Objective:

The purpose of the document is to prompt your thinking on how to understand and present a set of data to a client on behalf of Capgemini Invent. There are no hard and fast rules for this … deliberately so. The format and style of presentation are your choice, as long as the code is provided alongside the case study for review. How you package this is up you the only restriction is that our team must be able to run what you have built, our preference is Python, however anything can be used.

You will have 15 minutes to present your findings which you can structure as you see fit followed by a Q&A session. Slide count is up to you, however we suggest aiming for concise.

You will be presenting to a Panel consisting of members of the Capgemini Invent Business, Insights Driven Enterprise team, amongst others.

Client problem:

The dataset provided is for a client who is a utility in the USA named Waterco, they want to understand if there is an issue with customers paying a given invoice late, and what factors influence this.

Areas to cover:

- What are the key findings from the dataset that are important and relevant to the clients problem?
- How did you arrive at these findings?

Problem:

- Supervised Classification

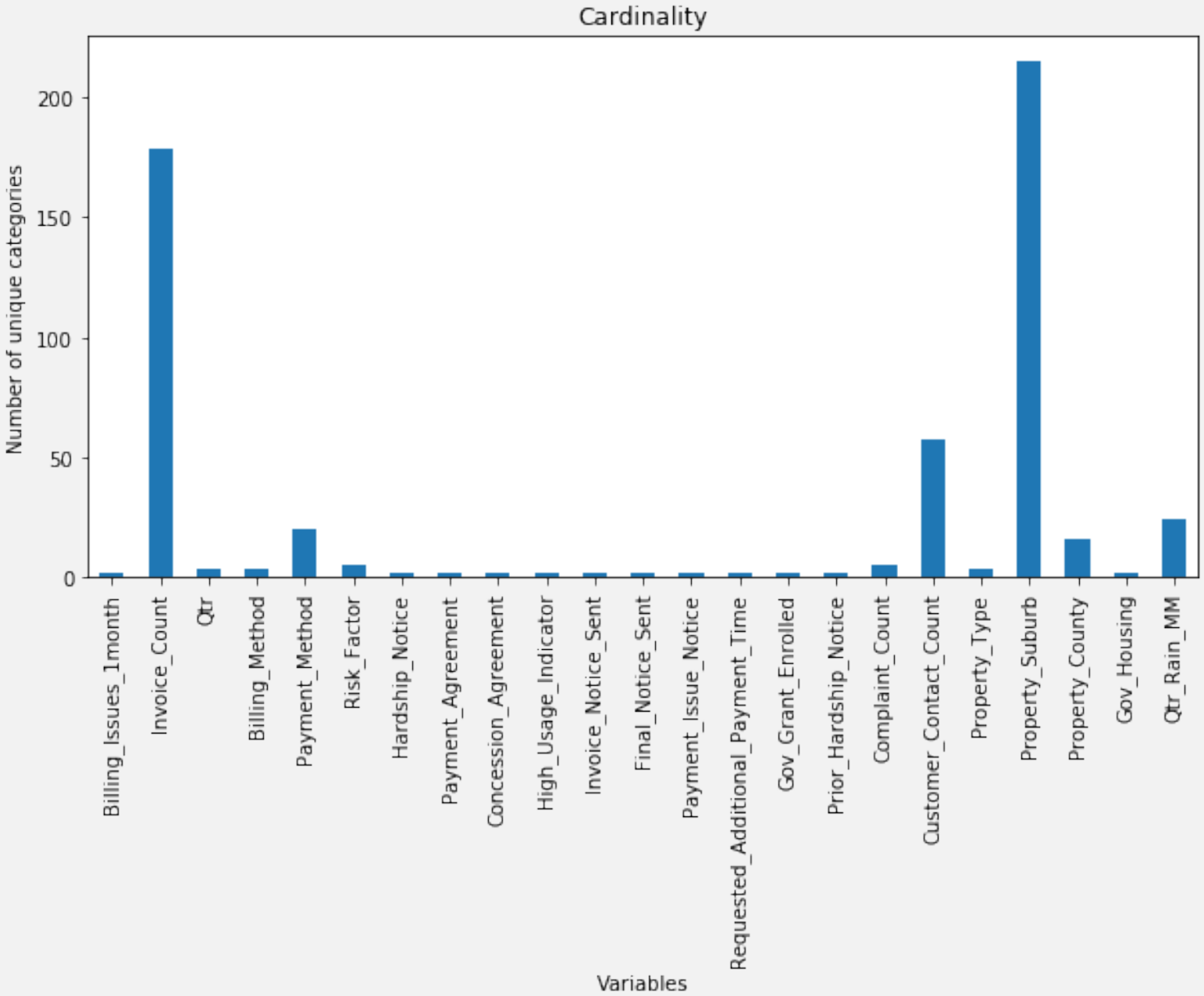
Data Description

- Dataset shape: (313962, 24)
- Target: Billing_Issues_1month (Yes/No)
- Categorical (Nominal):
  - Yes/No: Hardship_Notice, Payment_Agreement, Concession_Agreement, High_Usage_Indicator, Invoice_Notice_Sent, Final_Notice_Sent, Payment_Issue_Notice, Requested_Additional_Payment_Time, Gov_Grant_Enrolled, Prior_Hardship_Notice, Gov_Housing
  - Billing_Method (Mail, Email, BPAY View)
