ MB
```

```
cleaned_data_df.info()
✓ 0.0s
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14606 entries, 0 to 14605
Data columns (total 44 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   id                              14606 non-null  object
 1   channel_sales                   14606 non-null  object
 2   cons_12m                        14606 non-null  int64
 3   cons_gas_12m                    14606 non-null  int64
 4   cons_last_month                 14606 non-null  int64
 5   date_activ                      14606 non-null  datetime64[ns]
 6   date_end                        14606 non-null  datetime64[ns]
 7   date_modif_prod                 14606 non-null  datetime64[ns]
 8   date_renewal                    14606 non-null  datetime64[ns]
 9   forecast_cons_12m               14606 non-null  float64
 10  forecast_cons_year              14606 non-null  int64
 11  forecast_discount_energy        14606 non-null  float64
 12  forecast_meter_rent_12m         14606 non-null  float64
 13  forecast_price_energy_off_peak  14606 non-null  float64
 14  forecast_price_energy_peak      14606 non-null  float64
 15  forecast_price_pow_off_peak     14606 non-null  float64
 16  has_gas                         14606 non-null  object
 17  imp_cons                        14606 non-null  float64
 18  margin_gross_pow_ele            14606 non-null  float64
 19  margin_net_pow_ele              14606 non-null  float64
...
 42  var_6m_price_mid_peak           14606 non-null  float64
 43  churn                           14606 non-null  int64
dtypes: datetime64[ns](4), float64(29), int64(7), object(4)
memory usage: 4.9+ MB
```

Exploratory Data Analysis

Pandas – data analysis

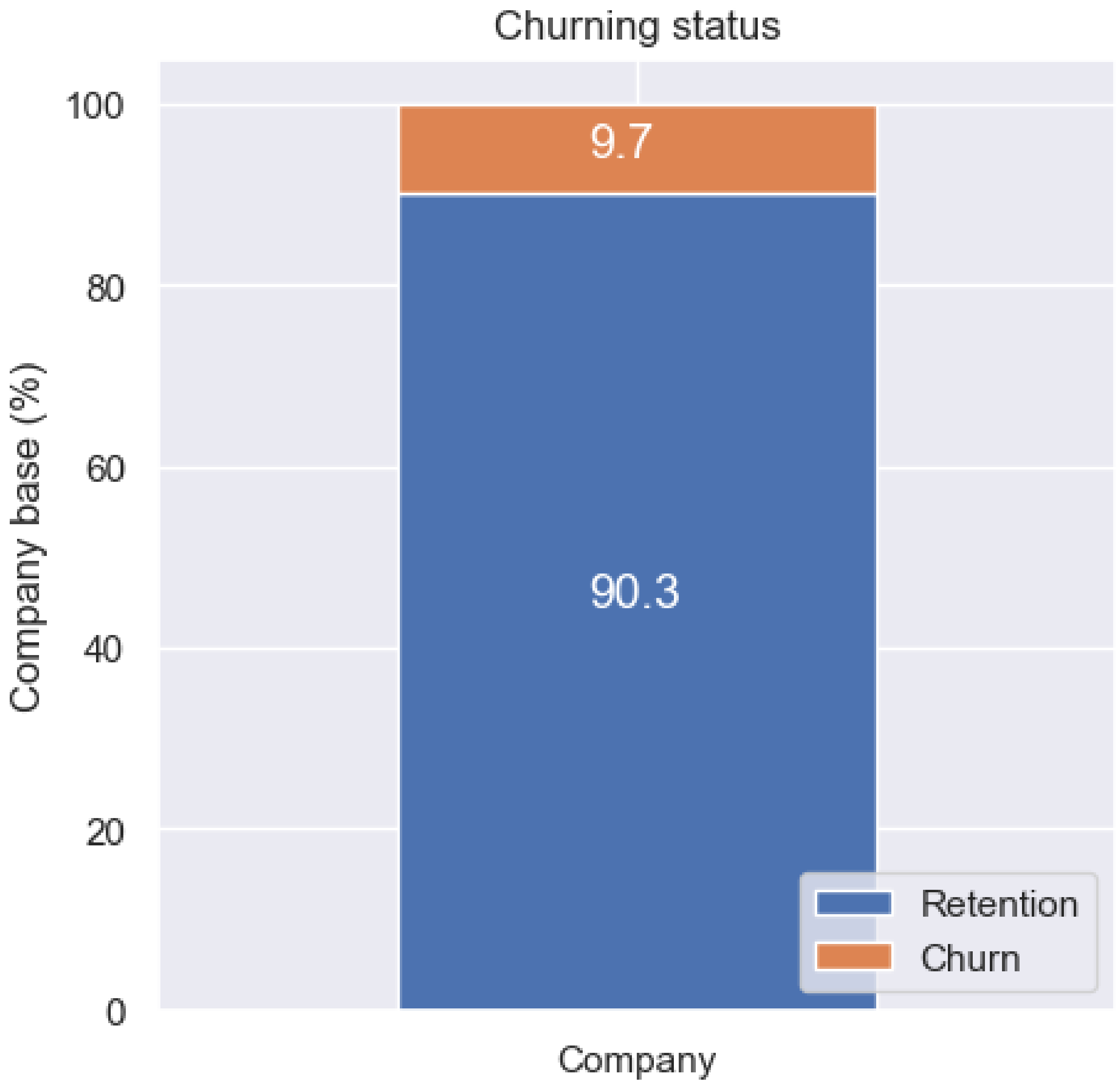
Seaborn & matplotlib - visualization

Techniques used:

- Summary statistics
- Visualization using:
  - Stacked bars
