Churn Optimization

PowerCo

Meeting agenda

Discuss

- Insights learned form the data
- First model draft
- Suggestions on churn strategy

“ In recent years, post-liberalization of the energy market in Europe, PowerCo has had a growing problem with increasing customer defections above industry average. The churn issue is most acute in the SME division and thus PowerCo want it to be the first priority

PowerCo has asked whether it is possible to predict the customers which are most likely to churn so that they can trial a range of pre-emptive actions.

There is a hypothesis that clients are switching to cheaper providers so the first action to be trialed will be to offer customers with high propensity of churning a 20% discount.

“

During the meeting BCG will propose the solution to the problem as well as will share the insights learned from the data provided by Power Co.

Churn model results as well as recommendation on churn strategy will be covered during the meeting.

Next project steps will be discussed at the end of the meeting.

Project Design and Data

Data

Data has been provided in 5 csv files

-ml_case_training_data.csv (16096 records)

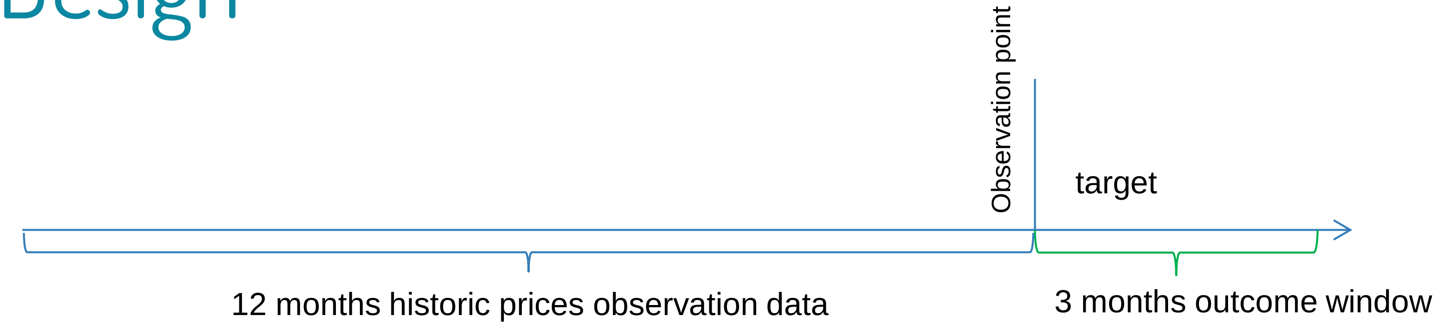
-ml_case_test_data (4024 records)

-ml_case_training_hist_data (193002 records)

-ml_case_test_hist_data (48236 records)

-churn target data

Design



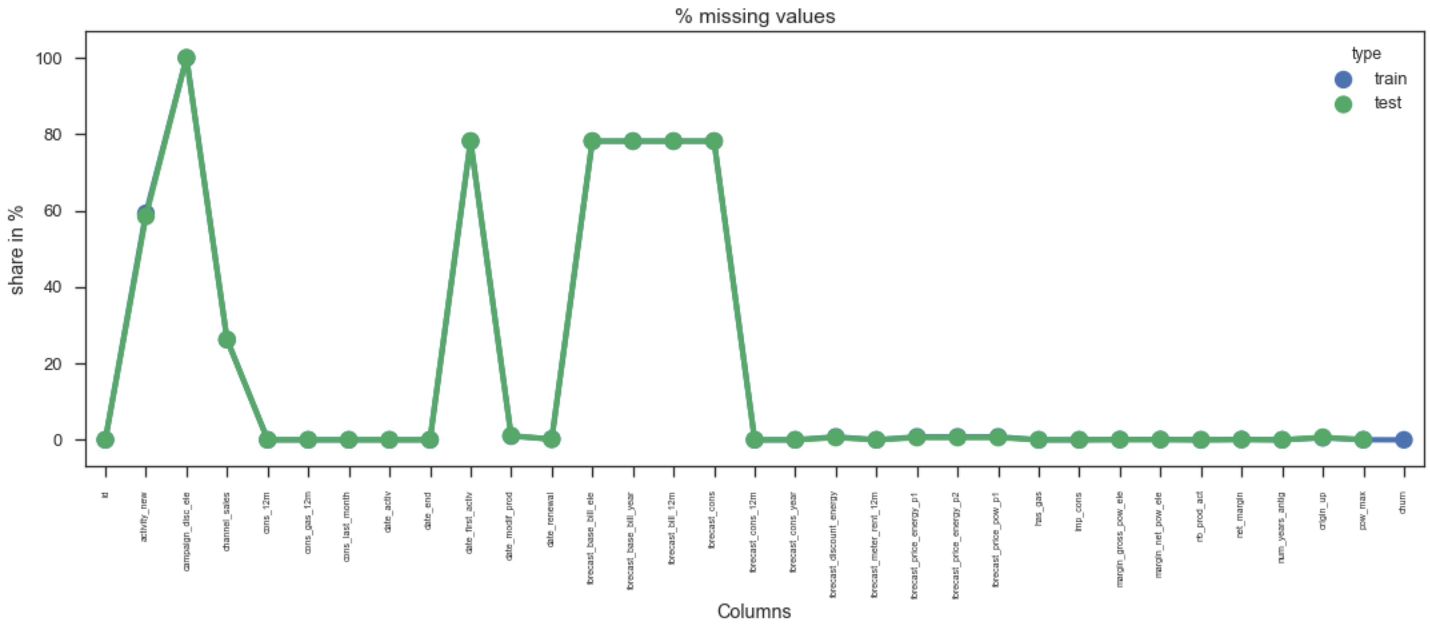
Observation point - Data for the clients provided at observation point (January 2016)
Outcome window - Churn events were captured at 3 months time window after observation point (from January 2016 to March 2016)
Prices data - was provided 12 months back from observation point