  - Payment_Method (Payment_Method 1-20)
  - Risk_Factor (Low Risk, Excluded, Medium Risk, High Risk, Risk Not Calculated)
  - Property_Type (Dwelling, Apartment, Unit)
  - Property_Suburb (Property_Suburb 1-248)
  - Property_County (Property_County 1-15 & 17)
- Numerical (Nominal in reality):
  - Invoice_Count (2-276)
  - Qtr (1, 2, 3)
  - Complaint_Count (0, 1, 2, 3, 4)
  - Customer_Contact_Count (1-76)
  - Qtr_Rain_MM (24 values from 0 to 277.4)

Data Cleaning

- Missing values: Complain_Count
  - Null values are in fact 0 → null changed to 0
- Row duplicates removal
  - new data shape: (288187, 24)
- Features with constant values: Year
  - Year only has one value (2017) → removed from dataset
    - Data shape: (288187, 23)

Cardinality Analysis



Feature Hashing (Binning)

- Features with high cardinality:
  - 'Property_Suburb’
  - 'Customer_Contact_Count’
  - 'Invoice_Count’
  - 'Property_County’
  - 'Qtr_Rain_MM’
  - 'Payment_Method’
- Object type features → array → dataframe → concat with original data → removal of original columns
- Feature hashing does the encoding and creates dummy variables (based on the chosen n_features) at the same time

One Hot Encoding

- Features with Yes/No items
- Billing_Method (Mail: 1, Email: 2, BPAY View: 3)
- Risk_Factor (Low Risk: 1, Excluded: 2, Medium Risk: 3, High Risk: 4, Risk Not Calculated: 5)
- Property_Type (Dwelling: 1, Apartment: 2, Unit: 3)
- Qtr (1, 2, 3)
- Complaint_Count (0, 1, 2, 3, 4)

Dummy variables

- get_dummies from Pandas library
  - The functionality creates dummy features for each item within a feature after the encoding step
  - Data shape: (288187, 78)