  - Distribution plots
  - Boxplots
  - KDE plots

- **Client data**
  - Columns with consumption, forecast, margins data - positively skewed, with outliers
  - Columns with data on subscribed power, number of active products & services, antiquity of client in years – has outliers
  - All features have no linear separability (non-linear)
- **Price data** - not skewed, no outliers

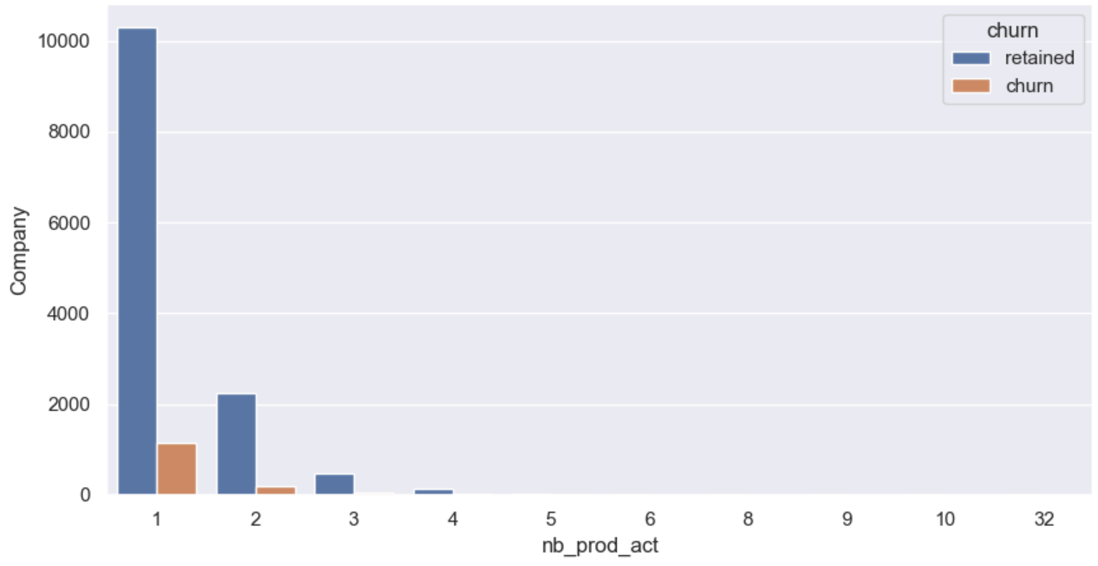
Exploratory Data Analysis - ii

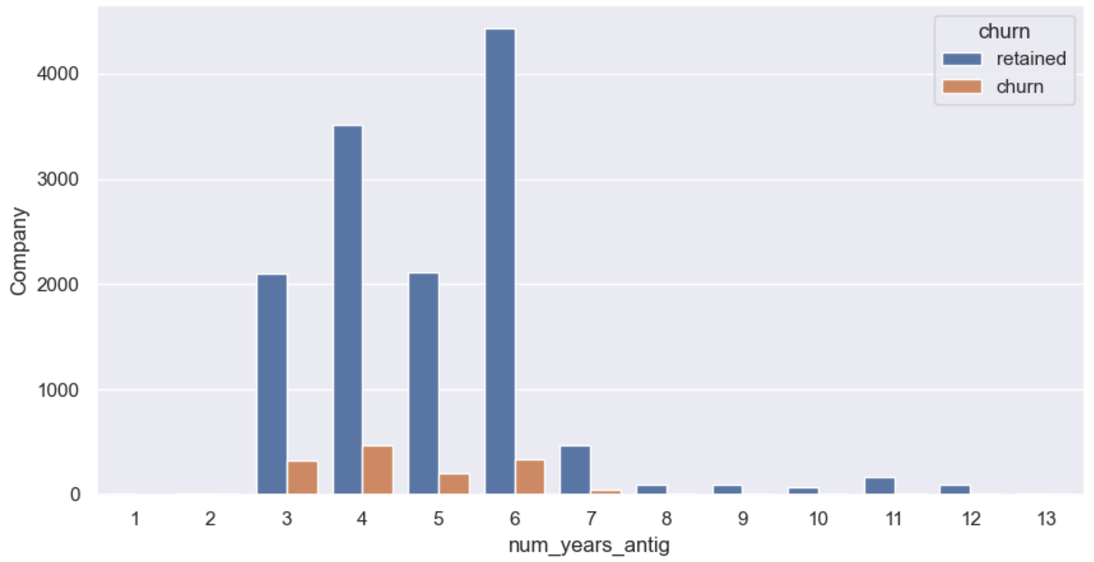


- 9.7% of customers have churned

```
        Company
churn
0         13187
1          1419
          Company
```

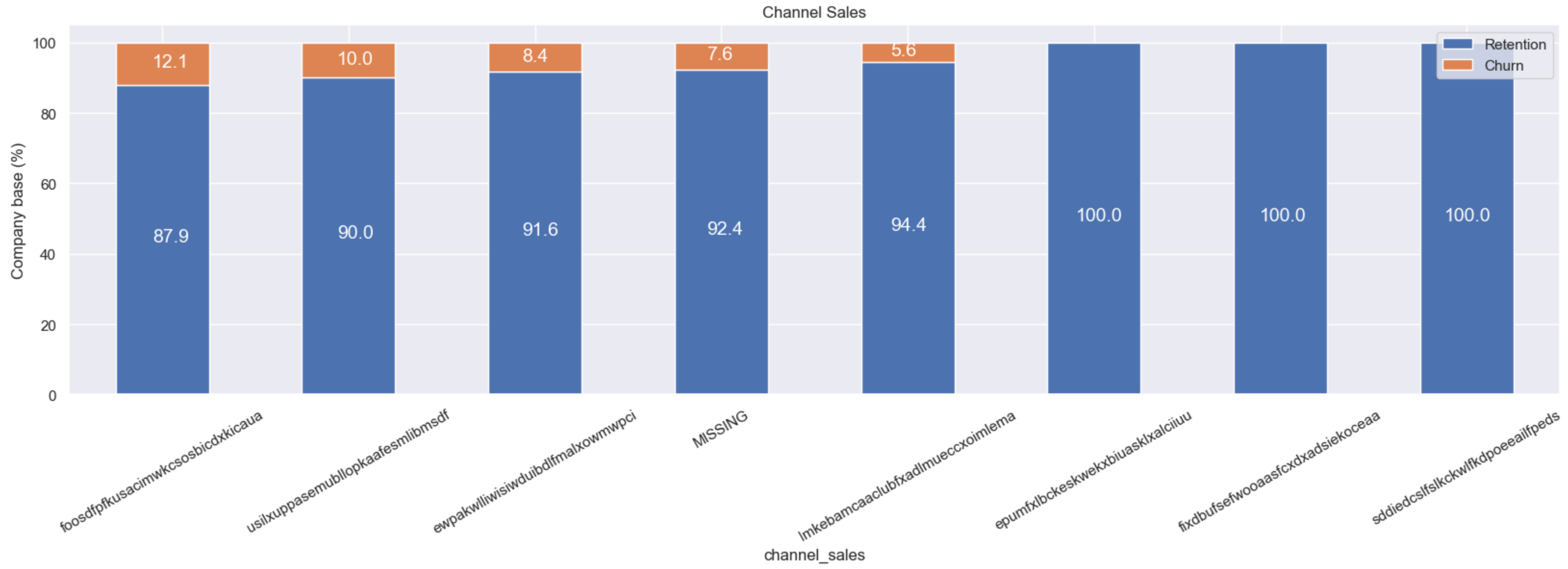
Exploratory Data Analysis - iii

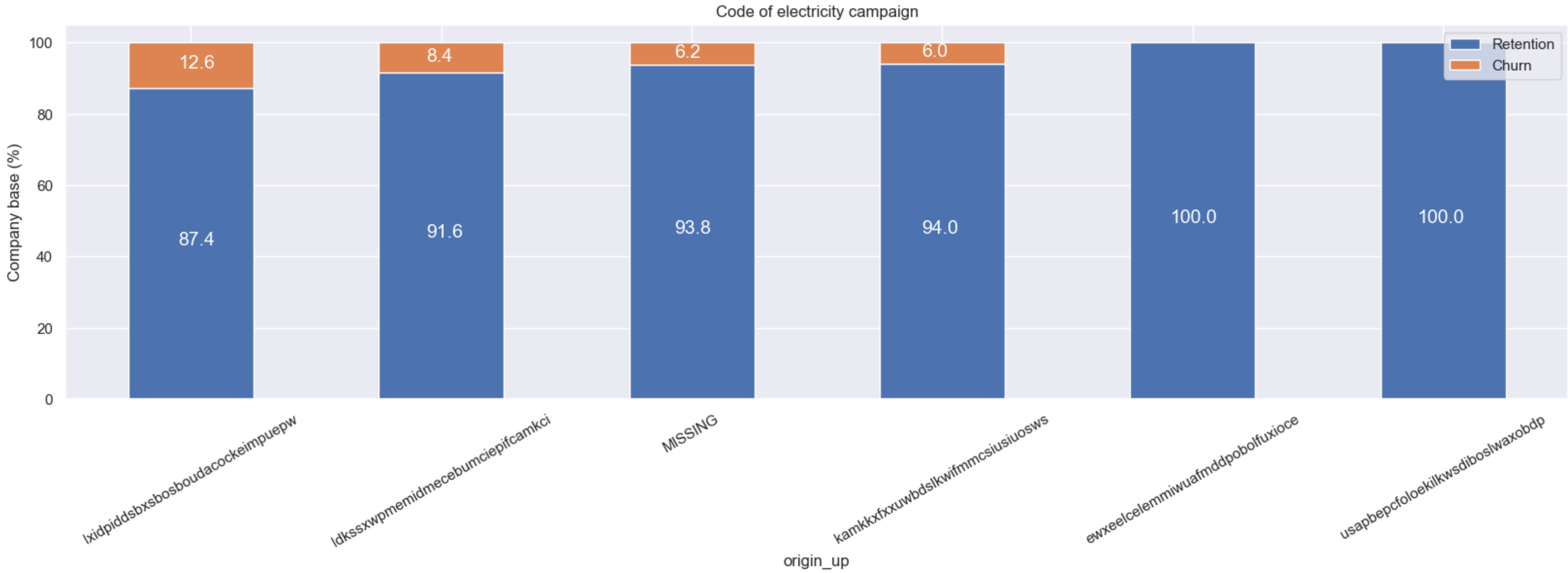




- Customers with 1 active product & services - highest churn rate
- Customers with > 2 active products & services - did not churn
- Customers with 4 years of antiquity – highest churn rate