Exploratory Data Analysis

Exploratory Data Analysis

- Missing values counts
- Categorical variables exploration
- Continuous variables exploration
- Insights learned from the data

Missing values



Missing values

Column activity_new has 59.3004% missing values.
Column campaign_disc_ele has 100.0000% missing values.
Column channel_sales has 26.2053% missing values..
Column date_first_activ has 78.2058% missing values.
Column forecast_base_bill_ele has 78.2058% missing values.
Column forecast_base_bill_year has 78.2058% missing values.
Column forecast_bill_12m has 78.2058% missing values.
Column forecast_cons has 78.2058% missing values.

Column activity_new has 58.4990% missing values.
Column campaign_disc_ele has 100.0000% missing values.
Column channel_sales has 26.2425% missing values.
Column date_first_activ has 78.2058% missing values.
Column forecast_base_bill_ele has 78.2058% missing values.
Column forecast_base_bill_year has 78.2058% missing values.
Column forecast_bill_12m has 78.2058% missing values.
Column forecast_cons has 78.2058% missing values.

Missing values –Historic Price Data

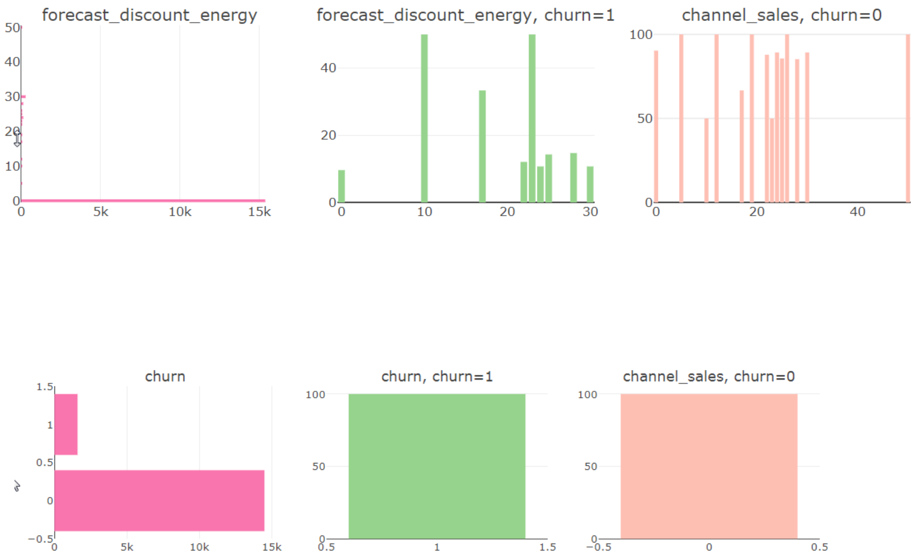
There is about 0.7% of missing values for all types of prices for train_hist file

There is about 0.6% of missing values for all types of prices for test_hist file

Data-Categorical variables by churn

- Distributions for channel sales and nb_prod_act are quite different for churned and non churned clients. Non churned clients have variable set of services. Some channels are not present for churned clients. Churned clients consume gas less

