CHARGE-OFF IN COVID19

DATATHON23

by team raining (Junyeong, Gayeon, Yuan, Jian)

OUR
MAIN GOAL

Using historical data and macroeconomic data, predict the accurate number of "charge-offs" in each month during the covid19 period

OUR APPROACH

Preprocessing

- Exploratory Data Analysis
- Data cleaning and aggregation

Training

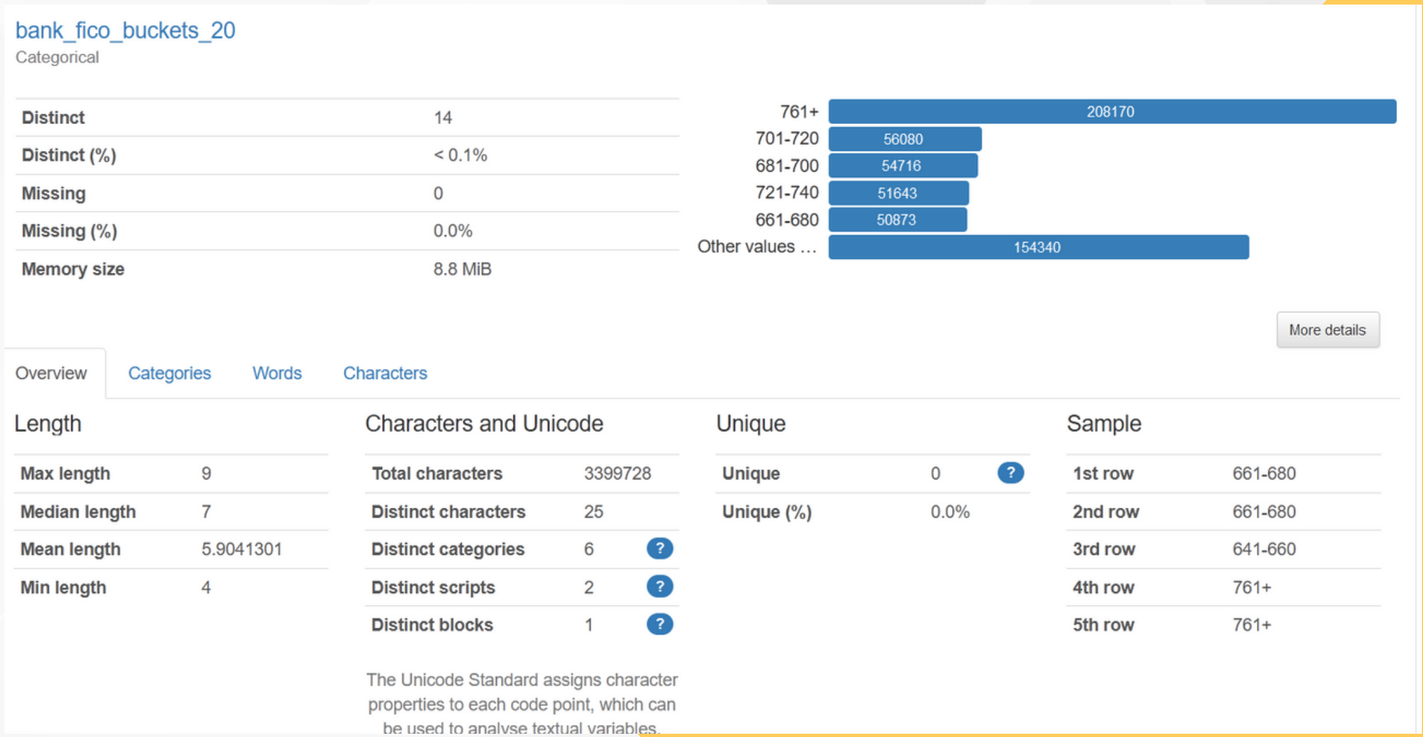
- Model training and selection

Testing on unseen data

- Forecasting

EXPLORATORY DATA ANALYSIS

- Conducted pandas profiling on the train data
- Obtained a high-level understanding of each variable and the values
- Correlation matrix to identify important features to predict charge_off



bank_fico_buckets_20

Categorical

Distinct	14
Distinct (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	8.8 MiB

More details

Overview | Categories | Words | Characters

Length

Max length	9
Median length	7
Mean length	5.9041301
Min length	4

Characters and Unicode

Total characters	3399728
Distinct characters	25
Distinct categories	6
Distinct scripts	2
Distinct blocks	1

The Unicode Standard assigns character properties to each code point, which can be used to analyse textual variables.