- Customers with 8 - 12 years of antiquity - did not churn
- Customers with no gas service churned more than the ones with electricity & gas

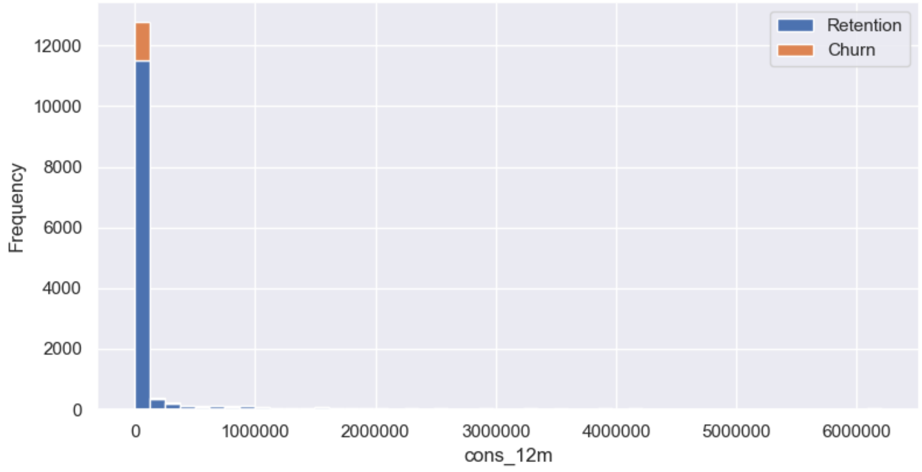
Exploratory Data Analysis - iv

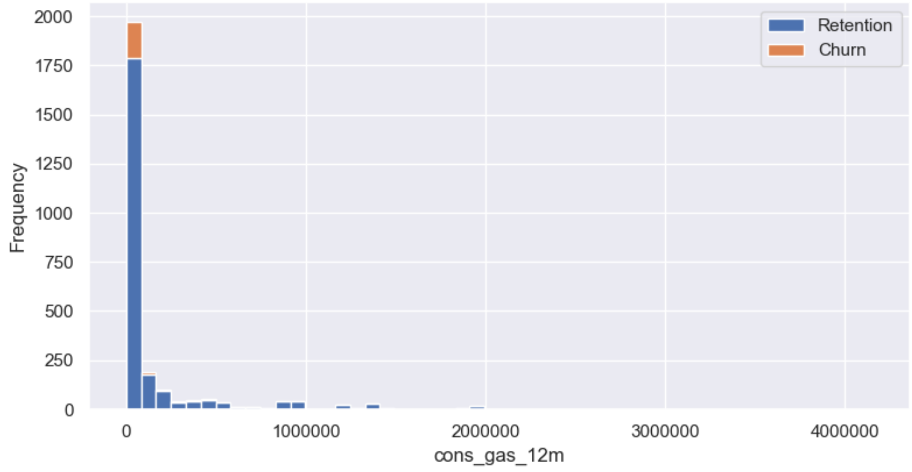


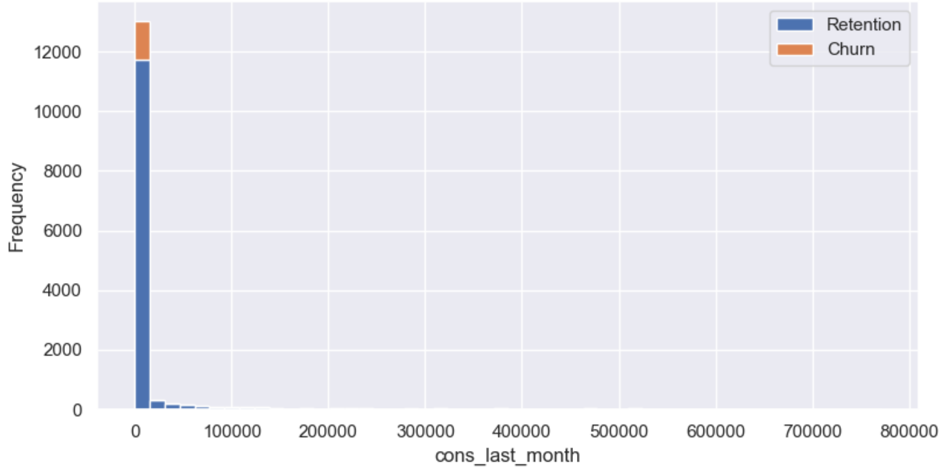


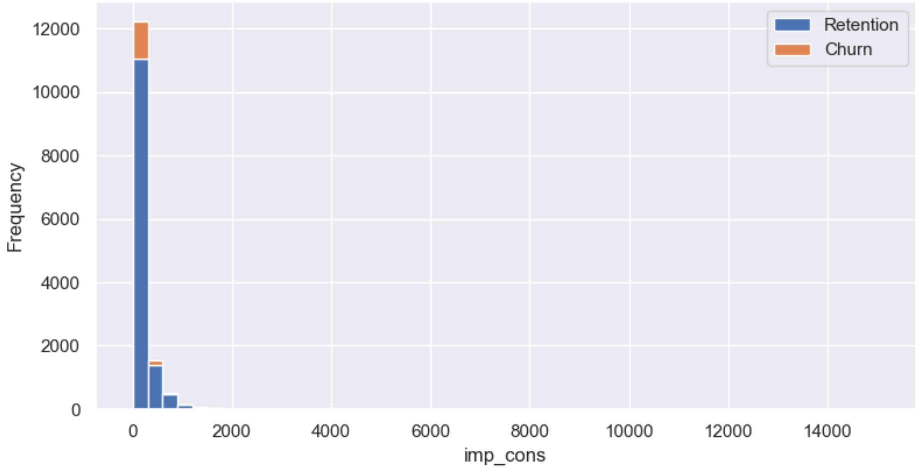
- Churning customers distributed over 5 different categories of channel_sales
- Churning customers distributed over 4 different categories of code of electricity campaign of first subscription

Exploratory Data Analysis - v



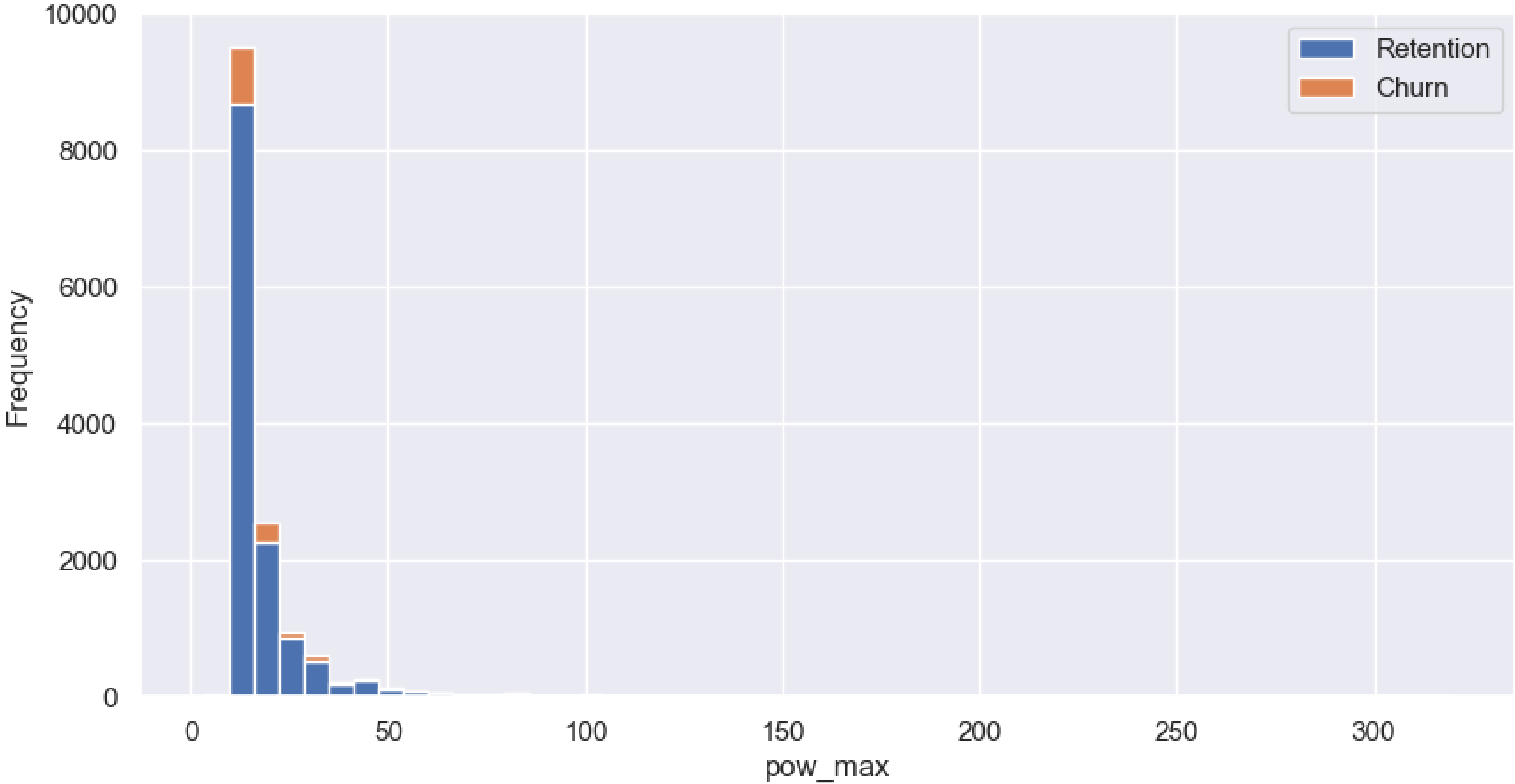






- The largest proportion of customers that have churned had lower consumptions
- With increase in consumption, there is decrease in churning
- At high consumption, there is no churning

Exploratory Data Analysis - vi



- Low values of subscribed power – highest proportion of churned customers