Feature Selection: Chi-square

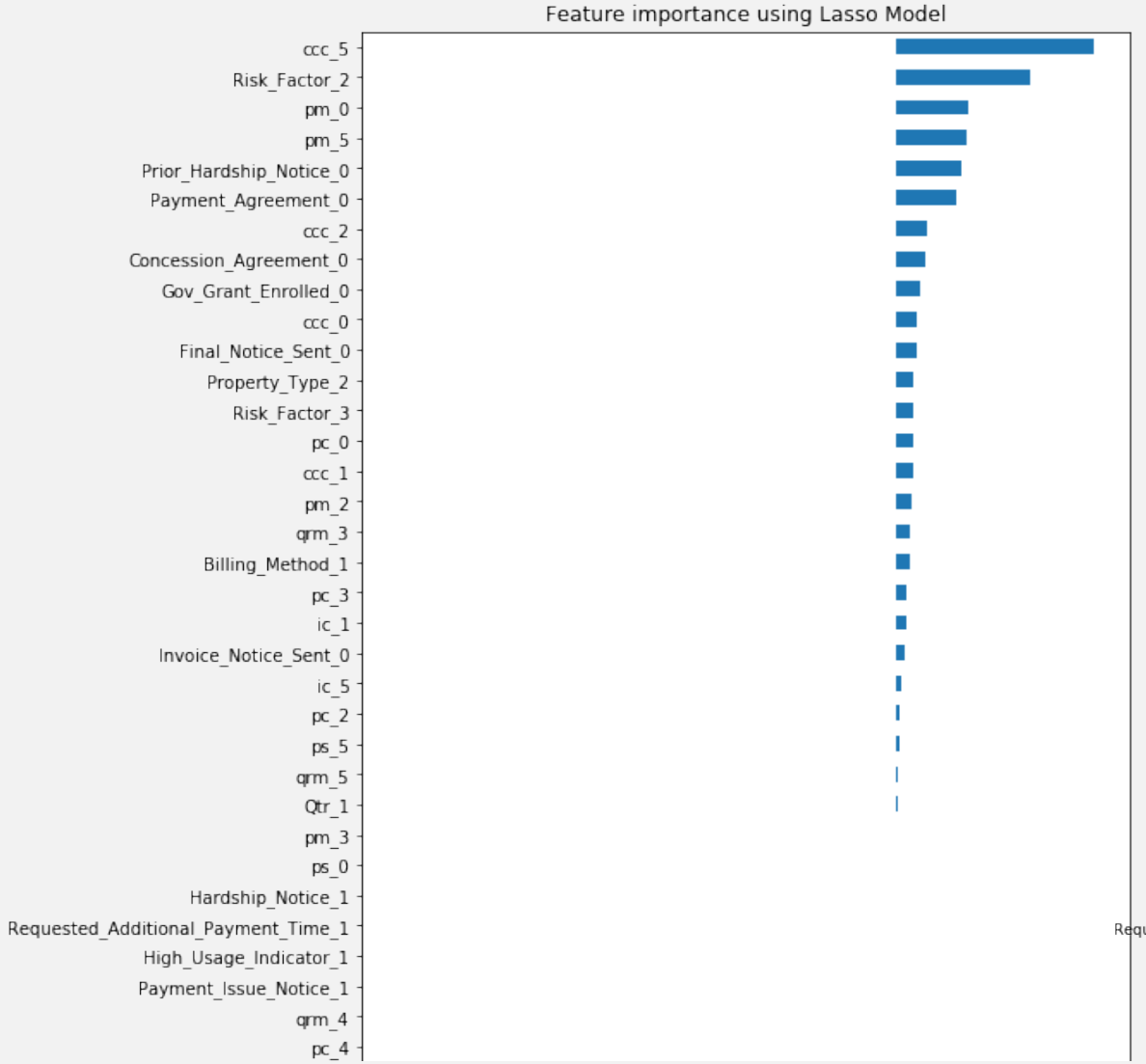
- The Chi-Square: to determine if there is a significant relationship between two categorical (nominal) variables. Null Hypothesis (H0): There is no relationship between the variables
- Alternative Hypothesis (H1): There is a relationship between variables
- If the p-value > 0.05 null hypothesis is rejected