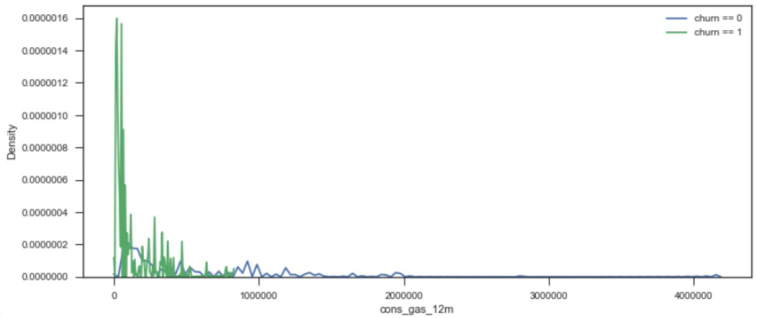
Data-Categorical variables by churn

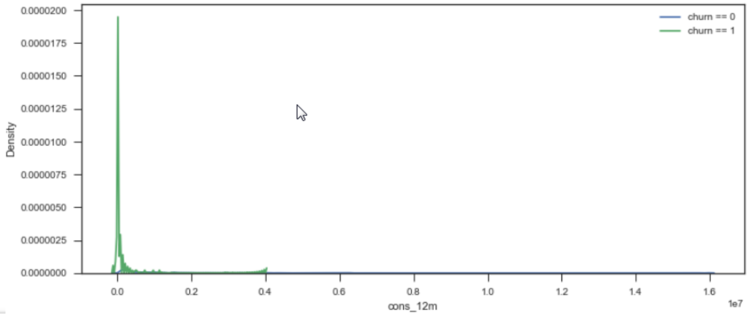


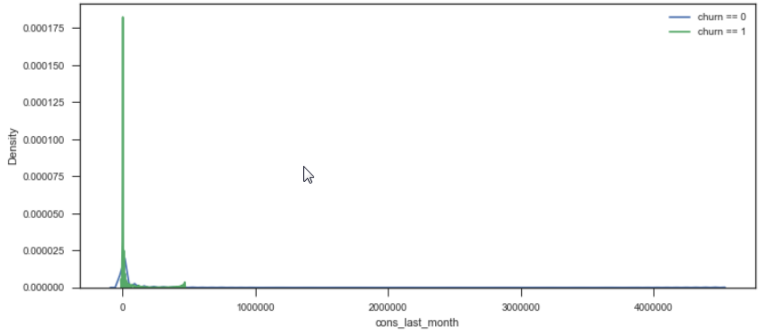
- More discount predicted options for churned customers

“Current churn rate is about 10%

Data-Continuous variables by churn- consumption

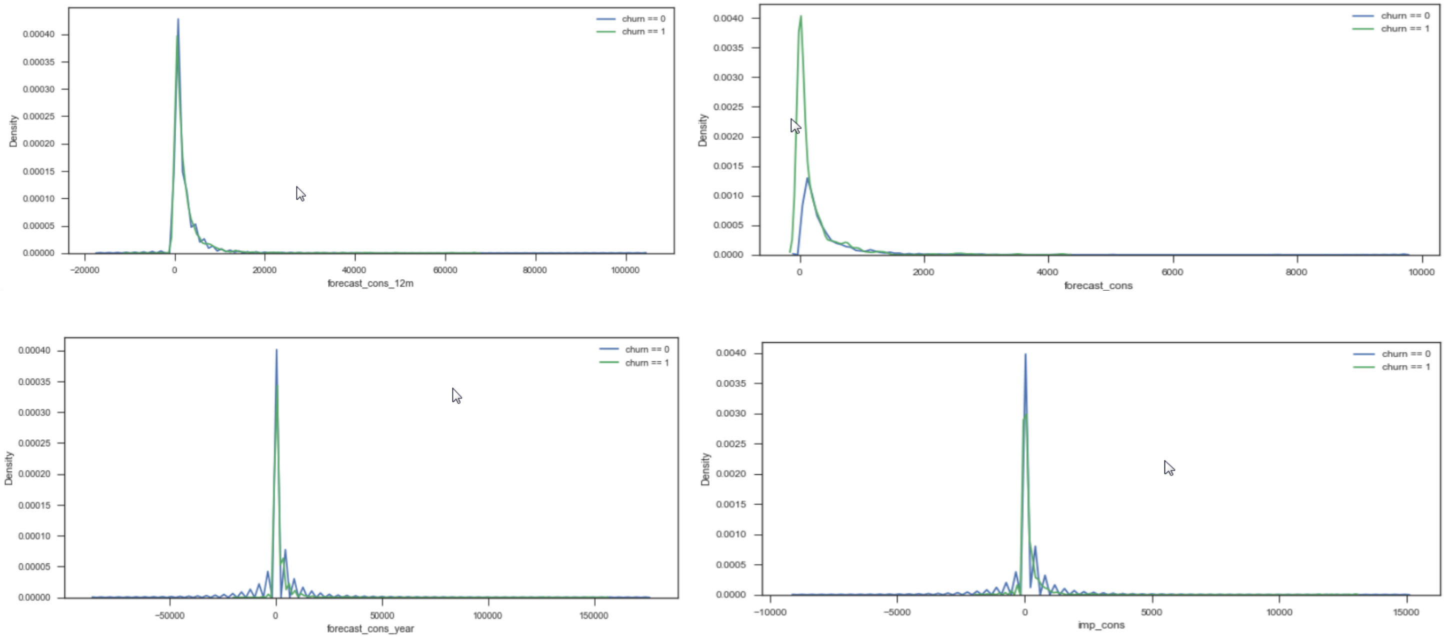






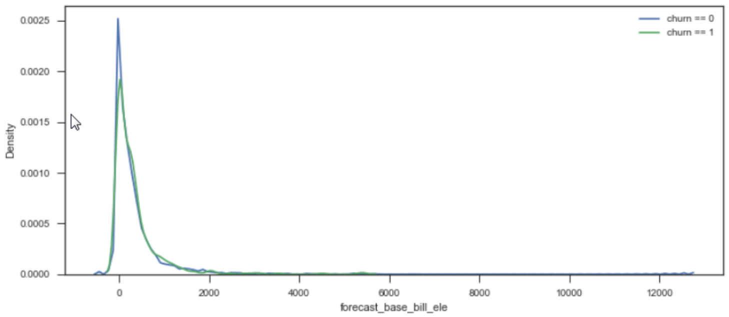
- Gas consumption (cons_gas_12_m) trend is quite different for churn customers (more peak values on the right tale)
- -A lot of values concentrated near 0

