Customer Churn Analysis and Prediction

Stacey (Shiqing) Wang

1

Introduction

Information about the background of the problem and business objectives

2

Data Preparation

Explanatory data analysis, data cleaning and feature selection

3

Modeling and Evaluation

Algorithm selection and comparison, evaluation of the selected models

4

Conclusions and Recommendations

Findings of the study, interpretation of results and recommendations for next step

01

Introduction

Introduce the background of the problem, and the business objectives of this study

Introduction

Background

- AtCo has a growing problem with increasing customer defections above industry average
- The churn issue is most acute in the small business division

Objectives of the study

- Predict which customers are most likely to churn
- Identify the drivers of the churn problem
- Whether 20% discount will keep customers from churning
- Explore the correlation between subscribed power and consumption
- Examine the link between sales channel and churn

Data Preparation

Explanatory data analysis, data cleaning and feature selection

Data Preparation

Missing data imputation

- Replace numerical missing data with mean
- Assign categorical missing data a new level
- Drop data that is 100% missing

Data cleaning

- Remove outliers
- Remove low probability (<5%) observations
- Convert datetime data to numerical data

Explanatory data analysis (EDA)

- Visualize overall churn rate
- Explore churn VS. price, churn VS sales channel, churn VS. activity, etc.
- Examine collinearity between variables

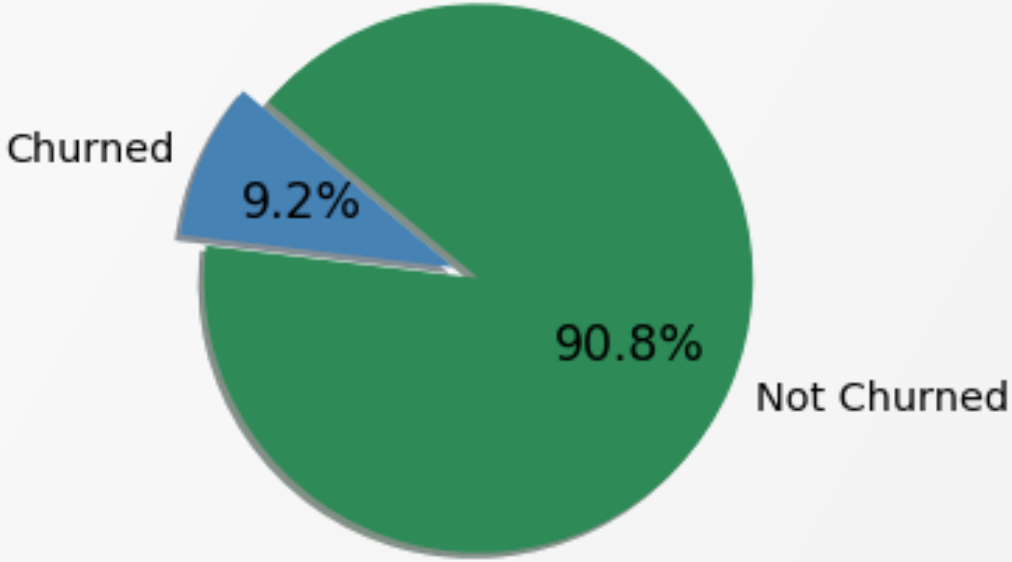
Model input data

- Encode categorical variables
- Remove redundant (highly correlated) variables

Explanatory Data Analysis

Overall Churn Rate

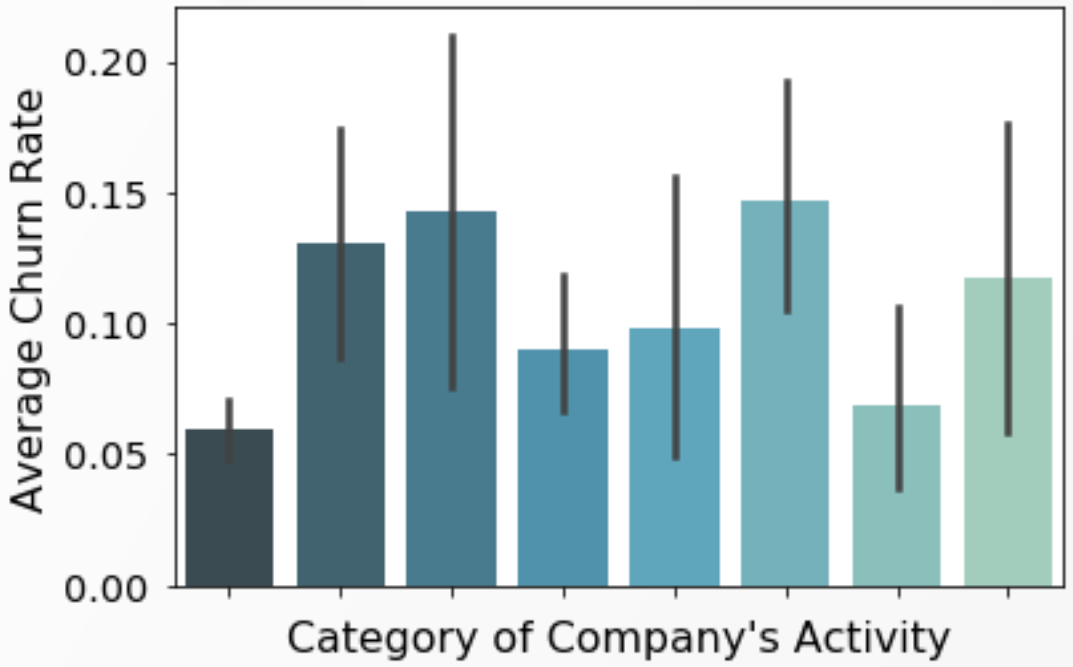
- According to the training data, about 9% of the customers have churned