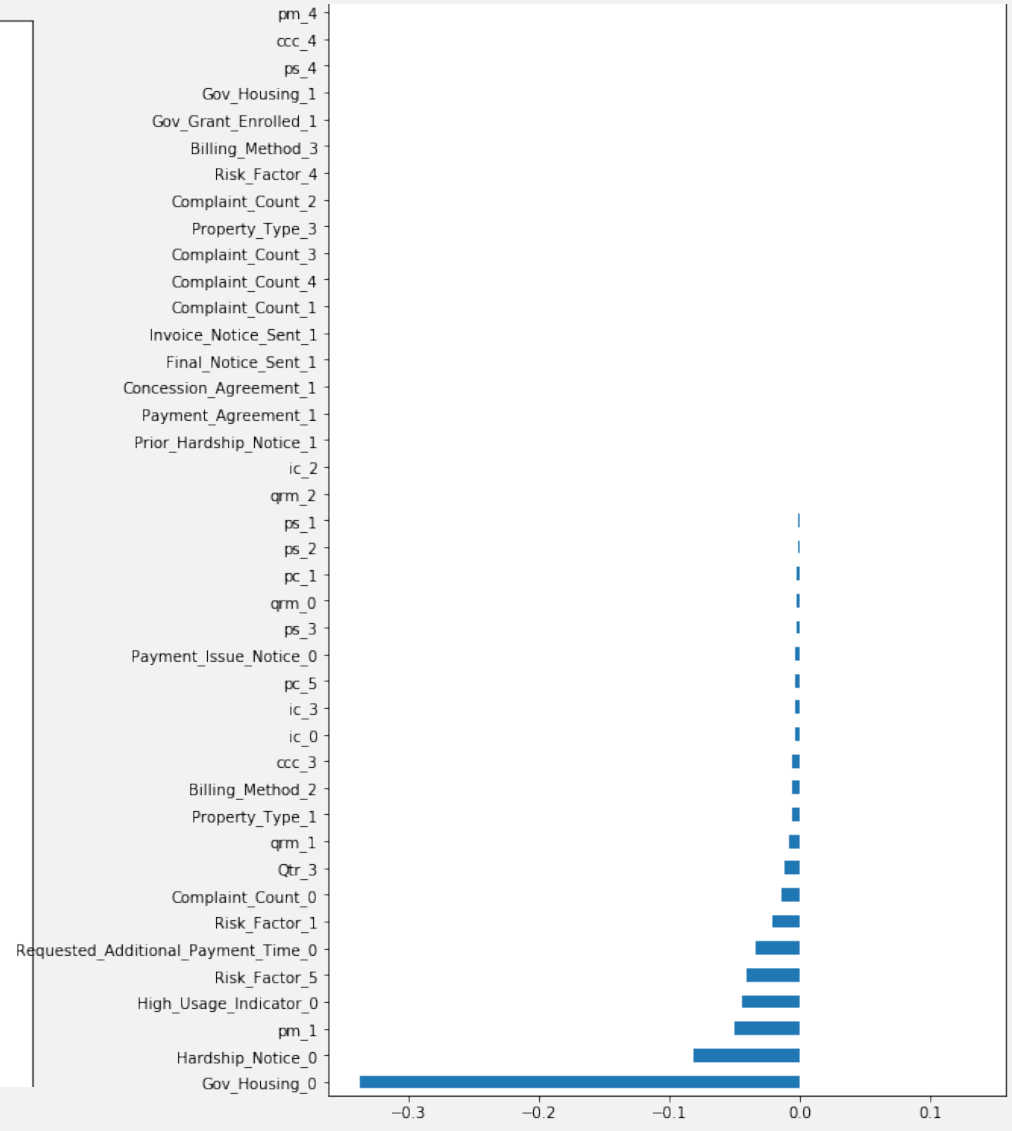
Non_Significant Features

	Column	Hypothesis
4	ps_4	Fail to Reject Null Hypothesis
5	ps_5	Fail to Reject Null Hypothesis
10	ccc_4	Fail to Reject Null Hypothesis
16	ic_4	Fail to Reject Null Hypothesis
22	pc_4	Fail to Reject Null Hypothesis
28	qrm_4	Fail to Reject Null Hypothesis
34	pm_4	Fail to Reject Null Hypothesis
36	Qtr_1	Fail to Reject Null Hypothesis
37	Qtr_2	Fail to Reject Null Hypothesis
49	Payment_Agreement_0	Fail to Reject Null Hypothesis
50	Payment_Agreement_1	Fail to Reject Null Hypothesis
69	Complaint_Count_2	Fail to Reject Null Hypothesis
70	Complaint_Count_3	Fail to Reject Null Hypothesis
71	Complaint_Count_4	Fail to Reject Null Hypothesis
74	Property_Type_3	Fail to Reject Null Hypothesis