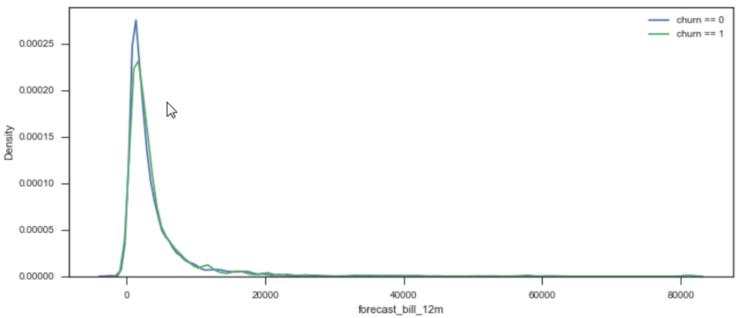
Data-Continuous variables by churn- forecast consumption

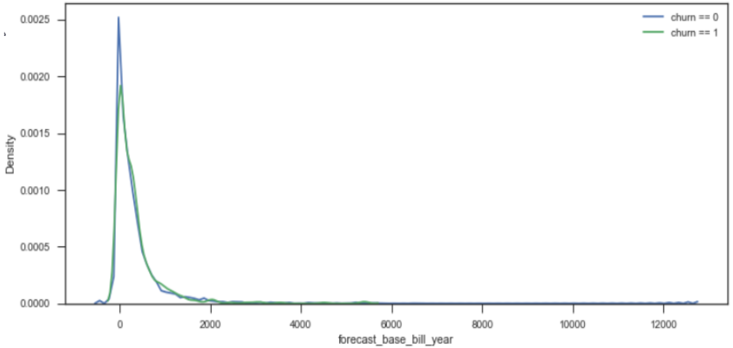


- Current paid consumption and forecasted consumption can take negative values
- A lot of ~0 values. Negative values can take place.

Data-Continuous variables by churn-forecast bill

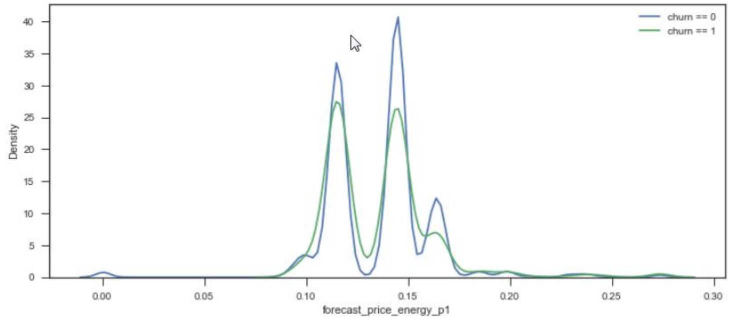


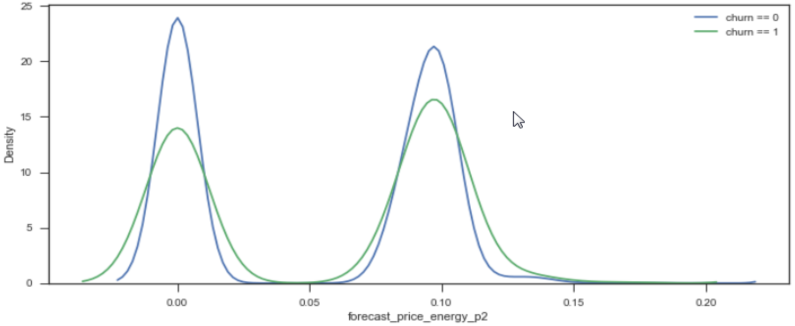


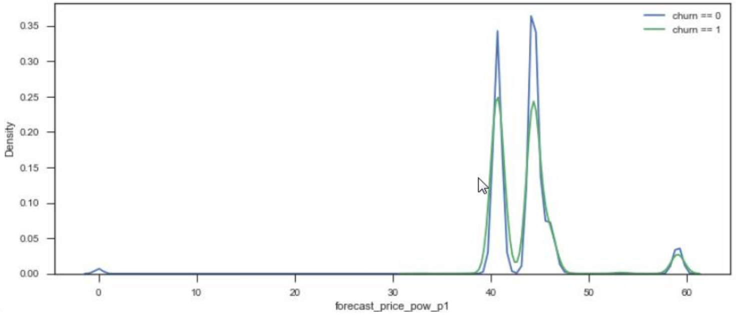


- A lot of forecasted values are equal to 0 or N/A .Negative values can take place.
- Only few number of clients have huge forecasted bills

