synchrony

Statistics Datathon Group Project

Team YYZZ

Yihao Chen
Yi Jin
Zheer Wang
Ziqi Xu

Mindset

Model One — Linear Regression Model (Using training_data.csv)

- By linear regression, calculate amount of charge off
- Using charge off amount from previous data, calculate 01/2020 charge off amount

Model Two — Linear Regression Model (Using Macro_data.csv)

- By least square matrix, calculate macro coefficients
- Using macro coefficients, predict charge offs during 02/2020 – 01/2021

Model One — Linear Regression Model

	Unnamed: 0	charge_off
6	open_closed_flag	-0.1303438961338125
8	nbr_mths_due	0.2118169491894203
18	active	-0.9987885430561247
19	charge_off	1.0
20	charge_off_aged	0.9405717785490472
21	charge_off_bk	0.3387755307539496
23	principal_amt_chrg_off	0.606442253032329
24	total_writeoff_amt	0.6870568001060232
25	fee_chg_off_reversal_amt	0.6839247790229768
30	aged_writeoff_amt	0.6325975813642509
31	bankruptcy_writeoff_amt	0.2402019810391827
32	fc_reversals	0.5299494138890559
33	fee_reversals	0.8161220582798733
35	other_writeoff_amt	0.2012625023625118
37	recovery_amt	0.1555765073795357
38	writeoff_type_bko	0.3387755307539483
41	writeoff_type_deceased	0.2200347469497959
43	writeoff_type_aged	0.8945252384076223
44	writeoff_type_settlement	0.1877813543513465
47	writeoff_type_null	-0.9960707451368722