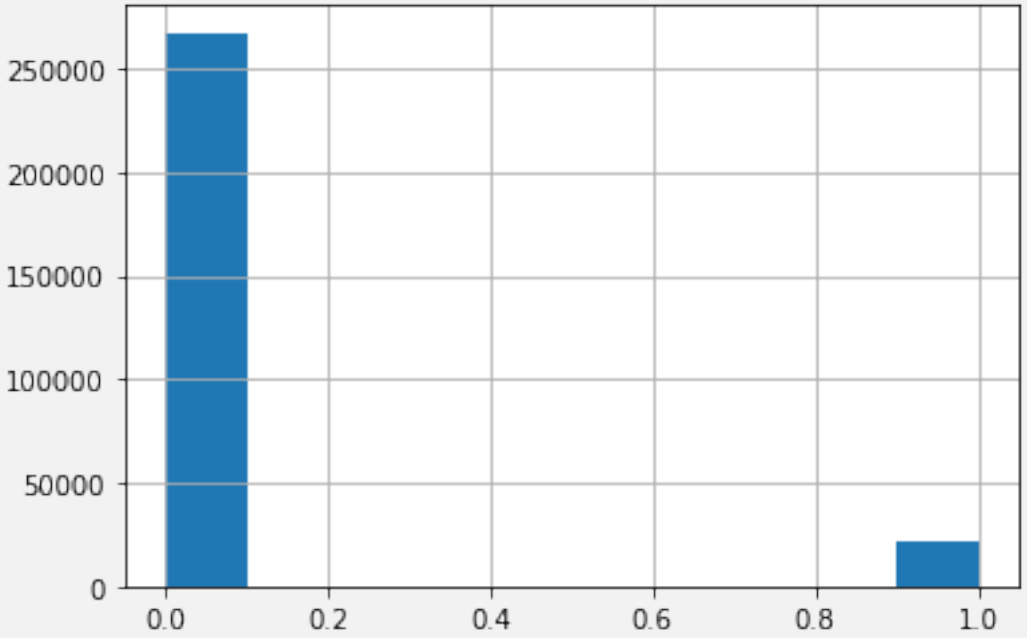
Feature Selection: Embedded

- Lasso regularization: penalize a feature given a coefficient threshold





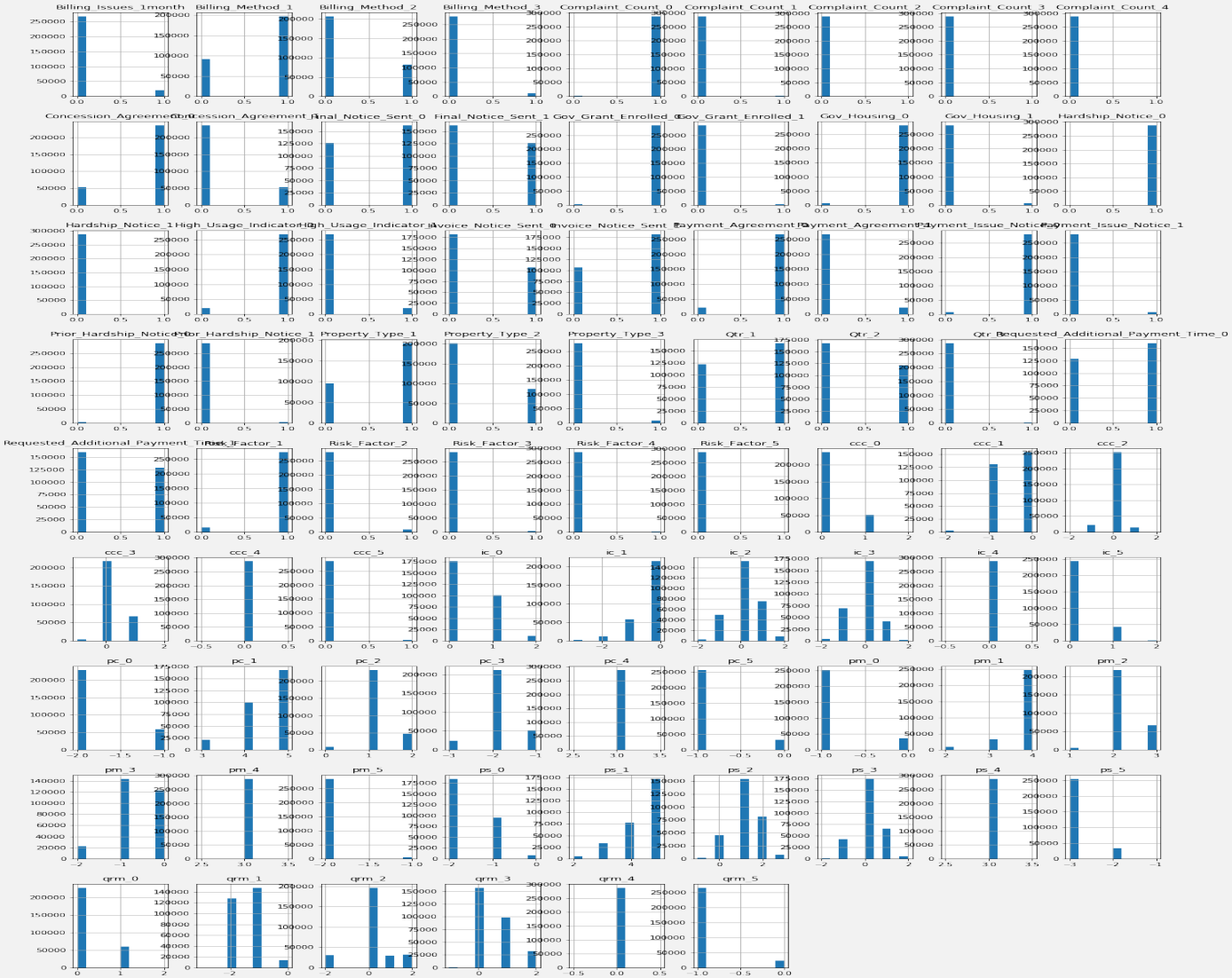
Data Distribution:



['Billing_Issues_1month'].value_counts():

- 0: 267094 → 0.926808
- 1: 21093 → 0.073192

→ Imbalanced data



Imbalanced Dataset

- It is important to have balanced data for model training to avoid overfitting.
  - Imbalanced data results in a good accuracy for the category with large numbers
  - But it fails to predict the category with small numbers

- Data Balancing techniques used:
  - Downsampling:
    - The larger category is downsampled to match the number of the smaller category
      - 0: 267094 → downsampled to 21093
      - 1: 21093
  - weighted approach
    - Weights and penalties are assigned as part of the modelling
    - {0: 7, 1:93}, which means if the model predicts a 1 wrongly, it will be penalized by 93/7 more times compared to the value 0.

Modelling – Logistic Regression

- Logistic Regression is a good model for binary classification problems
  - Baseline: the default model is done on the actual data prior to any dummy feature creation, feature selection or resampling steps.

```
Logistic test score: 0.20714670003072178
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     53459
           1       0.78      0.19      0.31      4179

    accuracy                           0.94     57638
   macro avg       0.86      0.59      0.64     57638
weighted avg       0.93      0.94      0.92     57638

Accuracy: 0.9374197577986745
Precision: 0.9287354326761028
Recall: 0.9374197577986745
F score: 0.9192488566984257
Area Under Curve: 0.5928121530728648
```

Baseline: without dummy variables

```
Logistic test score: 0.1984089839055091
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     53459
           1       0.82      0.17      0.28      4179

    accuracy                           0.94     57638
   macro avg       0.88      0.58      0.62     57638
weighted avg       0.93      0.94      0.92     57638

Accuracy: 0.9370380651653423
Precision: 0.9303028079562115
Recall: 0.9370380651653423
F score: 0.9171899721425771
Area Under Curve: 0.5825697352622932
```

