Payment Default Detection

Team TabWeau

22 March 2023

Team TabWeau

Sandeep Kumar

Mounika Yallamandhala

Divisha Jain

Naveen Parthasarathy

Sriya Kondabathula

Agenda

Overview | Methodology | Business Implications | Questions

Overview

- What is the problem?
- What do we have?
- What do we do?

01	**Predicting Behavior**	• What makes a customer likely to make on time payments?
02	**300,000 x 55**	• Large dataset - high dimensionality
03	**Balanced Data**	• Default vs. Non-Default
04	**Binary Classification**	• Determining YES or NO

Methodology

- What is our approach?

1 **Exploratory Data Analysis**

Understanding what we are working with

2 **Data Cleaning**

Performing transformations and feature engineering to prepare the data

3 **Model Development**

Utilizing data for training, culminating in the selection of the best performing model

4 **Model Evaluation**

Measuring model effectiveness on validation data

5 **Prediction**

Predicting targets of test data using the final model

Exploratory Data Analysis

Missing data

PrevAccountStatus1	95.81%
AccountStatus1	66.73%
PrevAccountDetail1	58.9%
AccountDetail6	47.46%
AccountStatus2	35.57%
PrevAccountStatus2	31.67%

11 columns - <6%

30 columns - <1%

Date columns

	CurrentDate	AccountDetail2	AccountStatus2	PrevAccountDetail1	AccountDetail8	Payment2	Payment4
0	11/1/2017	5/1/2016	NaN	NaN	7/28/2017	7/7/2017	11/30/2017
1	11/1/2017	4/1/2015	9/1/2017	4/1/2015	11/2/2017	10/24/2017	11/3/2017
2	11/1/2017	8/1/2016	4/3/2017	NaN	11/3/2017	11/26/2017	11/4/2017
3	11/1/2017	7/1/2017	NaN	NaN	9/25/2017	9/5/2017	11/26/2017
4	11/1/2017	5/1/2016	6/7/2017	5/1/2016	11/7/2017	11/2/2017	11/8/2017

Categorical Columns

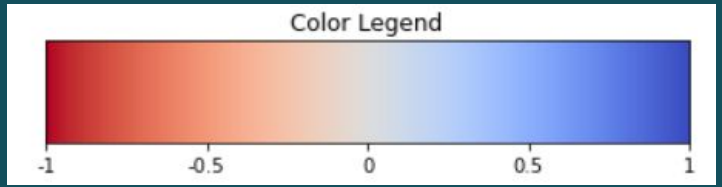
	AccountDetail5	AccountDetail6	AccountStatus1	PrevAccountStatus1	PrevAccountStatus2
0	NaN	NaN	NaN	NaN	NaN
1	X	X	NaN	NaN	D
2	X	NaN	NaN	NaN	O
3	NaN	NaN	NaN	NaN	NaN
4	X	NaN	NaN	NaN	D

Exploratory Data Analysis

Correlation matrix

	Balance2	HistoricalAccountDetail2	HistoricalAccountStatus3	AccountActivity7	AccountDetail3	HistoricalAccountStatus13	Target
Balance2	1.00	0.40	0.63	-0.54	0.12	0.66	0.40
HistoricalAccountDetail2	0.40	1.00	0.41	-0.11	-0.24	0.30	0.33
HistoricalAccountStatus3	0.63	0.41	1.00	-0.43	-0.16	0.63	0.38
AccountActivity7	-0.54	-0.11	-0.43	1.00	0.13	-0.32	-0.32
AccountDetail3	0.12	-0.24	-0.16	0.13	1.00	-0.10	-0.18
HistoricalAccountStatus13	0.66	0.30	0.63	-0.32	-0.10	1.00	0.29
Target	0.40	0.33	0.38	-0.32	-0.18	0.29	1.00

Note: This is a snippet of the correlation matrix. Range is from -1 to 1 (Strong negative - Strong positive)



Exploratory Data Analysis

Summary

- 4 out of 5 Categorical columns have >30% missing values
- 2 out of 7 Date columns have >30% missing values
- Occurrence of redundant data due to correlation values
- Need to handle missing and redundant data

Data Cleaning

Initial Cleaning Process

- Convert date columns to date type
- Split date columns into 2 - Year and Month
- Label encode categorical columns

Levels	AccountDetail5	AccountDetail6	AccountStatus1	PrevAccountStatus2
O	121509	-	27793	121509
D	49262	-	14961	49262
X	26360	157607	55663	26360
N	7853	-	1401	7853