Data pre-processing & feature engineering

client data

- **25 features**

- date columns to datetime
- Addition of new features
- Merging data sets
- Deletion of features
- One hot encoding
- Log transformation - skewed data

Price data

8 features

Feature engineered data

- **59 features**

Robust scaling & deletion of features during model evaluation for improvement of metrics

Final data for final model

54 features

Data pre-processing & feature engineering - ii

- Creation of new features – calculation of differences, mean, max, conversion to months
- Calculation of correlation matrix
- Deletion of features – those with high correlation (9 features), no feature importance, unnecessary features i.e. id, datetime columns

Modelling & Evaluation

- Class imbalance – handled via SMOTE
- Data split into train & test tests: 75/25
- Scikit learn module used for modelling
- Random forest classifier model used as dataset has;
  - Many features & rows
  - Non-linear relationship – features & target variable
  - Can generate feature importances so as to know features predicting churn
- Base random forest model with default parameters

Modelling & Evaluation - ii

- Evaluation metrics used:
  - Accuracy
  - Precision
  - Recall
  - f1 score
  - PR-AUC(average precision score)
  - Confusion matrix
  - Correlation report

- Hyperparameter tuning - RandomizedsearchCV – indentify best parameters
- Threshold tuning – determine probability threshold to optimize precision/recall balance

Modelling & Evaluation - iii

- Best Parameters
  - 'n_estimators': 300
  - 'min_samples_split': 5
  - 'min_samples_leaf': 1
  - 'max_features': 'sqrt’
  - 'max_depth': None
  - 'criterion': 'entropy’

- Threshold: 0.5061

- Final model evaluation metrics:
  - Accuracy: 0.9510
    Precision: 0.9774
    Recall: 0.9241
  - F1-Score: 0.9500
  - PR-AUC: 0.9873

Insights

- Churn rate is high, **9.7%** (1419/14606 customers)
- Model can predict churn; however, churn is not driven by customers' sensitivity to prices
- Top 5 features predicting churn are:
  - Subscribed power
  - The last month's electricity consumption
  - Gross margin on power subscription
  - Forecasted bill of meter rental for the next 12 months
  - Yearly electricity consumption

Recommendations

- Check the past 12 months consumption, gross margins on power subscription, subscribed power & forecast meter rentals for the next 12 months for the following:
  - 5 sales channels that had customers who churned in the next 3 months
  - 4 codes of the electricity campaign that companies first subscribed to that had customers who churned in the next 3 months

Next Steps

- Explore other supervised classification models on the dataset

References/ Attributes

1. https://www.theforage.com/simulations/bcg/data-science-ccdz