Baseline: with dummy variables

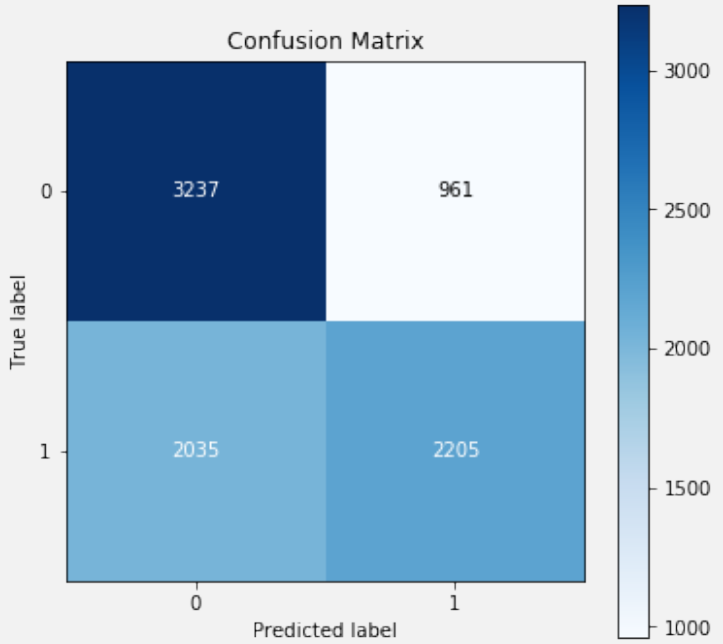
→ The impact of Imbalanced data

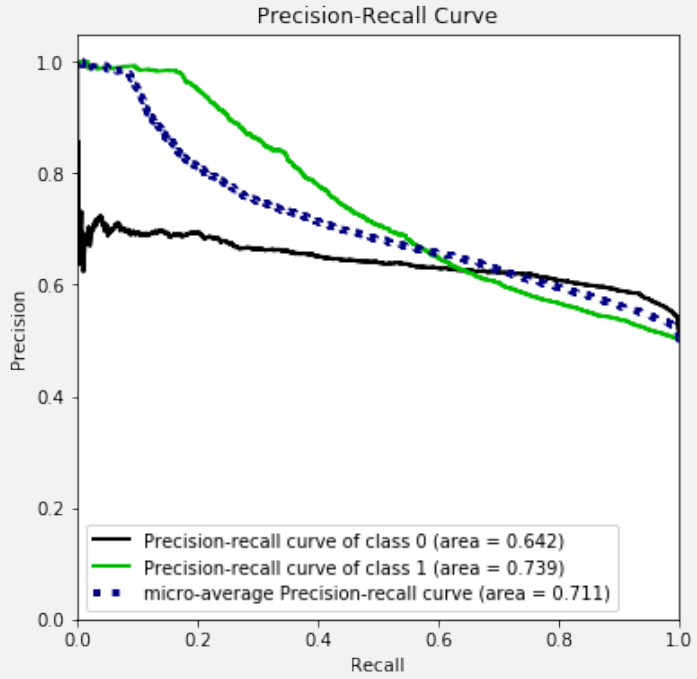
Modelling – Logistic Regression

```
Logistic test score: 0.6038021862674593
              precision    recall  f1-score   support

           0       0.61      0.78      0.68      4198
           1       0.70      0.52      0.59      4240

    accuracy                           0.65      8438
   macro avg       0.66      0.65      0.64      8438
weighted avg       0.66      0.65      0.64      8438

Accuracy: 0.6450580706328514
Precision: 0.6562923008767773
Recall: 0.6450580706328514
F score: 0.6390632700777389
Area Under Curve: 0.645703479644395
```

With dummy variables and resampling

→ Performance improved





Modelling – Logistic Regression

```
Logistic test score: 0.10879248963820064
              precision    recall  f1-score   support

           0       0.95      0.74      0.83     53459
           1       0.14      0.54      0.22      4179

    accuracy                           0.72     57638
   macro avg       0.55      0.64      0.53     57638
weighted avg       0.89      0.72      0.79     57638

Accuracy: 0.7225094555675076
Precision: 0.8948351133366033
Recall: 0.7225094555675076
F score: 0.7869650774294304
Area Under Curve: 0.6406316548792712
```

With dummy variables and weighting

```
Logistic test score: 0.19461872112229836
              precision    recall  f1-score   support

           0       0.94      1.00      0.97     53459
           1       0.77      0.18      0.29      4179

    accuracy                           0.94     57638
   macro avg       0.85      0.59      0.63     57638
weighted avg       0.93      0.94      0.92     57638

Accuracy: 0.9363614282244352
Precision: 0.9266812228673573
Recall: 0.9363614282244352
F score: 0.9173700359571478
Area Under Curve: 0.5859549278720557
```

With dummy variables and SMOTE approach

→ Poor Performance, There are methods to calculate more accurate weighting parameters, That can be investigated.
→ Downsampling sampling had a better performance compared to weighting

Modelling – Logistic Regression

```
Logistic test score: 0.603364200409893
              precision    recall  f1-score   support

           0       0.61      0.77      0.68      4198
           1       0.70      0.52      0.60      4240

    accuracy                           0.64      8438
   macro avg       0.66      0.65      0.64      8438
weighted avg       0.66      0.64      0.64      8438

Accuracy: 0.6449395591372363
Precision: 0.6554356792431824
Recall: 0.6449395591372363
F score: 0.6393284000764705
Area Under Curve: 0.6455643185883664
```

With dummy variables, downsampling and chi-square feature selection

```
Logistic test score: 0.6028239619504981
              precision    recall  f1-score   support

           0       0.61      0.78      0.68      4198
           1       0.70      0.51      0.59      4240

    accuracy                           0.64      8438
   macro avg       0.66      0.64      0.64      8438
weighted avg       0.66      0.64      0.64      8438

Accuracy: 0.6437544441810855
Precision: 0.6552992546365093
Recall: 0.6437544441810855
F score: 0.6375219268478257
Area Under Curve: 0.6444110290614578
```

With dummy variables, downsampling and wrapper feature selection

→ Feature selection did not improve the model performance
→ Other encoding/feature selection/data balancing techniques should be investigated to improve The performance.