Unique

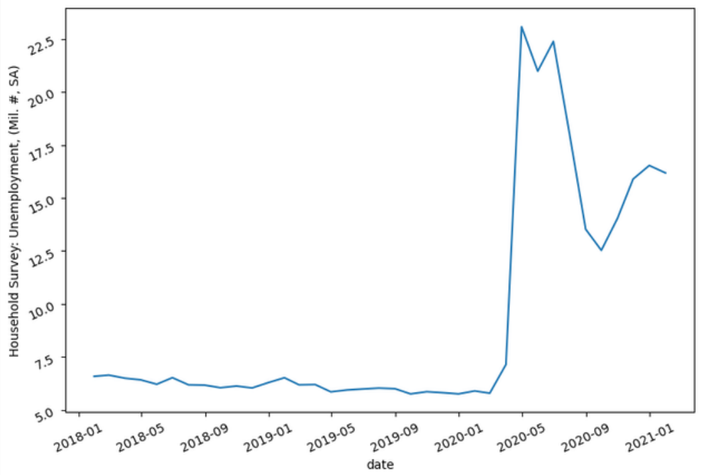
Unique	0
Unique (%)	0.0%

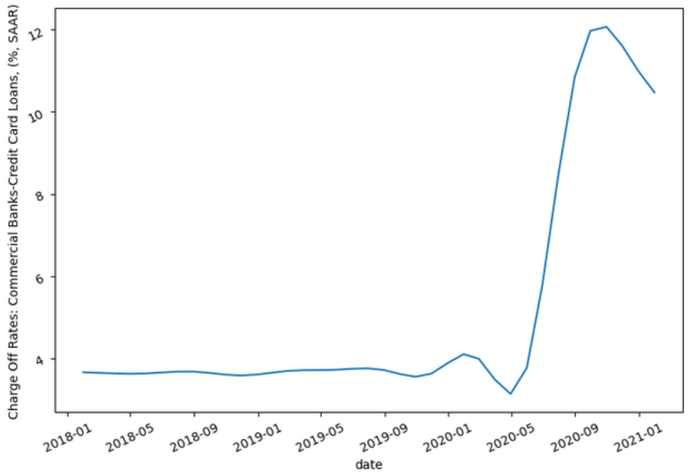
Sample

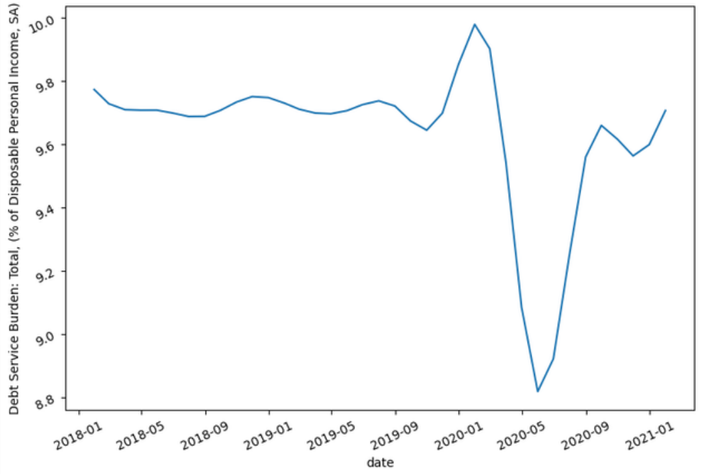
1st row	661-680
2nd row	661-680
3rd row	641-660
4th row	761+
5th row	761+

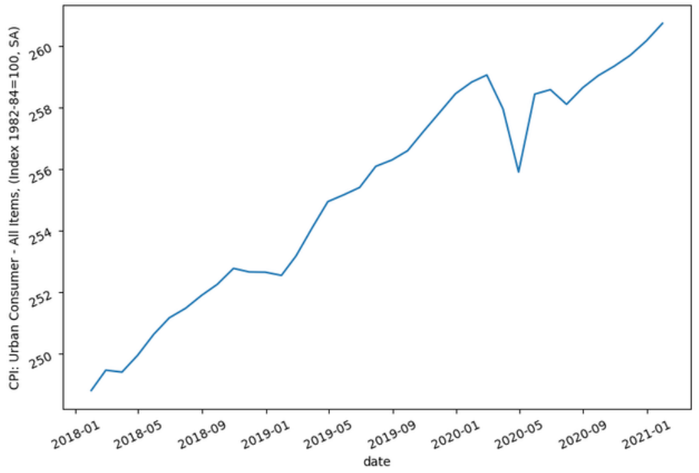
EXPLORATORY DATA ANALYSIS (EDA)

