CHURN MODEL AND ANALYSIS

CONTENTS

Flow Chart

Data Cleaning

- Duplicates
- Outliers
- Missing values

Feature Engineering

- Feature Creation
- Feature Selection
- Normalization and One-hot

Churn Model

- Model Comparison
- Grid Search
- Output

Data Analysis

- Distinguish Feature
- Crosstabs
- Suggestions

01

Data Cleaning

1.1 Data Cleaning

Duplicates

- No duplicate detected

Outliers

- Set negative consumptions and bills as 0
- Set Q1 – 1.5IQR as lower limit and Q3 + 1.5IQR as upper limit

Missing Values

- Exclude columns with more than 50% missing
- KNN imputer for multi-peak distributed columns
- For columns that are close to normal distribution, simply fill with median/mode

02

Feature Engineering

2.1 Feature Engineering

- Length of contract(date_end – data_activ)
- Length of satisfaction(date_end – date_mod)
- Degree of importance(3 if price_p3 > 0)

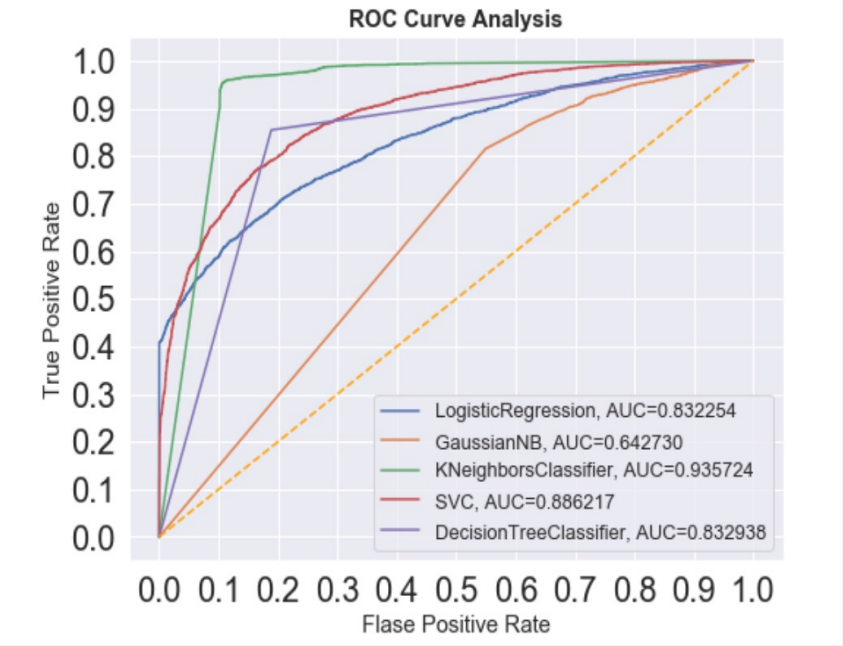
- Select several pairs of features with correlation > 0.75
- Dimension reduction for selected feautres by using PCA

- Apply Min-Max scaling for numeric features
- Create dummy variables for categorical features

03

Churn Model and Output

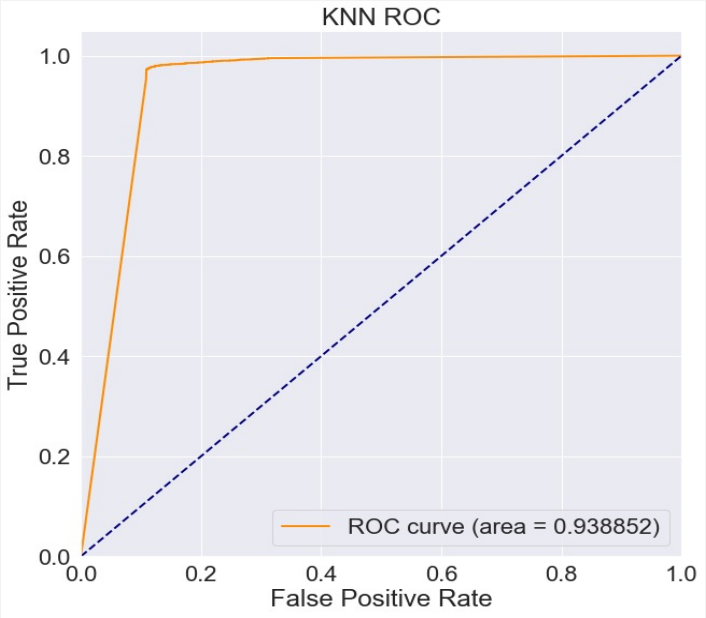
3.1 Model Comparison



Model	Recall	Precision
Logistic Regression	0.750118	0.755665
KNN	0.858499	0.881033
Naïve Bayes	0.517240	0.697621
SVM	0.800618	0.800734
Decision Tree	0.832938	0.833558