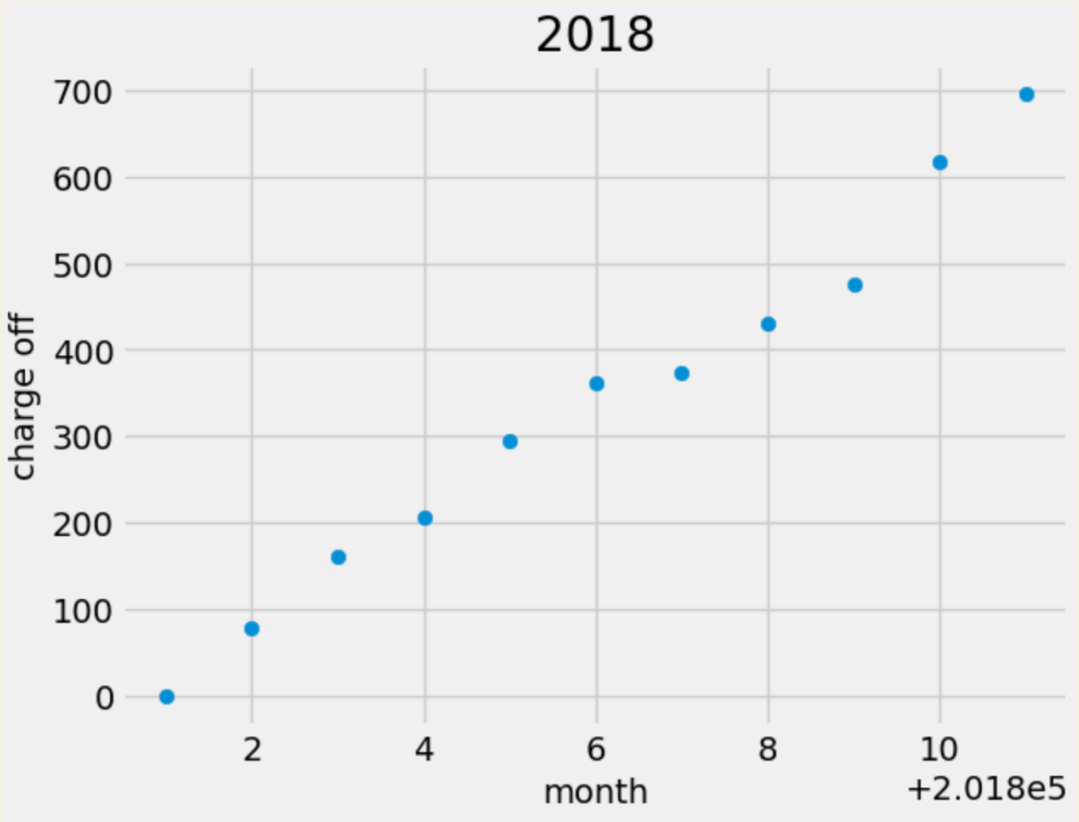
High correlation

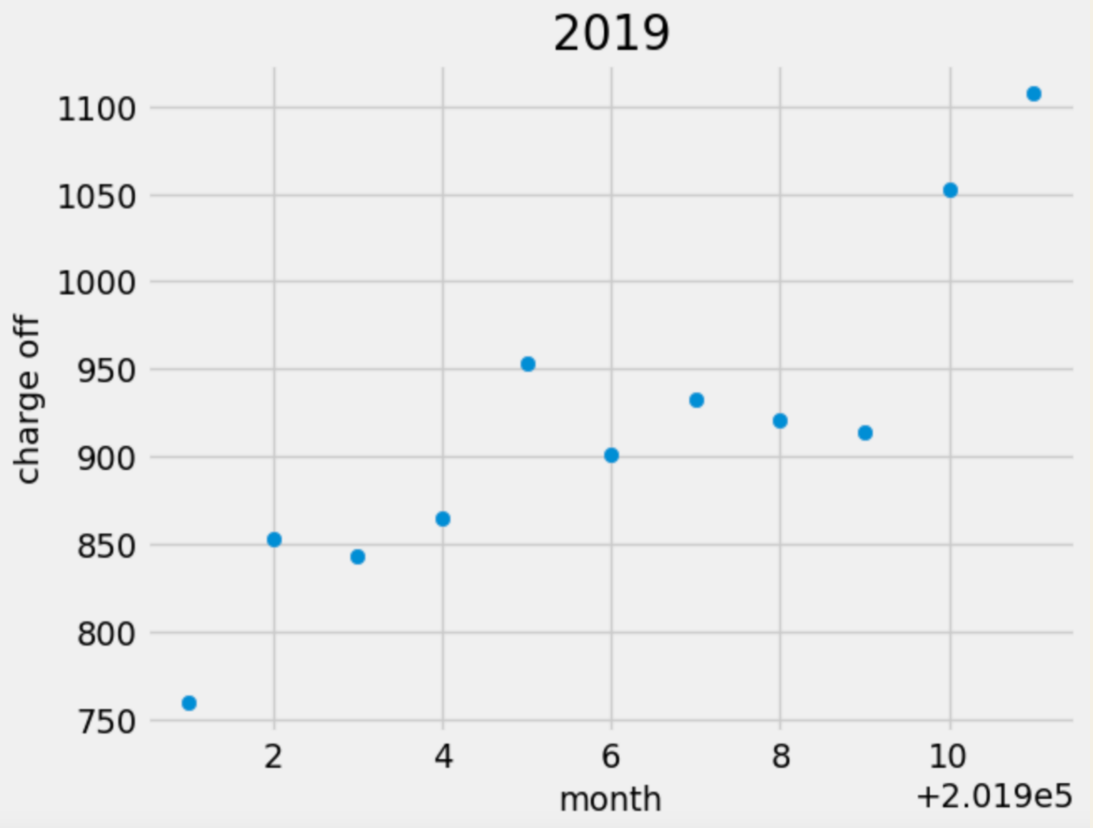
By calculating correlation coefficient, we find “open_closed_flag”, “active”, “fc reversals” and “fee reversals” have high correlation with charge offs.

Why we don’t choose the parameter below which also has high correlation, it is because without the existence of charge offs, these variables won’t exist. Hence, they should not be placed as independent variables in our model.

Model One — Linear Regression Model

The relationship of months and charge offs in 2018 and 2019





Model One — Linear Regression Model

According to Jin et al. (2021), We choose Linear Regression Model to predict charge offs.
The formula we set:

$$CO = a_1 \cdot ST + a_2 \cdot CLA + a_3 \cdot AC + a_4 \cdot FA + a_5 \cdot ED + a_6 \cdot OC + a_7 \cdot FC + a_8 \cdot FEE + \sum_{i=2}^{8} a_{9i} \cdot DB_i$$

CO = charge off
ST = stmt balance
CLA = credit limit amount
AC = active
FA = financial active
ED = ever delinquent flg
OC = open closed flag
FC = fc reversals
FEE = fee reversals
DB = due balance

In addition to “open_closed_flag”, “active”, “fc reversals” and “fee reversals” these two parameters we choose, the choices of other parameters are based on empirical experiences. For example, higher due balance means higher amount the borrower owe, which increases the risk of charge offs.