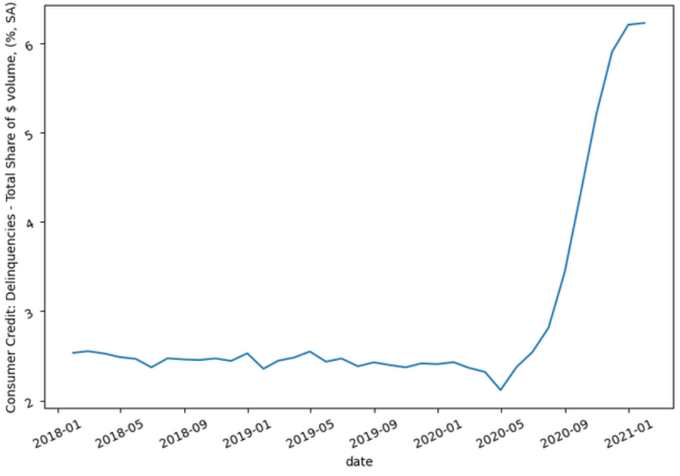
- Macroeconomic data shows drastic change in various indicators during Covid period
- Machine learning models might struggle to forecast during this period of unforeseen circumstances

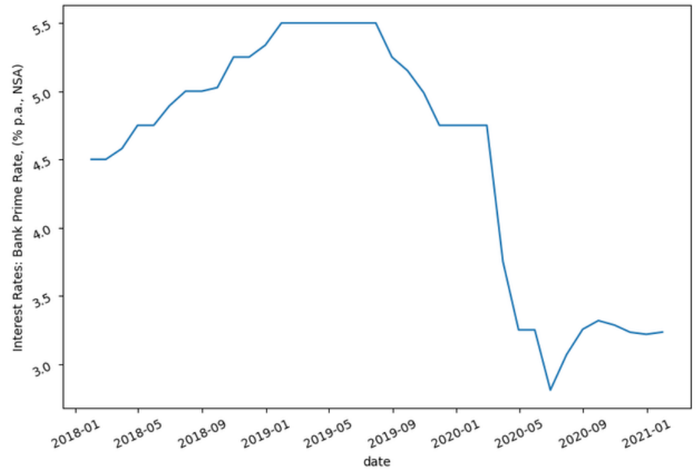












FEATURE ENGINEERING AND SELECTION

Feature	Description
Bank Fico	summarizes your credit risk based on your credit file at one of the three major consumer bureaus at a particular point in time.
Delinquency	A loan becomes delinquent when it does not provide the required payment by the contractual date
balance_decrease	current balance statement - previous balance
month_diff	The difference between the snapshot date and the mth_code

Feature	Description
Debt Service Burden	cost of interest payments on debt
Consumer Price Index	average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.
Bank Prime Rate	The prime rate is the interest rate that commercial banks charge their most creditworthy customers
Charge Off Rates: Commercial Banks-Credit Card Loans	Proportion of credit card loans charged off

DATA AGGREGATION

- Train data contains many irrelevant features
- In order to forecast for the future, aggregation by snapshot and mth_code is done
- Proportion of charge_off accounts is analyzed by month to predict the forecast data
- Train data is joined with macroeconomic data by matching the mth_code in train with Mnemonic in macroeconomic data

MODEL TRAINING AND SELECTION

Non-aggregated data

- Large dataset (~ 6 million data points) making it difficult to work with
- Utilized various models such as logistic regression, random forest classifier and support vector classifier
- Limited success using non-aggregate data due to low accuracy, recall and precision