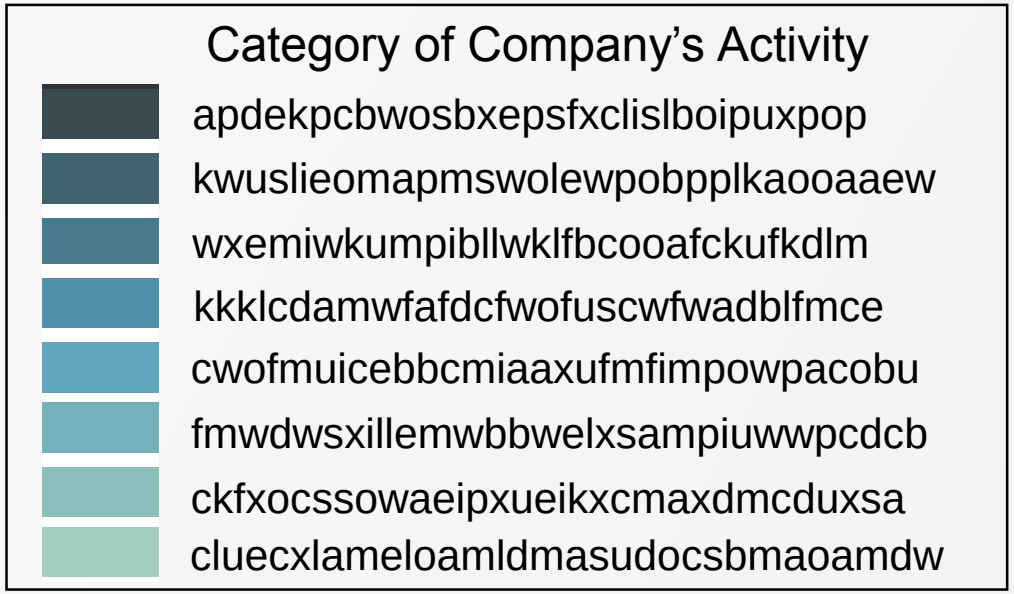


Explanatory Data Analysis

Category of Company's Activity VS. Churn Rate

- Companies with certain categories of activity show lower churn rate
- However, some categories have very low observations (<5%), and need to be excluded from the study
- 60% of the data is missing



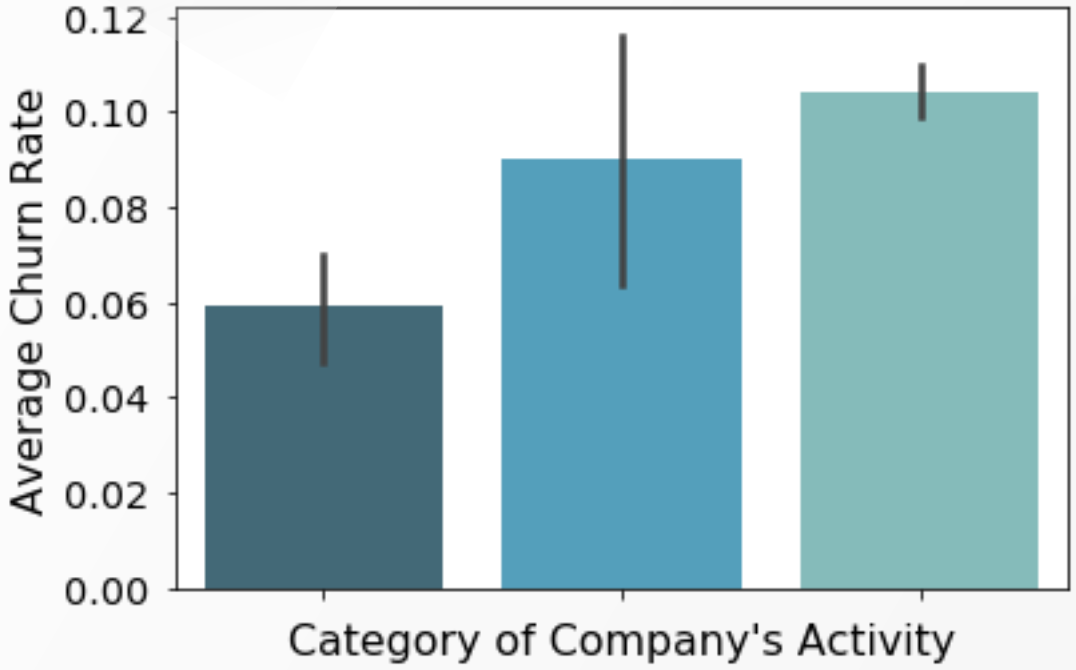


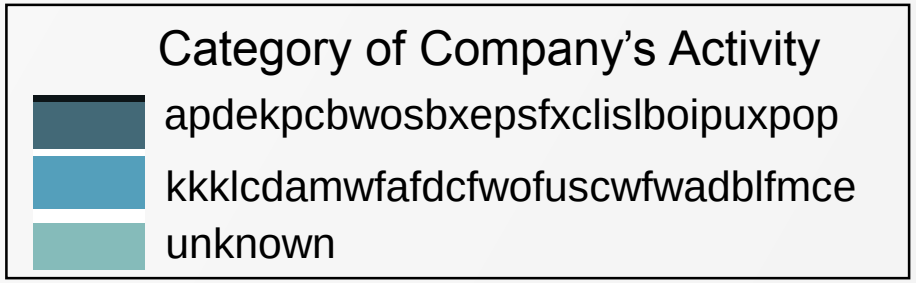
* Category of company's activity has been encrypted

Explanatory Data Analysis

Category of Company's Activity VS. Churn Rate

- Assign missing values and categories with low observations a new level: unknown
- The following bar plot shows the most common 2 categories of activity and all the rest are merged into category of "unknown"