Levels	PrevAccountStatus1
I	5065
C	4338
A	2955
E	174
Z	24
B	5
F	2

Data Cleaning

Imputing missing value

- Impute missing values in date columns using other column values (Number of similar values is high enough)
- Impute remaining missing data using interpolation or frequency

Feature Selection

- Used featurewiz to handle the problem of redundant data
- Correlation limit of 0.35
- The number of dimensions reduced from 55 to 20

Note - Cleaning process changes slightly later

Model Development

Initial Development

- Used standard scaler to standardize values
- 70-30 split into training and validation data
- Created 5 models with the following AUCROC scores -
  - Logistic Regression - 71.46%
  - Shallow NN - 71.62%
  - Support Vector Machine - 71.52%
  - Random Forest Classifier - 72.4%
  - Gradient Boosting Classifier - 72.3%
- Created an additional model using a slightly different approach

Model Development

Light Gradient Boosting Machine

- Similar to GBC, more efficient, faster and easy to scale
- Handle large datasets with high accuracy
- >60 hyperparameters, allowing fine-grained control
- LGBM has an in-built hyperparameter that allows missing data to be handled - whenever missing data is encountered, a split is created

Updates to approach

Data Cleaning - No interpolation or frequency based imputation

Feature Selection - featurewiz omitted, all dimensions utilised

Train-Test split - Utilized Stratified K-Fold cross validation

Hyperparameter tuning - Primary focus of preventing overfitting

```
params = {'boosting_type': 'gbdt',
          'n_estimators': 1000,
          'num_leaves': 50,
          'learning_rate': 0.05,
          'colsample_bytree': 0.9,
          'min_child_samples': 2000,
          'max_bins': 500,
          'reg_alpha': 2,
          'objective': 'binary',
          'random_state': 21}
```

Model Evaluation

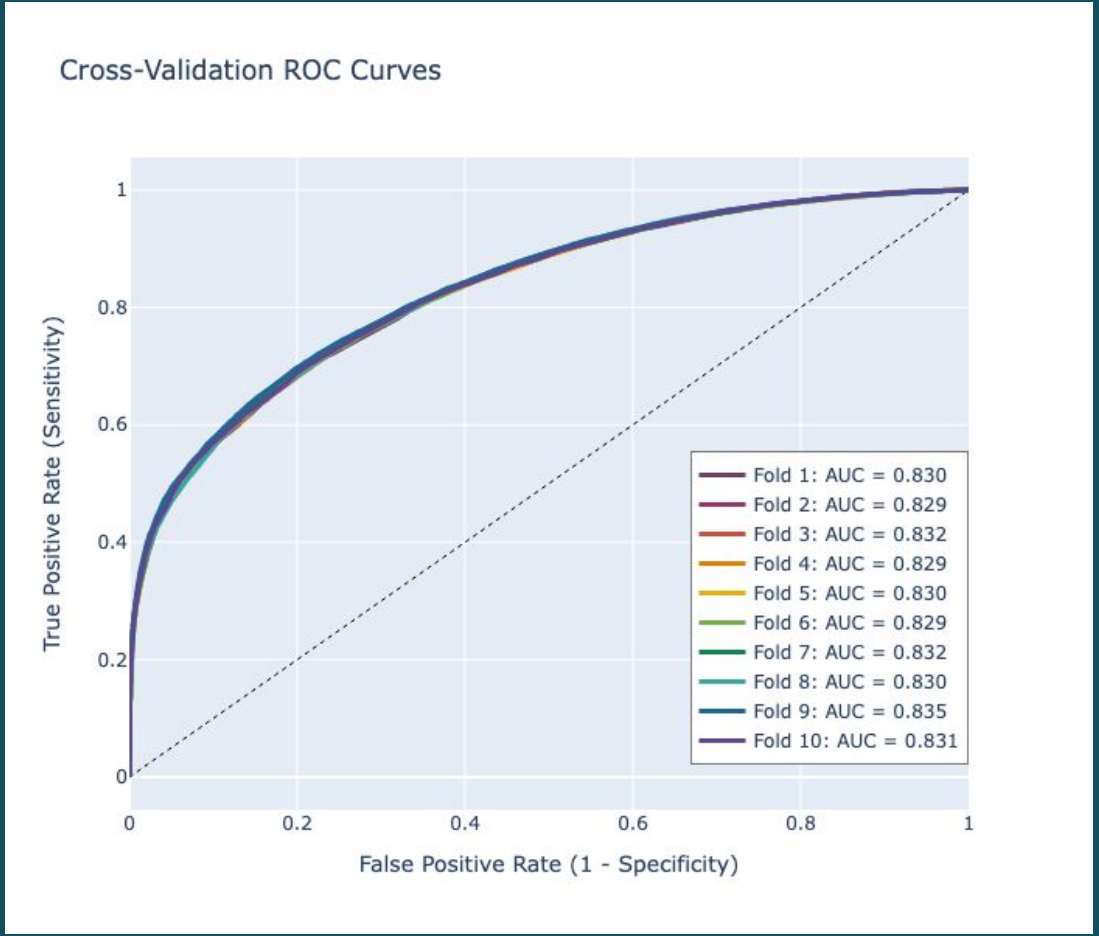
- Validation auc is computed during each round of each fold
- Total of 1000 rounds of training

```
Fold 9
Train shape: (270000, 63), (270000,), Valid shape: (30000, 63), (30000,)

Training until validation scores don't improve for 200 rounds
[500]   training's auc: 0.848715        training's binary_logloss: 0.473757     valid_1's auc: 0.834308 valid_1's binary_logloss: 0.488784
[1000]  training's auc: 0.86273 training's binary_logloss: 0.45663      valid_1's auc: 0.835112 valid_1's binary_logloss: 0.487785
Did not meet early stopping. Best iteration is:
[1000]  training's auc: 0.86273 training's binary_logloss: 0.45663      valid_1's auc: 0.835112 valid_1's binary_logloss: 0.487785
AUC: 0.8351
```