- KNN outperform other models in both ROC and Recall metrics
- Choose KNN as the final model and do Grid Search for further improvement

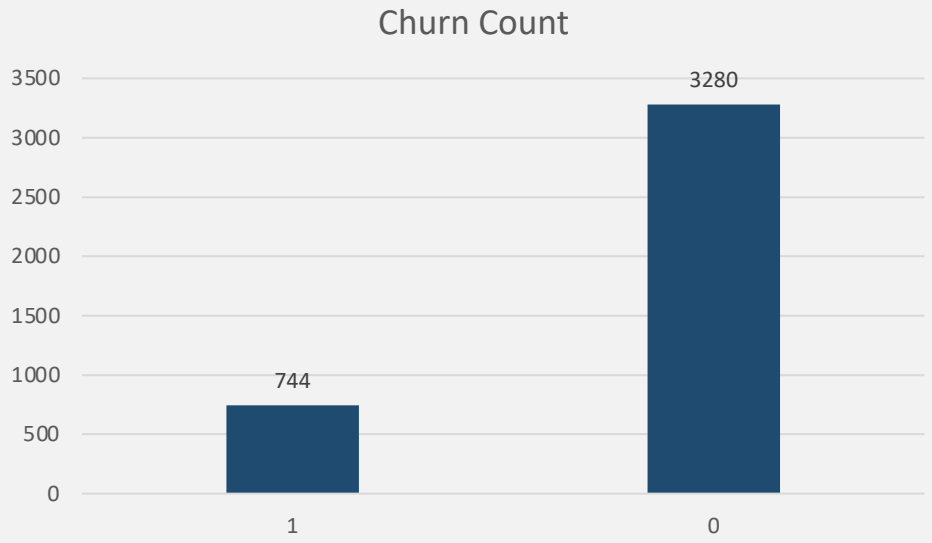
3.2 Grid Search



	Recall	Precision
Previous	0.858498	0.881033
After Grid Search	0.909074	0.918350

- Grid Search does not promote ROC of the model significantly
- Recall and Precision are improved