Data-Continuous variables by churn- forecast price

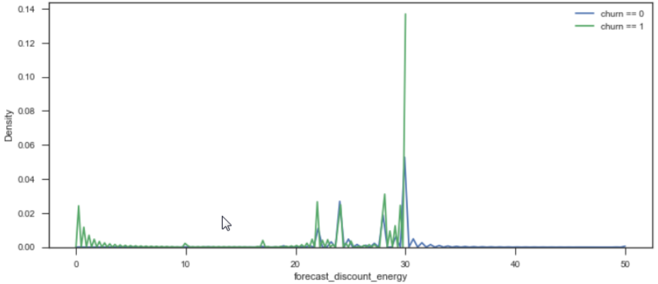


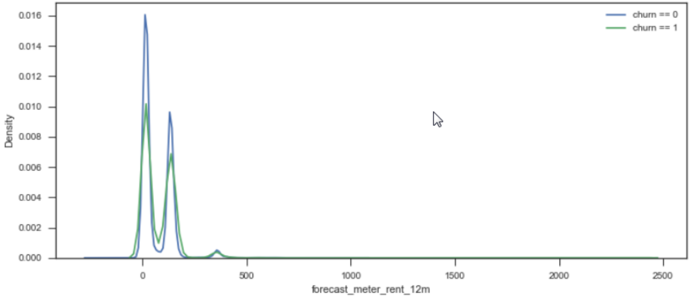


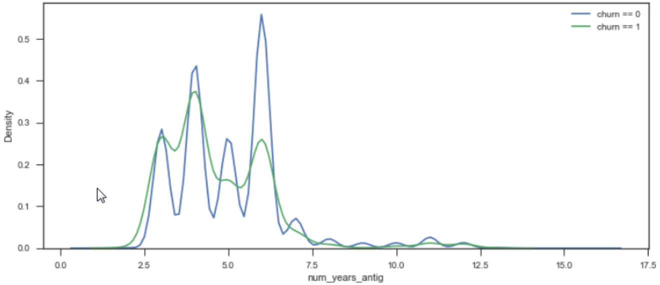


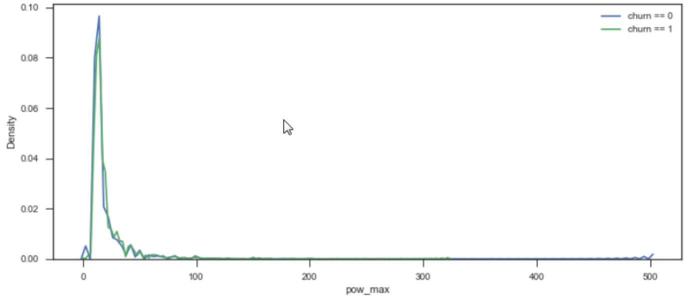
- There are several bell curves seen on all forecasted prices (with different parameters)
- Possibly this can be explained by various tariff plans. The same was observed analyzing historic pricing

Data-Continuous variables by churn



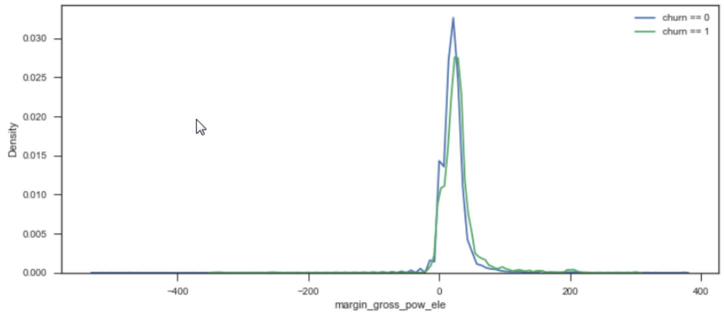


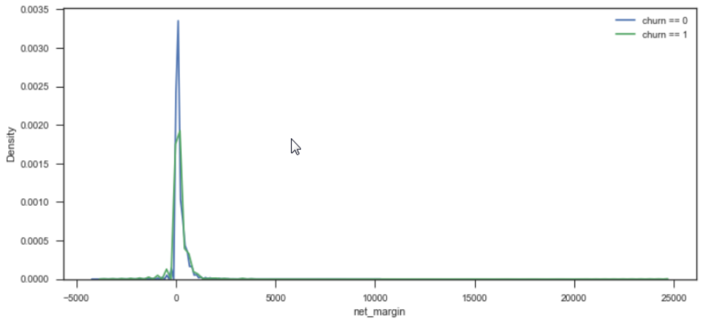


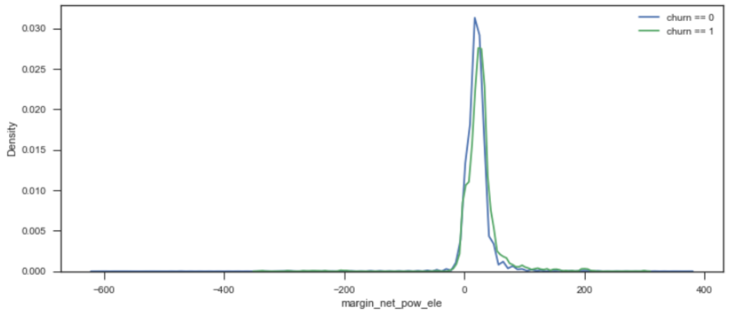


- There were peaks in time when clients started working with PowerCo
- By forecasted meter rent 2 main types of clients can be observed.