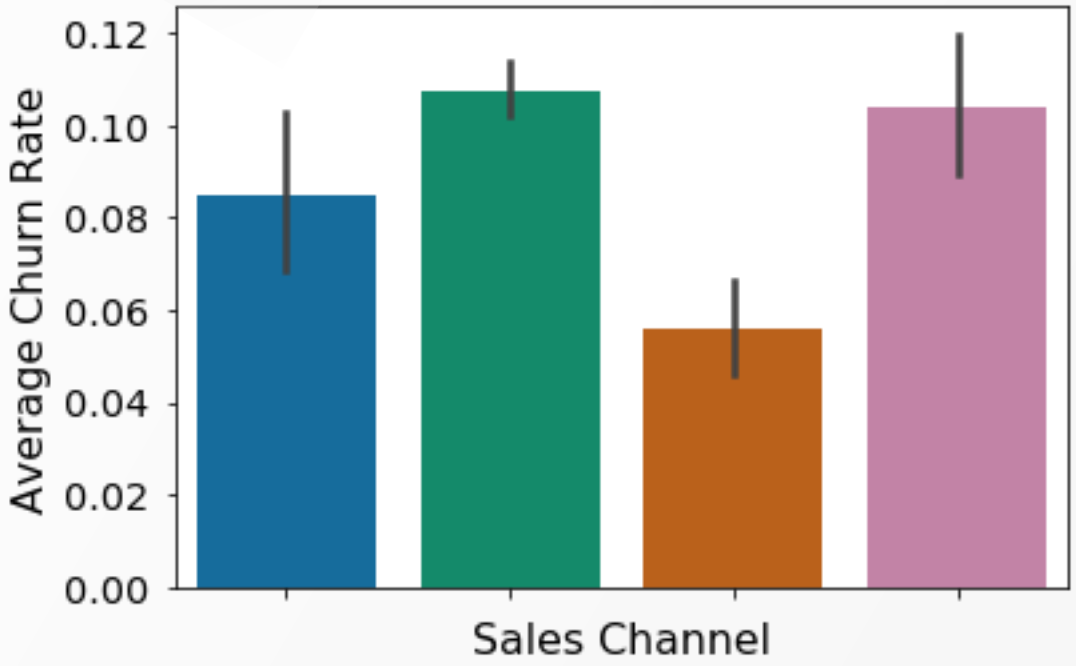


* Category of company's activity has been encrypted

Explanatory Data Analysis

Sales Channel VS. Churn Rate

- Customers from the 3rd sales channel (shortest bar below) exhibit lower churn rate



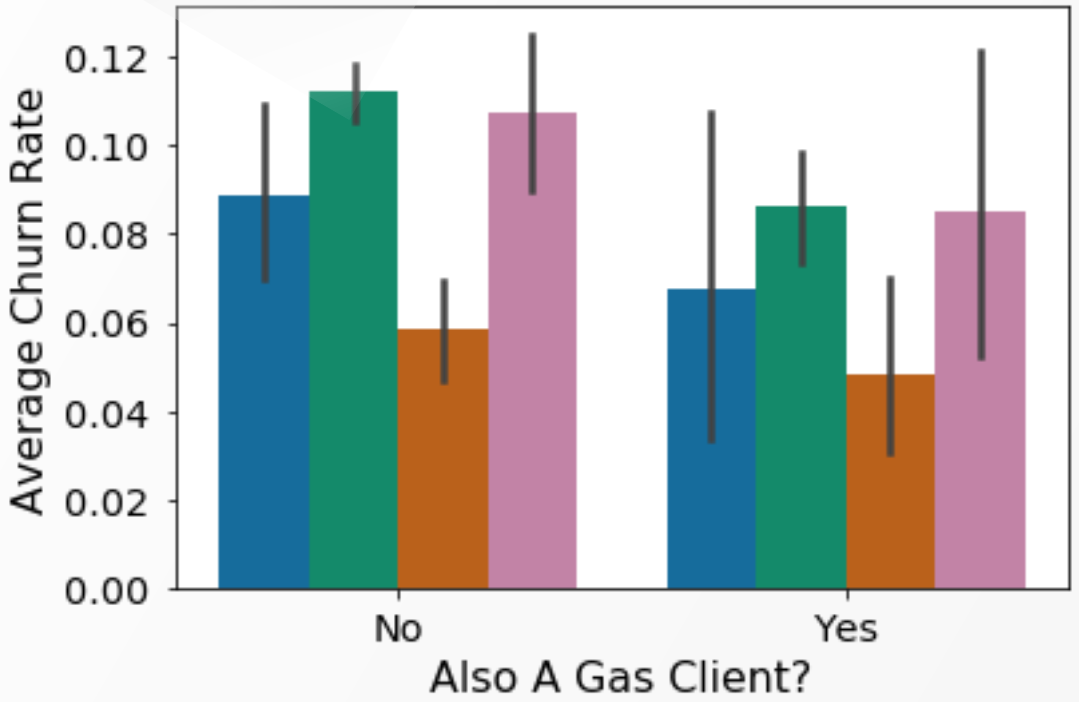
Sales Channel
- ewpakwlliwisiwduibdlfmalxowmwpci
- foosdfpfkusacimwkcsosbicdxkicaua
- imkebamcaaclubfxadlmueccxoimlema
- usilxuppasemubllopkaafesmlibmsdf