3.3 Output

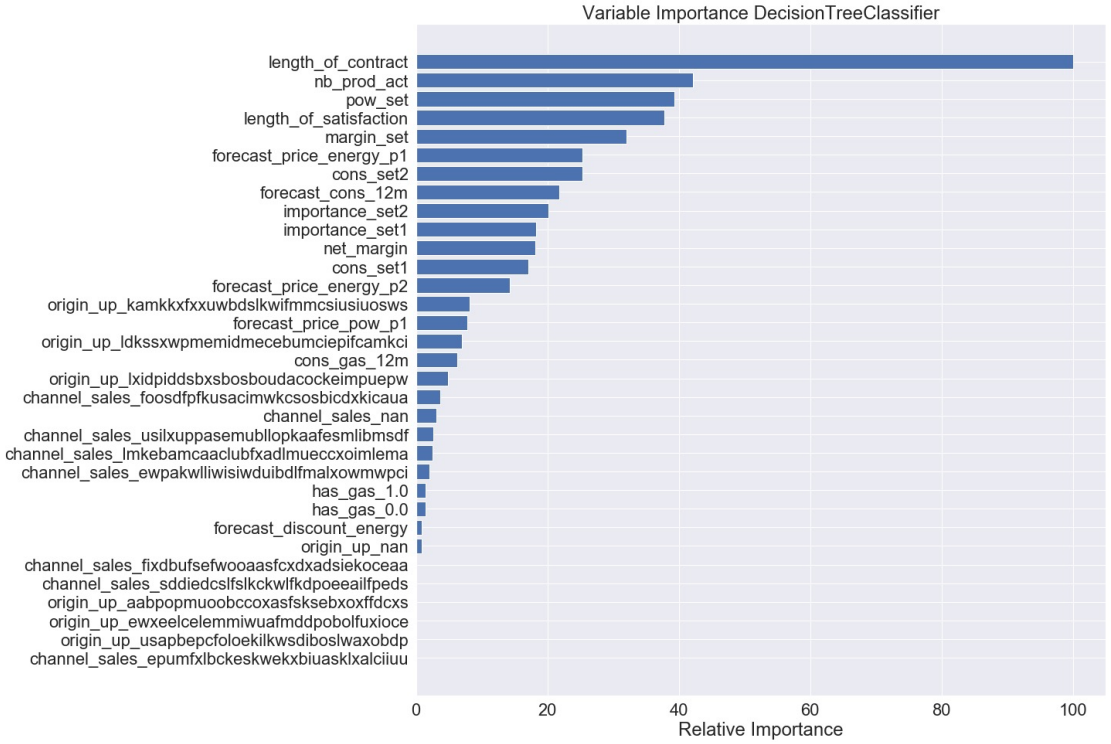


- The Churn problem will be even more serious if our model predicts correctly
- Simply handing clients discount could put a dent in company's bottom line

04

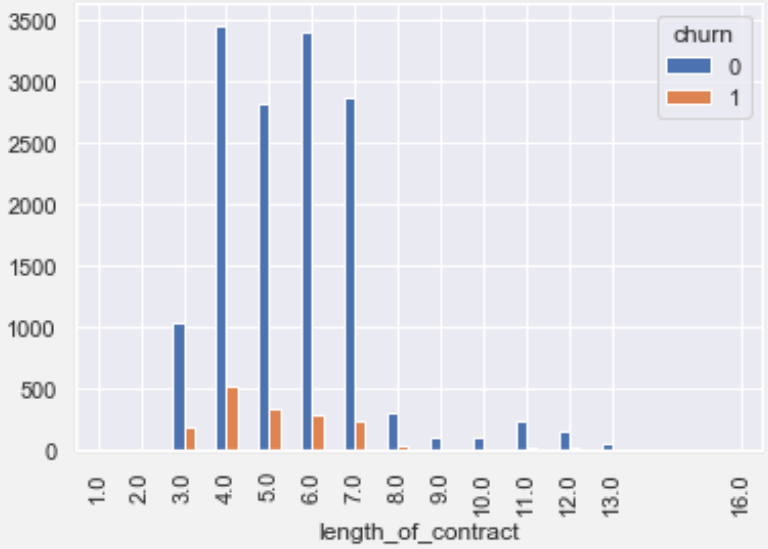
Data Analysis

4.1 Feature Importance



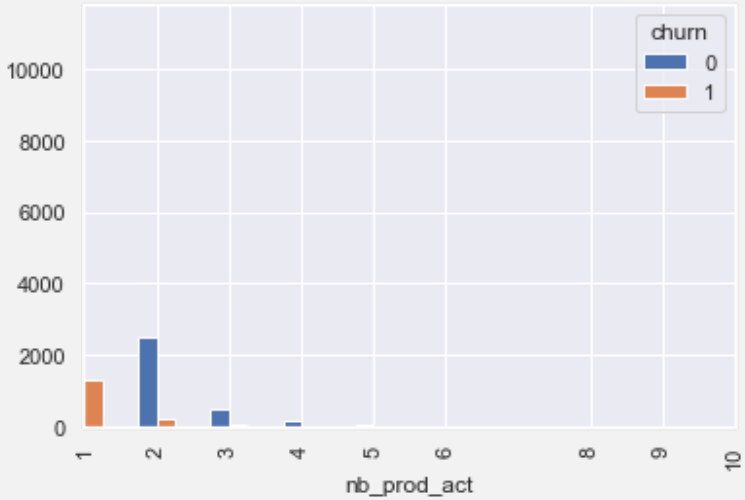
- Although we did not choose Decision Tree as our final model, the feature importance it provides is still useful
- We will do cross-tab analysis of some of those variables