Data-Continuous variables by churn







- Very few clients generate huge margins. There is no dramatic distinction in terms of churn separation.
- However for margin type variables more clients generate huge net_margins

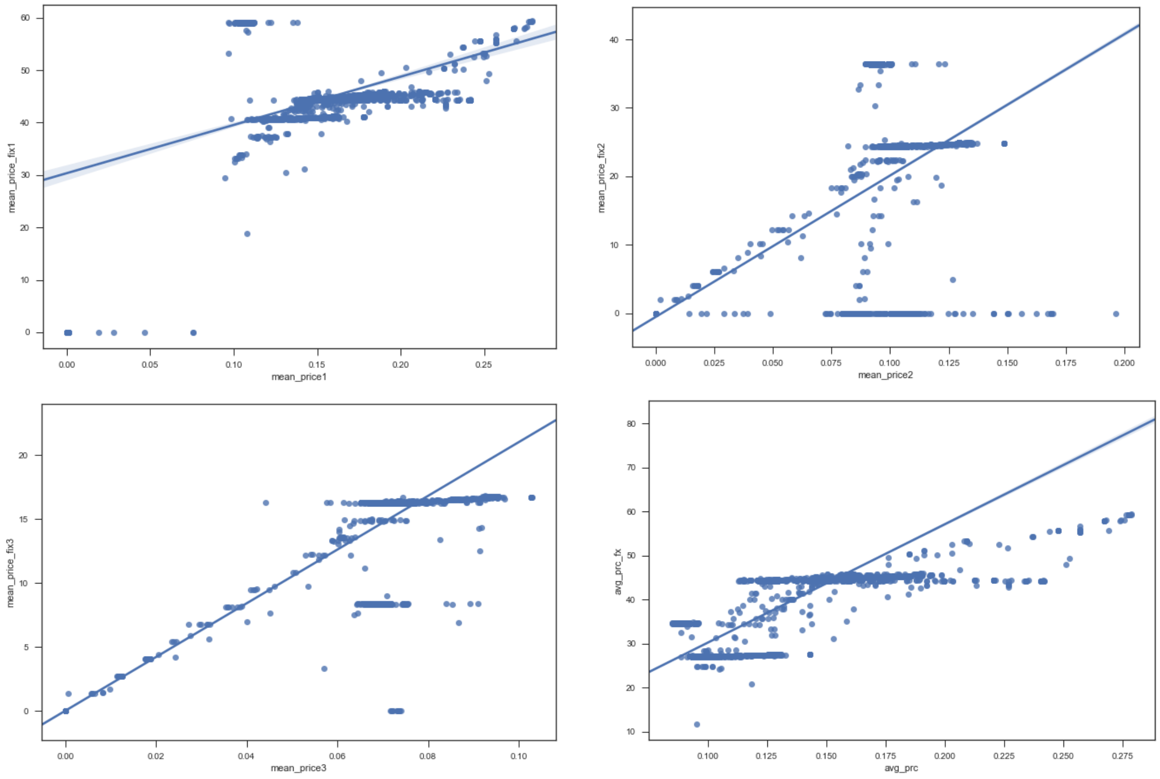
30% total company net_margin is done by 6% of the highest net_margin companies

50% total company net_margin is done by 14 % of the highest net_margin companies

80% total company net_margin is done by 37. % of the highest net_margin companies