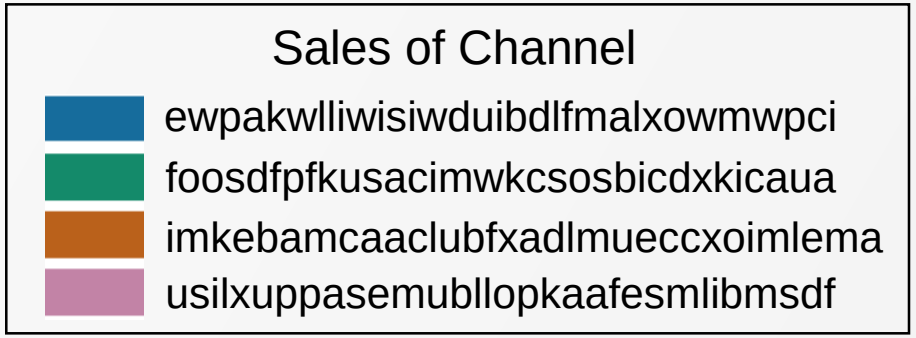
* Sales Channel has been encrypted

Explanatory Data Analysis

Customer with Multiple Services VS. Churn Rate

- Generally, customers who also have gas service are less likely to churn



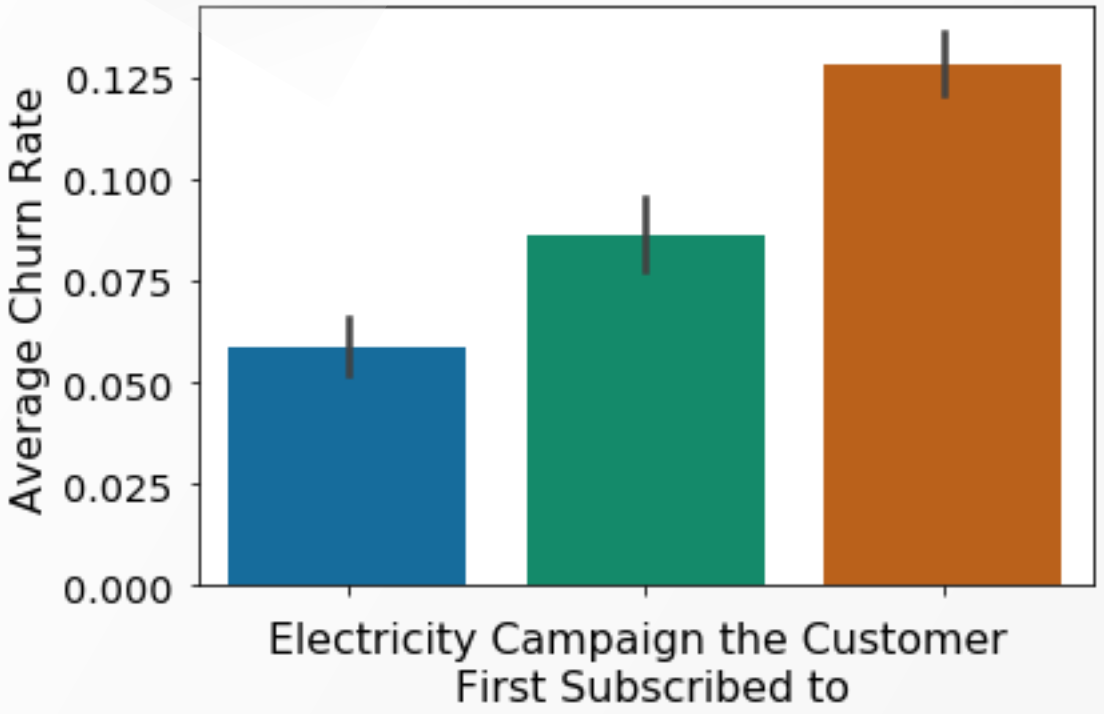


* Sales Channel has been encrypted

Explanatory Data Analysis

Electricity Campaign the Customer First Subscribed to VS. Churn Rate

- Customers subscribed to specific electricity campaign (as indicated by the shortest bar) show the lowest churn rate

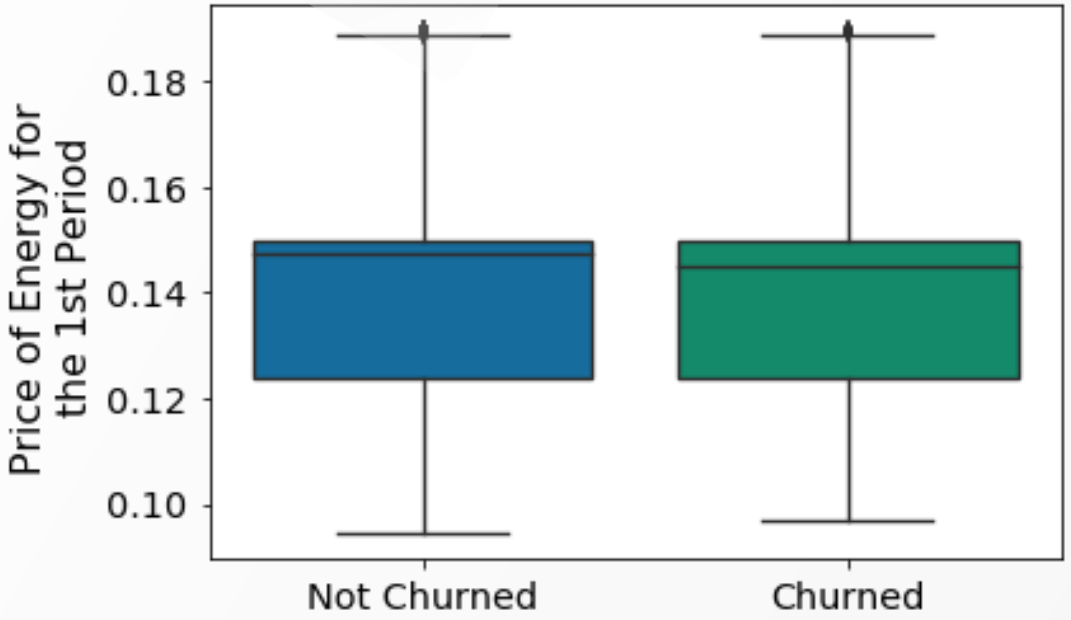


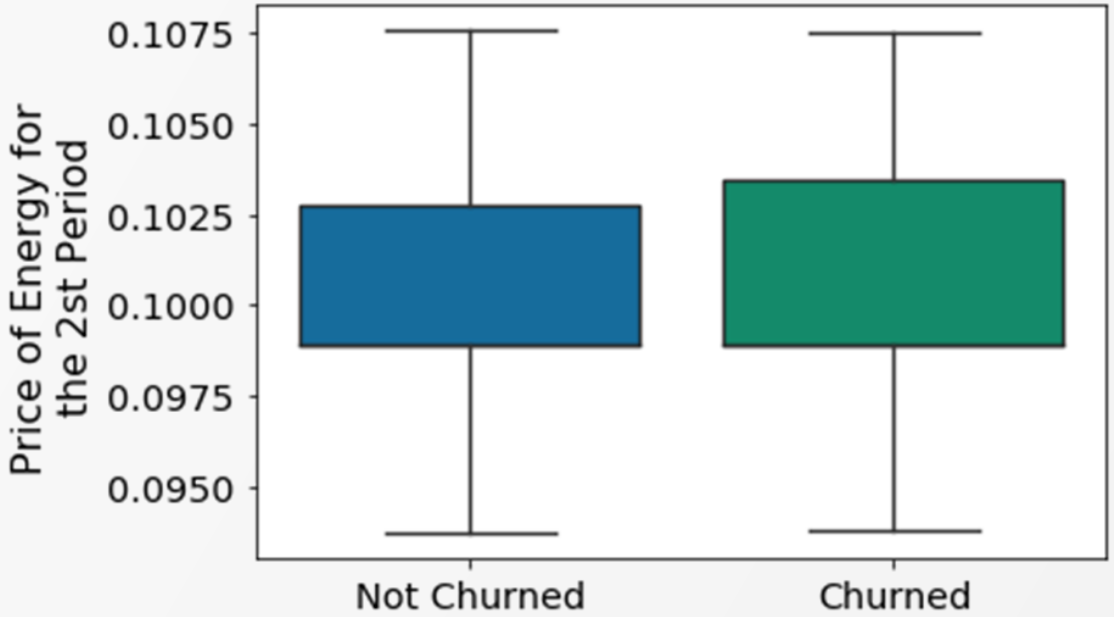
* Electricity campaign the customer first subscribed to has been encrypted