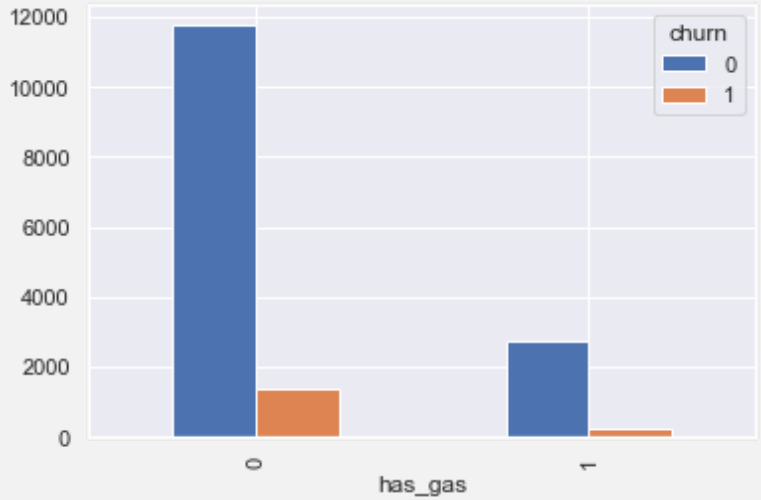
4.2 Cross-tab Analysis



- Clients of longer cooperation have lower churn rate
- The churn rate decreases as the ‘length of satisfaction’ increases

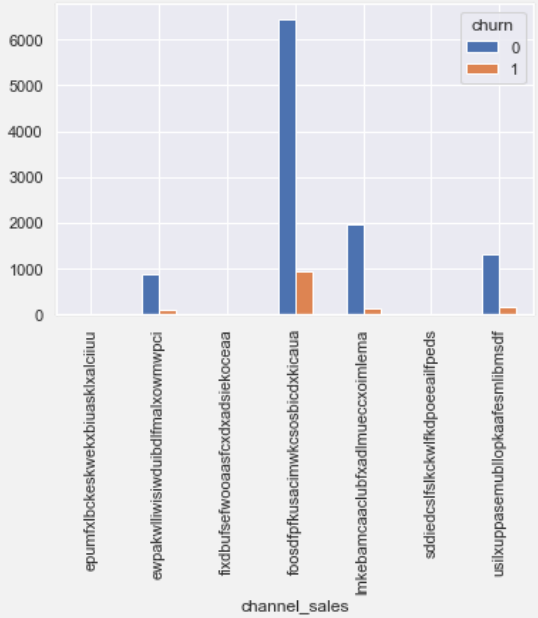
4.2 Cross-tab Analysis

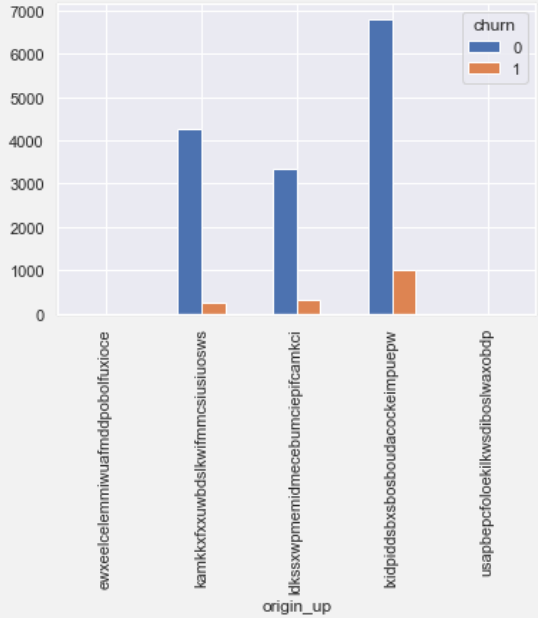




- Clients who have more contracts are less likely to churn
- Clients of both electricity and gas have a lower churn rate

4.2 Cross-tab Analysis





• Clients of ‘foosdfpfkusacimwkcsosbicdxkicaua’ channel_sales are more likely to churn, as well as origin_up is ‘lxidpiddsbxsbosboudacockeimpuepw’ .

4.3 Suggestion

- Do not offer a 20% discount to those clients who are predicted to churn, since they are too many of them. If we do that, it will lower our profitability dramatically
- Focus on clients that have longer cooperation and more contracts with us
- Investigate in ‘channel_sales’ and ‘origin_up’,
to find why ‘foosdfpfkusacimwkcsosbicdxkicaua’ and ‘foosdfpfkusacimwkcsosbicdxkicaua’ clients have a significantly higher churn rate, and to see whether there is a chance to smooth over it.

THANKS FOR LISTENING!