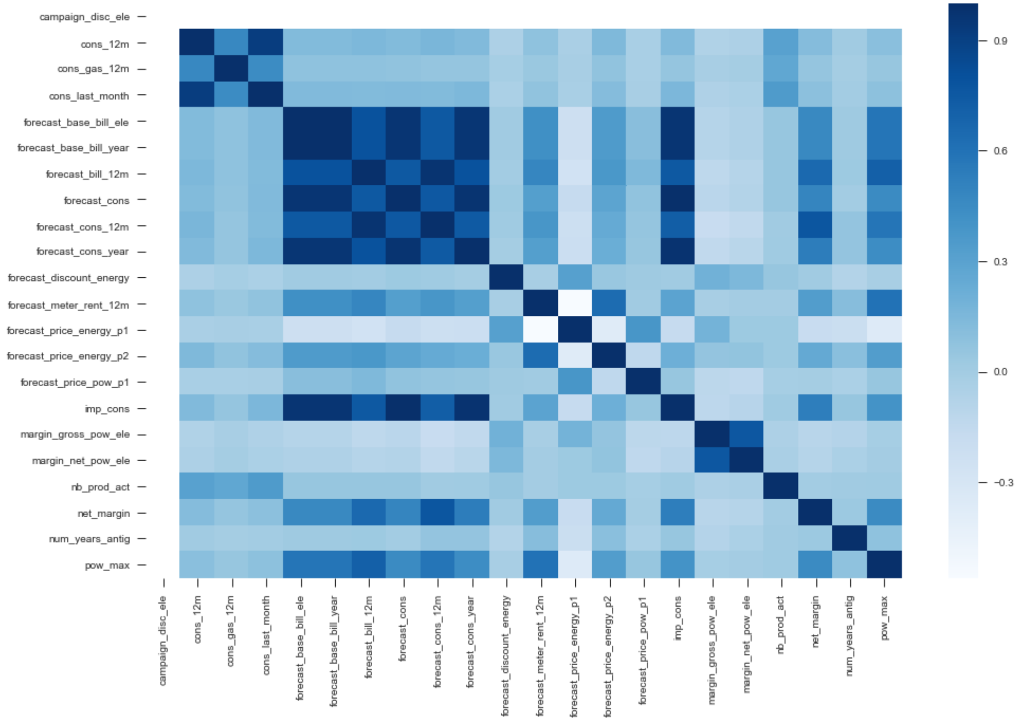
90% total company net_margin is done by 53 % of the highest net_margin companies

Data-Categorical - Prices data –Mean client yearly price_var to price_fix



- Respective price_var values are partialy linearly dependent on price_fix values

Correlation



- Consumption type and bill type variables are highly correlated among each other

Correlation with churn

Variable Name	Pearson Correlation with churn (positive)
origin_up	0.098807
margin_gross_pow_ele	0.080158
margin_net_pow_ele	0.063187
forecast_meter_rent_12m	0.029971
net_margin	0.029308
forecast_price_energy_p2	0.025597
forecast_discount_energy	0.012344
pow_max	0.009456
forecast_cons_12m	0.007395
forecast_bill_12m	0.006909
forecast_price_pow_p1	0.004034
imp_cons	0.003417
forecast_cons_year	0.002756

Variable Name	Pearson Correlation with churn (negative)
num_years_antig	-0.071565
cons_12m	-0.051759
cons_last_month	-0.046931
cons_gas_12m	-0.04088
channel_sales	-0.032198
has_gas	-0.032033
nb_prod_act	-0.023811
activity_new	-0.023541
forecast_cons	-0.005247
forecast_price_energy_p1	-0.003337
forecast_base_bill_year	0.000433
forecast_base_bill_ele	0.000433
forecast_cons_year	0.002756
imp_cons	0.003417
forecast_price_pow_p1	0.004034

Feature Generation

Sets of features

Features based on dates

- Days _since _activation
- Days_to_contract_end
- Days_since_first_contract
- Days_to_renewal
- Days_since_last_prod_mod

Features based on historic prices

-avg price

-mean price

-median price

-mean price

-max price

-etc.

Polynomial features

-Polynomial combinations of original features

^2,^3 from original features

-etc.