Explanatory Data Analysis

Price of Energy VS. Churn

- As shown in the graph, the median and distribution of energy price is very similar to both type of customers
- So not as assumed, customer turnover is not strongly related to price of energy

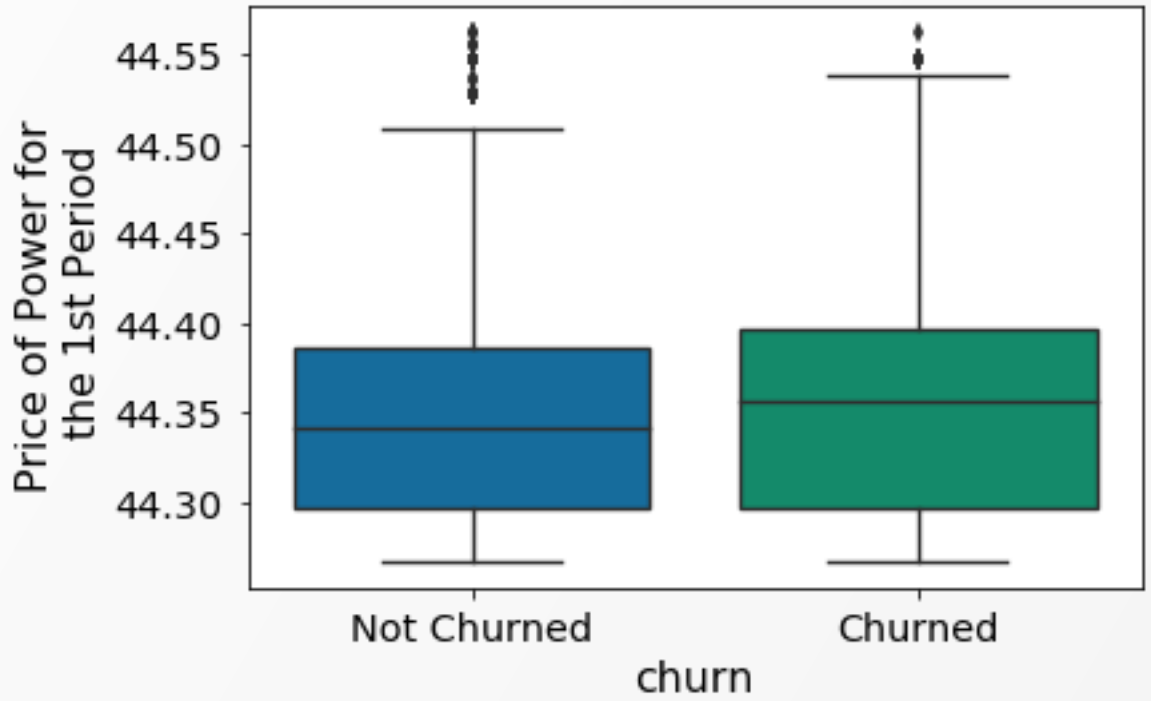




Explanatory Data Analysis

Price of Power VS. Churn

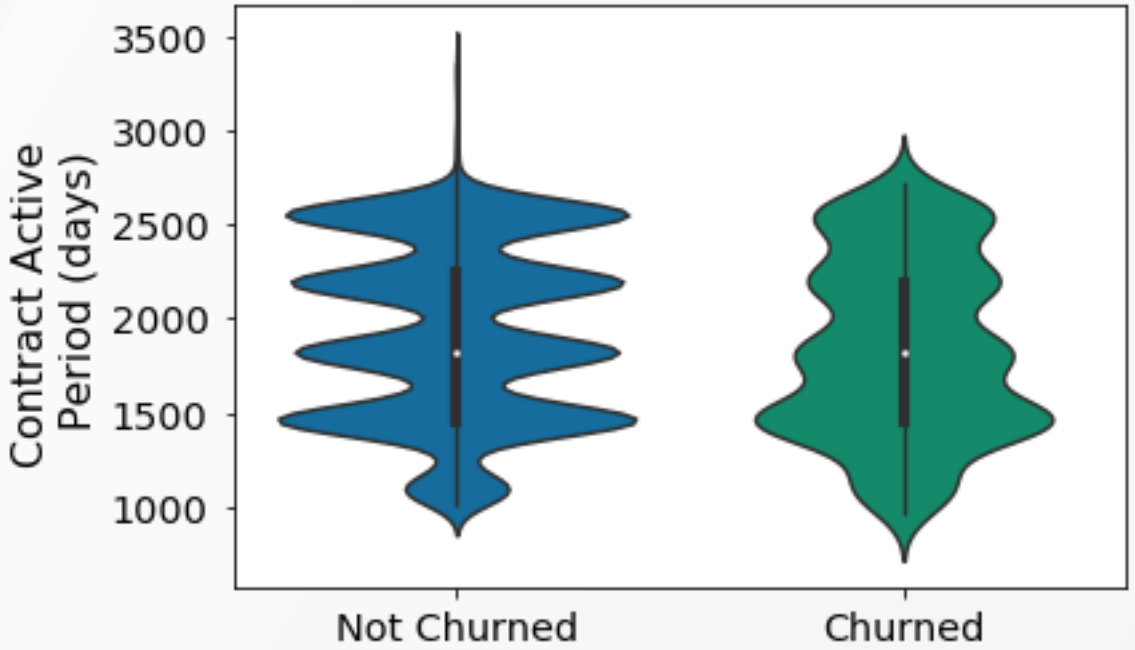
- Price of power for churned customer is slightly higher
- So customer turnover might related to price of power