Data cleaning: Large sample (5 million), so outliers may not affect the result of linear regression.

Model One — Linear Regression Model

How do we get the linear regression model:

Aggregate data by time,

```
timed_df = newdf.groupby("mth_code").agg(sum)
```

Fitting aggregated dataset into linear regression model from sklearn package

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```

Python

```
dep1 = timed_df["charge_off"]
```

Python

```
ind1 = timed_df[["stmt_balance",  "credit_limit_amt", "active", "financial_active", "ever_delinquent_flg", "open_closed_flag",
                 "fc_reversals", "fee_reversals",
                 'due_balance_2', 'due_balance_3', 'due_balance_4', 'due_balance_5', 'due_balance_6', 'due_balance_7', 'due_balance_8']]
```

Python

```
model1  = model.fit(ind1, dep1)
```

Python

Model One — Linear Regression Model

The coefficients for each parameter we get:

```
Intercept:
 -21.64636761362783
Coefficients:
 [-7.76585850e-07  4.40716803e-06  2.28652933e-02  1.07591009e-02
 -1.89951173e-03 -4.85448441e-02  5.61318862e-04  2.63582414e-03
 -5.04249166e-06  9.87193919e-06 -5.60422381e-05  1.29330039e-04
 -1.57923429e-05 -4.29948791e-05 -6.81122069e-03]
```

$$CO = a_1 \cdot ST + a_2 \cdot CLA + a_3 \cdot AC + a_4 \cdot FA + a_5 \cdot ED + a_6 \cdot OC + a_7 \cdot FC + a_8 \cdot FEE + \sum_{i=2}^{8} a_{9i} \cdot DB_i$$

CO = charge off
ST = stmt balance
CLA = credit limit amount
AC = active
FA = financial active
ED = ever delinquent flg
OC = open closed flag
FC = fc reversals
FEE = fee reversals
DB = due balance

Model One — Linear Regression Model

RMSE test equals 4.81- impressive

```
timed_df["charge_off_predict"] = model1.predict(ind1)
```

Python

+ Code + Markdown

```
import numpy
timed_df["residual"] = abs(timed_df["charge_off_predict"]-timed_df["charge_off"])
```

Python

```
#RMSE
timed_df['residual'].std()
```

Python

```
4.814850469322727
```

Using model one and data from forecast_starting.cvs to calculate charge-off for 01/2020 , we get 14.92

```
timed_df_forecast["charge_off_predict"] = model1.predict(timed_df_forecast[["stmt_balance",  "credit_limit_amt", "active",
                                                                            "financial_active", "ever_delinquent_flg", "open_closed_flag",
                                                                            "fc_reversals", "fee_reversals",'due_balance_2', 'due_balance_3',
                                                                            'due_balance_4', 'due_balance_5', 'due_balance_6', 'due_balance_7',
                                                                            'due_balance_8']])
timed_df_forecast["charge_off_predict"]
```

Python

```
0    14.920852
Name: charge_off_predict, dtype: float64
```

Model Two — Linear Regression Model

According to NCUA(2003), we choose 2 parameters to estimate charge off from 02/2020 to 01/2021, including unemployment and consumer credit.

```
df_macro_sel = df_macro[["Mnemonic","M_FLBR.IUSA", "M_FCCALLQ.IUSA"]]
df_macro_sel
```

✓ 0.0s

	Mnemonic	M_FLBR.IUSA	M_FCCALLQ.IUSA
0	Description	Baseline Scenario (October 2022): Household Su...	Baseline Scenario (June 2020): Consumer Credit...
1	Name	Household Survey: Unemployment Rate, (%, SA)	Consumer Credit: Total Outstanding, (Bil. USD,...
2	Source	U.S. Bureau of Labor Statistics (BLS): Current...	U.S. Board of Governors of the Federal Reserve...
3	Native Frequency	MONTHLY	MONTHLY
4	Geography	United States	United States
...	...	...	...
434	8/31/2035	3.937188576	NaN
435	9/30/2035	3.936435187	NaN
436	10/31/2035	3.936477376	NaN
437	11/30/2035	3.937037114	NaN
438	12/31/2035	3.937853289	NaN

Model Two — Linear Regression Model

By using least-square, the formula we used to calculate the coefficients is

Macro index matrix – macro_data.csv

Delta matrix: delta = CO-current – CO-previous_month

$$R = (A^T A)^{-1} A^T b$$

R = coefficient matrix
A = macro index matrix
b = delta matrix with $b_i = CO_i - CO_{i-1}$
i means time duration

Model Two — Linear Regression Model

$$CO = r_1 \cdot UR + r_2 \cdot CC$$

$\mathrm{UR} = \text{unemployment rate}$
$\mathrm{CC} = \text{consumer credit}$

By least square, we get two coefficients.