- Label encoding
- One hot encoding for categorical features

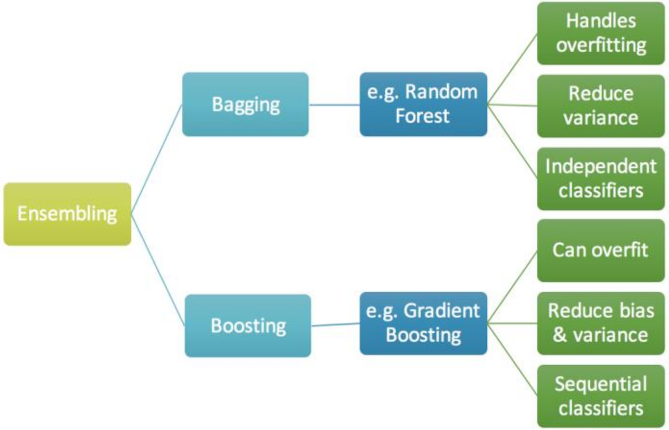
Model

Types of algorithms tested

3 algorithms has been tested

- Logistic Regression
- RANDOM FOREST
- LIGHT GBM

Cross validation on 5 stratified folds has been implemented



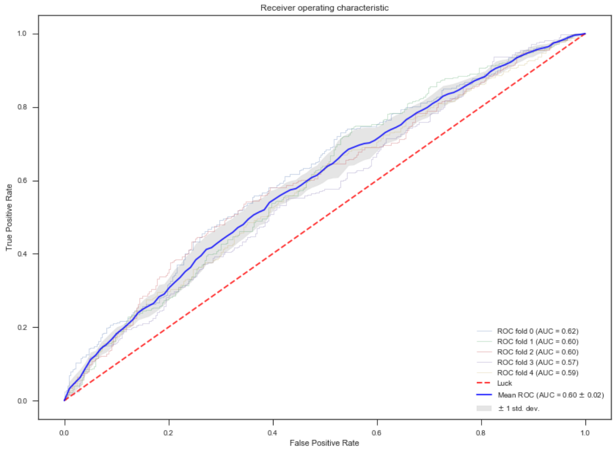
Model results

- Models have been trained on various sets of features

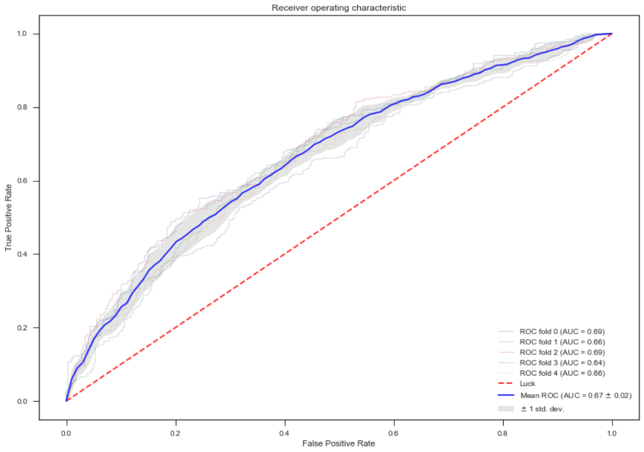
  (original set, with prices, with polynomial combinations)
- Polynomial subset of features has not shown significant uplift in the model performance
- Final set of features contained original set of features as well as historic price based historic variables (aggregations)
- Logistic regression has been taken as a baseline and ensemble models compared with it.
- As data is unbalanced the under sampling technique has been implemented. Model built on unbalanced data had problems with low precision and recall. Ubdersampling has improved the situation and model generalization has been tested on unbalanced data after training.

Model results- Logistic regression

Logistic regression



LightGBM

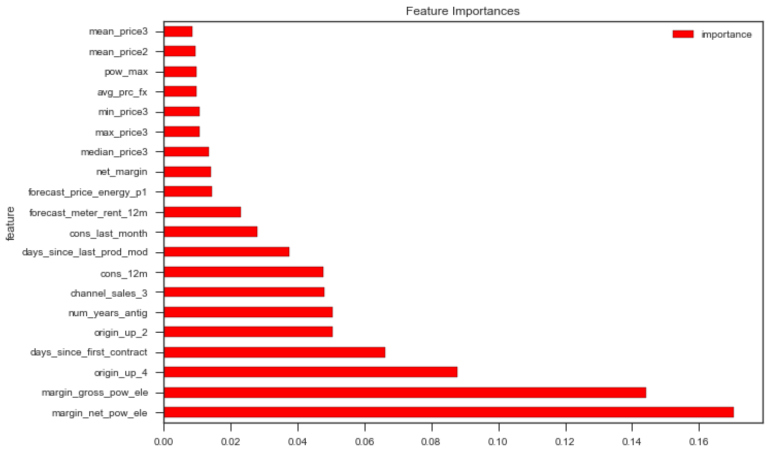


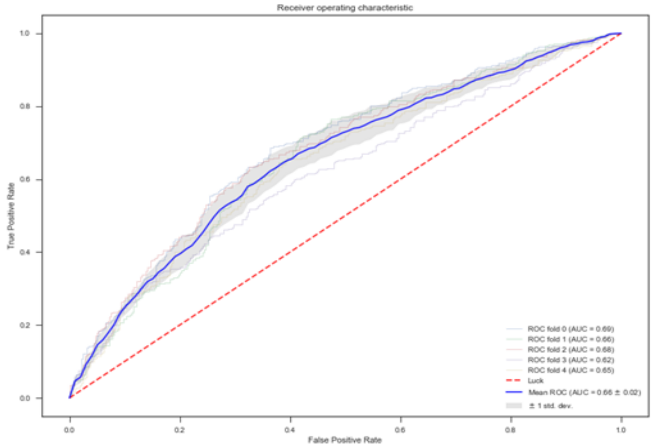
Logistic regression has been taken as a baseline model and shows the lowest performance (standard parameters)

However the model validated quite well on K-fold cross validatoin

LightGBM required parameters tuning as was overfitting heavily with standard parameters. For instance num_leaves hyper parameter has improved heavy overfitting problem.

Final model draft result





Random Forest showed the best result in terms of performance and generalization. The set of the most powerful variables can be seen on the graph. Variables are quite interpretable

However some variables should be discussed during the meeting.

Number of price based variables enter the model. However these variables are not top predictors.

Pricing Strategy

Approach suggested is to check the optimal price discount at which client will not churn based on the built model starting from very small discount values (in case churn model contains historic pricing components)

Provided data contains var and fix prices based on which various discount options can be simulated(e.g. 1,2,3,4,5,...20% discounts)

From the other hand optimization should consider margin which customer brings to the company. Hence optimization should consider the revenue that customer brings to a company. This means that 20% discount for all the clients may not be the optimal strategy. Each client can get individual discount taking margin prioritization into consideration.

THANKS!