Explanatory Data Analysis

Contract Active Period VS. Churn

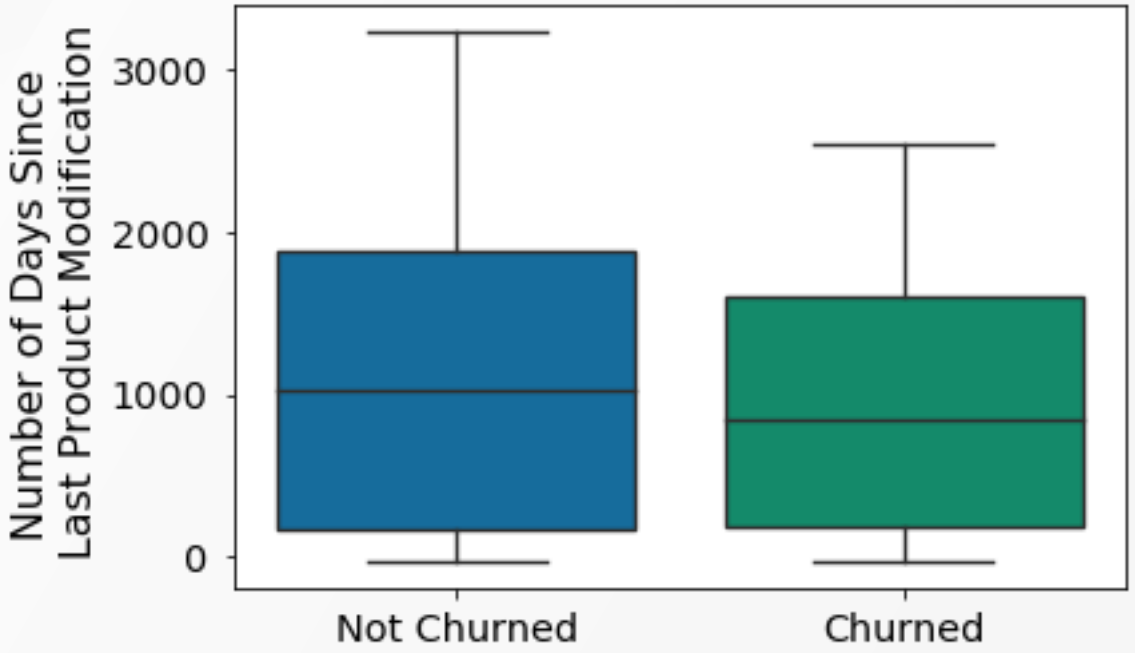
- Churned customers' contract active period mainly distributed between 1000 and 1500 days, while customers not churned have more evenly distributed contract active days
- So churned customers tend to have shorter contract active period



Explanatory Data Analysis

Product Modification VS. Churn

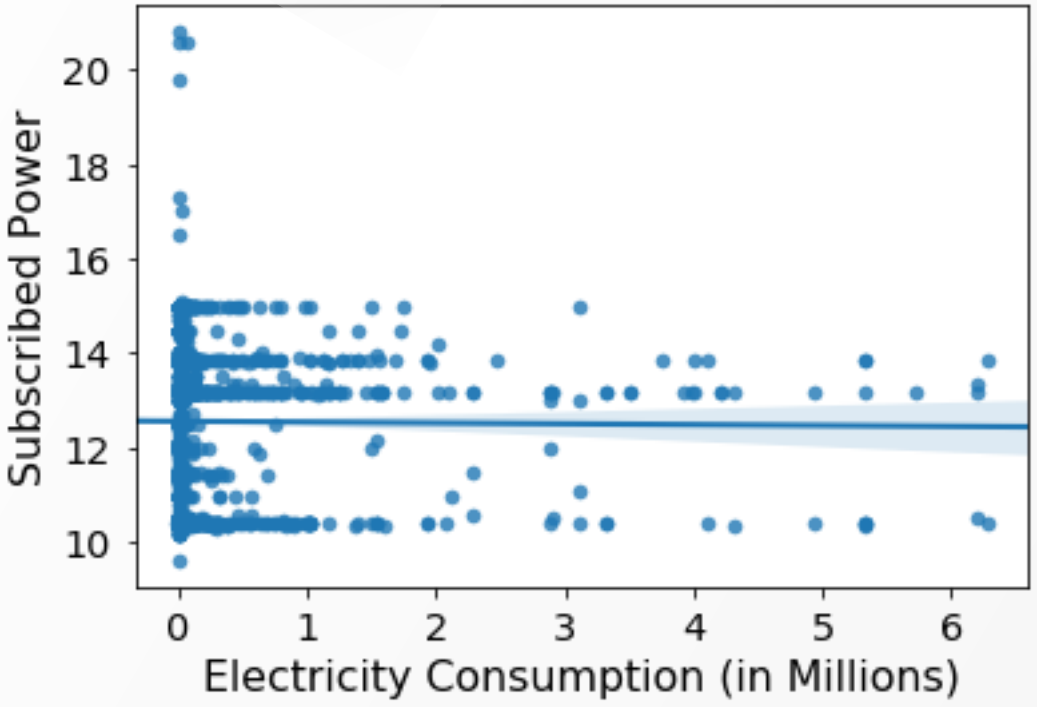
- Product modification may lead to customer churn

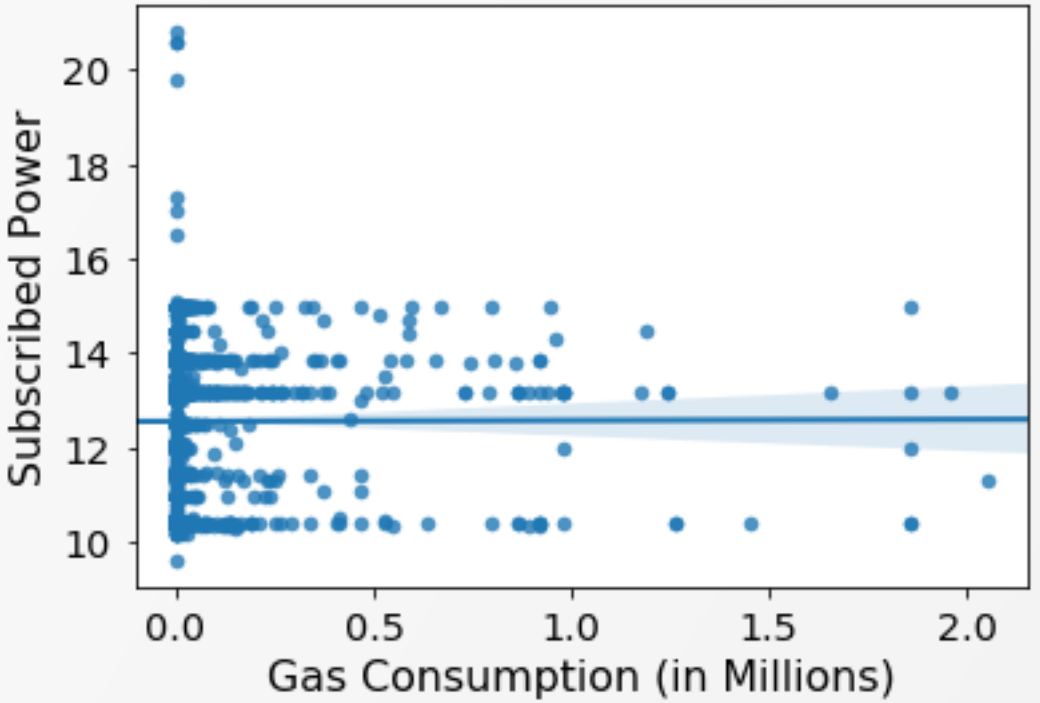


Explanatory Data Analysis

Subscribed Power VS. Consumption

- The figures below show there is no obvious correlation between subscribed power and consumption
- So there is no significant correlation between subscribed power and consumption