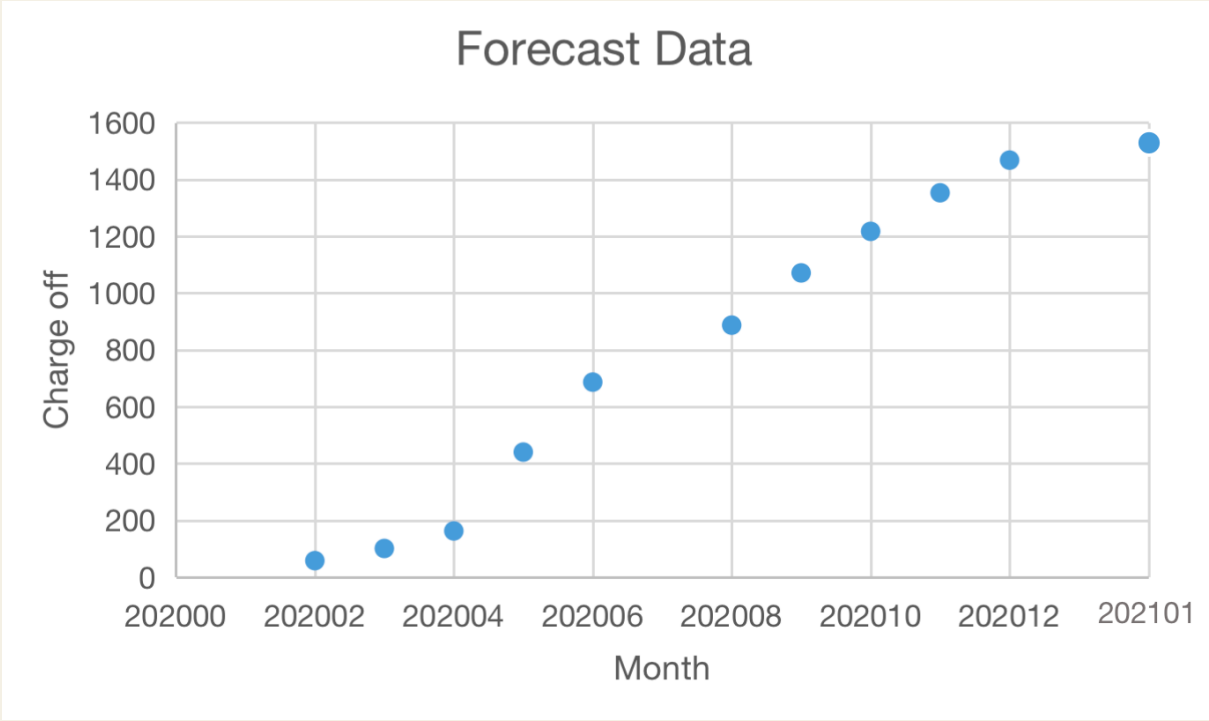
```
r = np.dot(np.dot(r, b.transpose()), deltamatrix)
r
```

✓ 0.0s

```
array([[ 2.09062051e+01],
       [-7.19262285e-03]])
```

Final Results — the Prediction

Month	accounts_charge_off
202002	57.92600303
202003	100.79114894
202004	162.55481147
202005	440.14708954
202006	686.51331905
202008	887.1552218
202009	1071.1547776
202010	1217.48325607
202011	1353.32020066
202012	1468.29731464
202101	1579.17371369



Reference List

- *Loan charge-off guidance*. NCUA. (2003, January 1). Retrieved March 26, 2023, from https: //ncua.gov/regulation-supervision/letters-credit-unions-other-guidance/loan-charge-guidance

- Justin Y. Jin, Mary L.Z. Ma, Victor Song, Mengyang Guo, Banks' loan charge-offs and macro-level risk, Journal of Behavioral and Experimental Finance, Volume 32, 2021, 100573, ISS N 2214-6350, https://doi.org/10.1016/j.jbef.2021.100573.

Thank you!

Team YYZZ

Yihao Chen
Yi Jin
Zheer Wang
Ziqi Xu