Modelling – Logistic Regression

- Summary:
  - Important features (table)
- Performance (f1-score =~60%)
- To improve performance:
  - The model needs further work with
    - Binning & Encoding
    - Feature Engineering
    - Feature Selection
    - Data Balancing

features	coef
ccc_5	1.230877
Gov_Housing_1	1.129117
Hardship_Notice_1	0.879311
Prior_Hardship_Notice_0	0.568679
equested_Additional_Payment_Time_1	0.563071
pm_5	0.483260
High_Usage_Indicator_1	0.477963
Risk_Factor_2	0.456599
Final_Notice_Sent_0	0.449320
Payment_Issue_Notice_1	0.418561

Modelling – Decision Tree

	precision	recall	f1-score	support
0	0.62	0.63	0.62	4198
1	0.63	0.62	0.62	4240
accuracy			0.62	8438
macro avg	0.62	0.62	0.62	8438
weighted avg	0.62	0.62	0.62	8438

Accuracy: 0.6239630244133682
Precision: 0.6239970918862818
Recall: 0.6239630244133682
F score: 0.6239623114202593
Area Under Curve: 0.6239802534000917

With dummy variables, downsampling and chi-square feature selection

Modelling – Decision Tree

```
|--- Gov_Housing_0 <= 0.50
|    |--- Concession_Agreement_0 <= 0.50
|    |    |--- ccc_0 <= 0.50
|    |    |    |--- ccc_1 <= -1.50
|    |    |    |    |--- Risk_Factor_4 <= 0.50
|    |    |    |    |    |--- class: 1
|    |    |    |    |--- Risk_Factor_4 >  0.50
|    |    |    |    |    |--- class: 0
|    |    |    |--- ccc_1 >  -1.50
|    |    |    |    |--- pm_1 <= 2.50
|    |    |    |    |    |--- pc_3 <= -2.50
|    |    |    |    |    |    |--- class: 1
|    |    |    |    |    |--- pc_3 >  -2.50
|    |    |    |    |    |    |--- Prior_Hardship_Notice_1 <= 0.50
|    |    |    |    |    |    |    |--- ps_2 <= 1.50
|    |    |    |    |    |    |    |    |--- Billing_Method_2 <= 0.50
|    |    |    |    |    |    |    |    |    |--- qrm_3 <= 1.50
|    |    |    |    |    |    |    |    |    |    |--- qrm_1 <= -0.50
```

Important features

Summary

- From LR, DT and Feature Selection approaches, it is evident that:
  - Government Housing
  - Customer Contact Count
  - Hardship Notice
  - Risk Factor

  → are the most important feature to be considered for prediction purposes

Data Definitions

Billing_Issues_1month – This is the primary column to predict, this indicates if the account has raised distress with the invoice provided

Invoice_Count – Count of invoices sent to this account

Qtr – Quarter the invoice was sent in

Year – Year the invoice was sent in

Billing_Method – Method used to communicate the invoice

Payment_Method - Method used to pay for the invoice

Risk_Factor – Predetermined risk factor of the account

Hardship_Notice – If the account is currently placed on hardship (unable to pay a bill, etc)

Payment_Agreement – If the account has an agreement for payment of previous outstanding invoices

Concession_Agreement – If the account has any agreed concession for payment of invoices

High_Usage_Indicator – If the account has been determined to be a high-water user

Invoice_Notice_Sent – Invoice sent to the account

Final_Notice_Sent – Final invoice reminder sent to the account

Payment_Issue_Notice – Notice issue to the account if there are payment issues with reoccurring payment plans

Requested_Additional_Payment_Time – If the account has requested additional time to pay a bill

Gov_Grant_Enrolled – If the are enrolled in the Welfare or Temporary Assistance for Needy Families (TANF) provided by the USA Government

Prior_Hardship_Notice - If the account has been placed on hardship (unable to pay a bill, etc) previously

Complaint_Count – Count of complaints made by the account to the service team

Customer_Contact_Count – Count of the amount of contact between account and service team

Property_Type – Type of property

Property_Suburb – Suburb of property location

Property_County – County of property location

Gov_Housing - If the are enrolled in the Welfare or Temporary Assistance for Needy Families (TANF) provided by the USA Government in Government housing

Qtr_Rain_MM – Measurement of rainwater for County quarterly