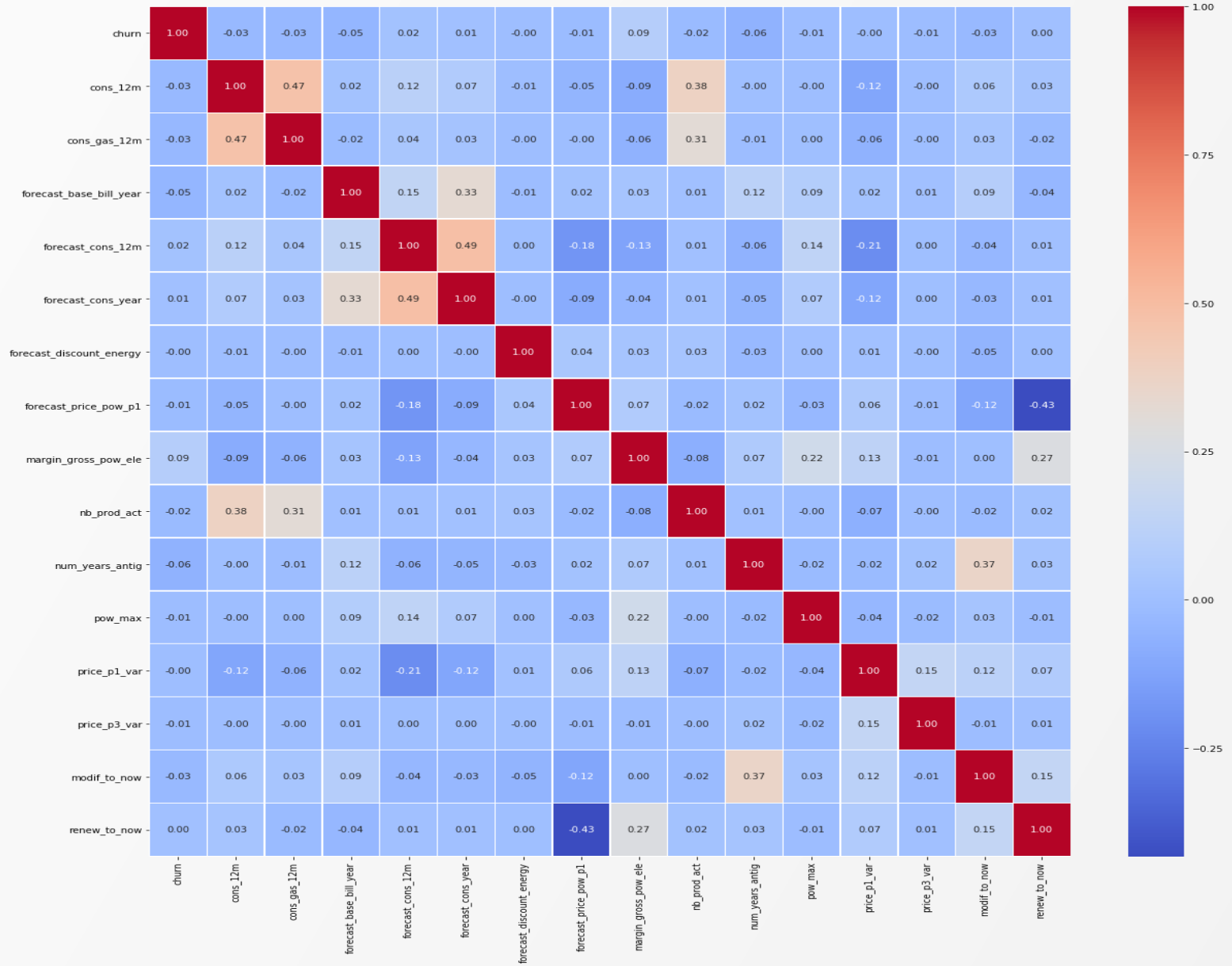




Explanatory Data Analysis

Feature selection

- Remove redundant predictors
- As the heat map shows, the correlations between variables are very small
- So the model input data have low collinearity



Modeling and Evaluation

Algorithm selection and comparison, evaluation of the selected models

Algorithm Selection

Logistic Regression

- A statistical method to predict whether an event will happen or not

Pros:

- Robust algorithm, easier to inspect and less complex

Cons:

- May over fit the data when the training set is high dimensional
- Sensitive to outliers and missing values

Random Forest

- Construct multiple decision trees and use the mode of results from individual trees to make prediction

Pros:

- Higher classification accuracy
- Less prone to overfitting

Cons:

- Difficult to analyze theoretically
- Large number of decision trees may slow down the algorithm

Algorithm Selection

K Nearest Neighbor

- Use the K-Nearest Neighbors of X to vote on the label of X

Pros:

- Simple and powerful
- Naturally handles multi-class cases

Cons:

- Longer computation time
- Need more space to store all training examples

Support Vector Machine (SVM)

- Constructs a hyperplane to separate data points into different classes

Pros:

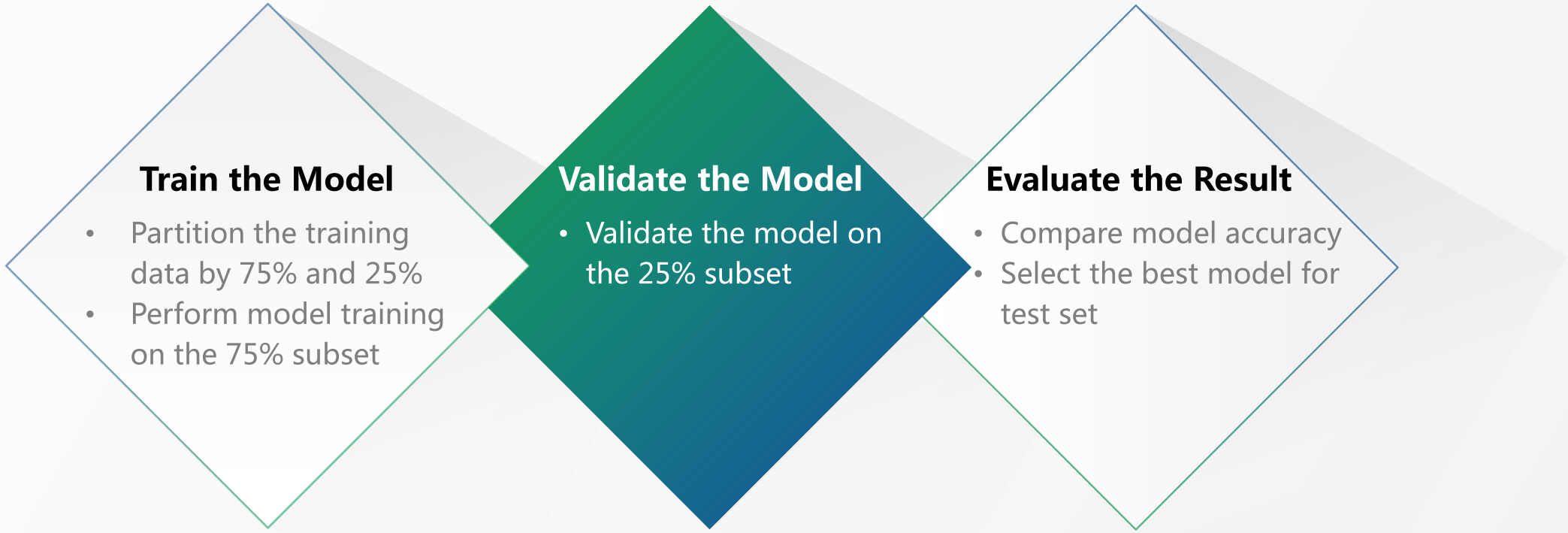
- Excellent performance on the training data
- High accuracy on the test data
- Does not over-fit the data

Cons:

- Longer computation time
- Sensitive to noise

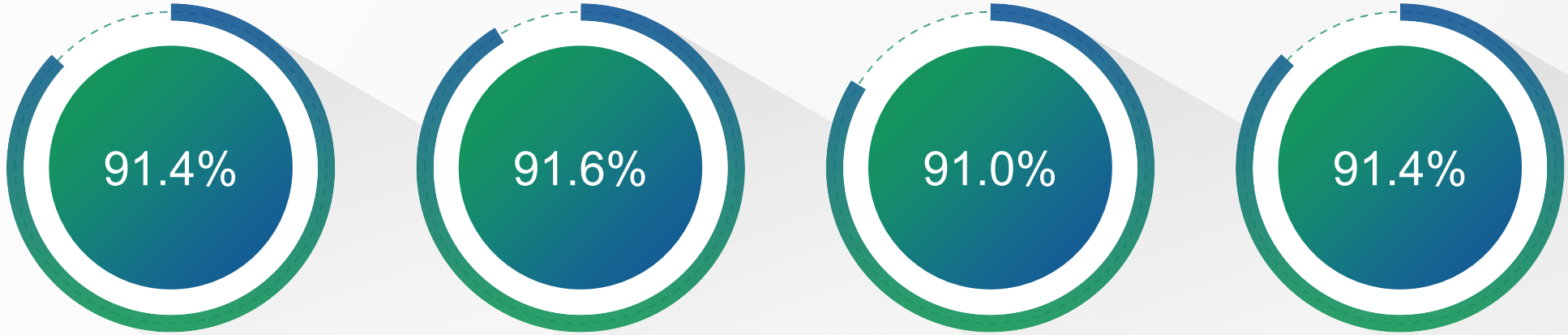
Modeling

The process can be summarized into 3 major steps:



Modeling Results

Accuracy of Models with Different Algorithms:



- Generally, all models performed very well as indicated by high accuracy
- Random forest model is selected for predicting customer churn on test data

04

Conclusions and Recommendations

Findings of the study, interpretation of results and recommendations for next step

Conclusions

The most explicative variables for churn

- Sales channel
- Customer’s activity
- Electricity campaign the customer first subscribed to

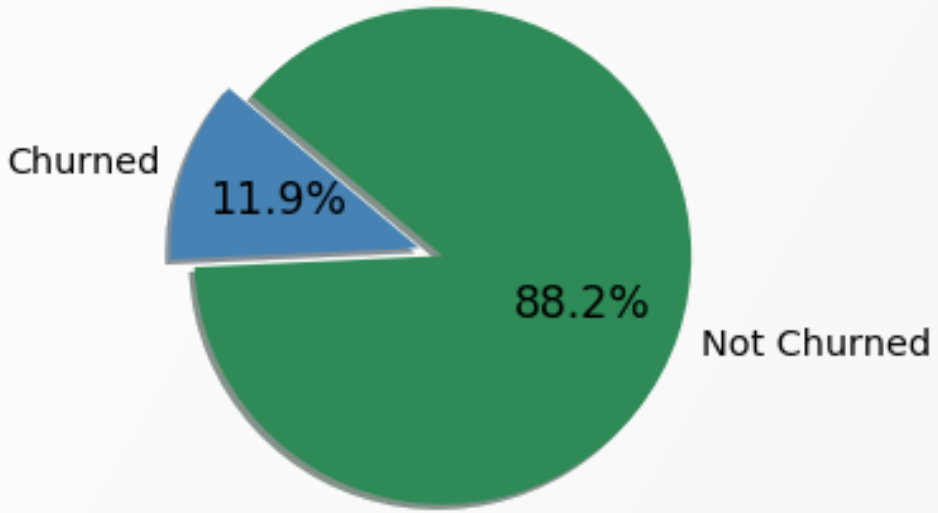
Collect more comprehensive information and dig deeper

- Customer churn is not specifically linked to one single factor
- Need more data on customer’s activity and sales channel
- May conduct experiments on new sales and marketing strategies

Recommendations for Next Step

Churn Prediction on Test Data

- Run random forest model (the model with highest accuracy) on test data, the result shows there are about 12% of customer churned.



Know the customer better

- Pay attention to company's activity
- Conduct customer experience survey
- Collect reasons of churning directly from customers

Retain Existing Customers

- Ramp up customer service
- Product bundling strategy

Attract New Customers

- Select the most effective sales channel to attract new customers

THANK YOU

Stacey (Shiqing) Wang