Model Evaluation

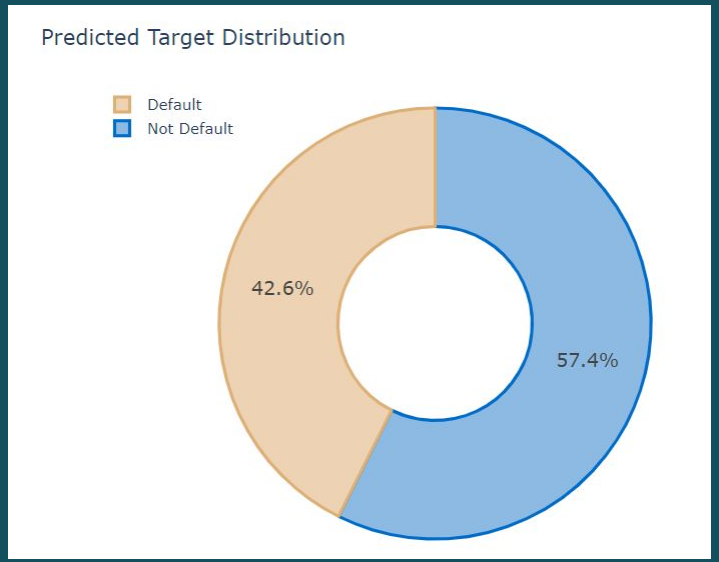
- ROC Curves of best iteration for each fold
- Fold 9 - 83.5%
- True positive rate - TP/(TP+FN)
- False positive rate - FP/(FP+TN)



Prediction

- Model utilised on test data to predict how likely a customer defaults
- Distribution of Default vs. Not Default is 57.4% and 42.6%

	UniqueID	CurrentDate	Prediction
0	32958481	11/1/2018	0.118525
1	33899630	11/1/2018	0.177980
2	28273111	11/1/2018	0.036368
3	34486080	11/1/2018	0.571529
4	31963197	11/1/2018	0.482933



Business Implications

- Identify high risk credit card applicants and make informed decisions
- Minimize the risk of credit card defaults and reduce the losses incurred by the organization
- Ability to utilize data “as-is” without the need for much data-cleaning, saving time and resources

Questions

1. **Exploratory Data Analysis**

   Understanding what we are working with

2. **Data Cleaning**

   Performing transformations and feature engineering to prepare the data

3. **Model Development**

   Utilizing data for training, culminating in the selection of the best performing model

4. **Model Evaluation**

   Measuring model effectiveness on validation data

5. **Prediction**

   Predicting targets of test data using the final model

Appendix

- Code
- Correlation wiki
- LGBM vs. XGBoost
- AUCROC

Thank you!