MODEL TRAINING AND SELECTION

Aggregated data

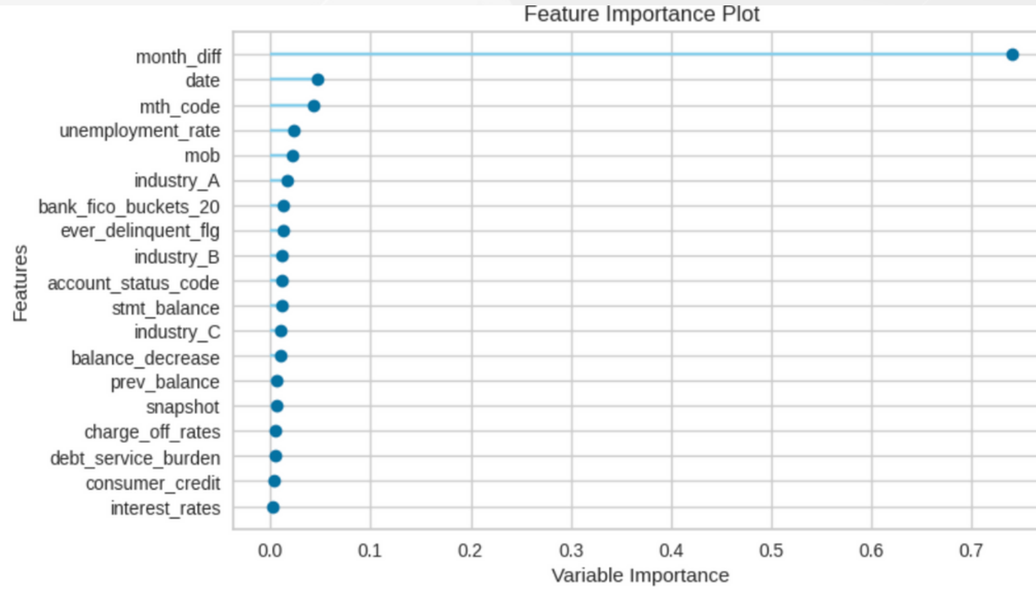
- Used aggregated data as a train data to solve the problem of too large data size
- joined with 2 version of macro data (using manually selected features only & using PCA vectors)
- trained various ML models for each version of data and tuned the hyper parameters
- finalized the models and compared them

MODEL PERFORMANCE

Feature Importance plot indicates that the further the mth_code is from the snapshot, the more likely an account is to be charged off

Utilized autoML to test out various models and picked the one with the lowest RMSE

Determined that the random forest regressor has the best performance out of all of the models




```
best = compare_models(sort='RMSE',n_select=4)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	0.0003	0.0000	0.0004	0.8154	0.0004	0.1157	0.3690
gbr	Gradient Boosting Regressor	0.0003	0.0000	0.0004	0.7880	0.0004	0.1254	0.4310
xgboost	Extreme Gradient Boosting	0.0003	0.0000	0.0004	0.8125	0.0004	0.1157	0.2600
ada	AdaBoost Regressor	0.0003	0.0000	0.0004	0.8114	0.0004	0.1218	0.2640
dt	Decision Tree Regressor	0.0004	0.0000	0.0005	0.6949	0.0005	0.1502	0.3350
et	Extra Trees Regressor	0.0004	0.0000	0.0005	0.7549	0.0005	0.1138	0.3100
lightgbm	Light Gradient Boosting Machine	0.0005	0.0000	0.0007	0.4509	0.0007	0.1596	0.2540
omp	Orthogonal Matching Pursuit	0.0007	0.0000	0.0009	0.0234	0.0009	0.1612	0.2360
lr	Linear Regression	0.0007	0.0000	0.0009	0.0038	0.0009	0.1840	0.4680
br	Bayesian Ridge	0.0007	0.0000	0.0009	0.0602	0.0009	0.1754	0.2820
knn	K Neighbors Regressor	0.0007	0.0000	0.0009	0.0006	0.0009	0.1724	0.2490
ridge	Ridge Regression	0.0007	0.0000	0.0009	0.0570	0.0009	0.1766	0.2370
lasso	Lasso Regression	0.0007	0.0000	0.0010	-0.0393	0.0010	0.1517	0.2440
en	Elastic Net	0.0007	0.0000	0.0010	-0.0393	0.0010	0.1517	0.2790
llar	Lasso Least Angle Regression	0.0007	0.0000	0.0010	-0.0393	0.0010	0.1517	0.2430
huber	Huber Regressor	0.0006	0.0000	0.0010	-0.0329	0.0010	0.1473	0.2520
dummy	Dummy Regressor	0.0007	0.0000	0.0010	-0.0393	0.0010	0.1517	0.2650
par	Passive Aggressive Regressor	0.0027	0.0000	0.0029	-8.9987	0.0029	1.0000	0.3730
lar	Least Angle Regression	12.7295	821.1753	14.9535	-737492983.2965	1.2170	4966.1808	0.3680

MODEL FORECASTING

Duplicated the 19866 accounts for all 12 months

Changed the mth_code and corressponding month_diff to create data to forecast from 2020 Feb to 2021 Jan

Aggregated each month and joined the macroeconomic data to the corresponding month

Used the model to forecast the number of charge_off to happen in a month within the year

MODEL FORECASTING

	Month	accounts_charged_off
0	202002	2.291248
1	202003	60.780062
2	202004	62.678149
3	202005	62.695475
4	202006	62.993995
5	202008	63.256278
6	202009	63.265189
7	202010	63.265189
8	202011	62.269069
9	202012	62.414680
10	202101	62.350368

Total proportion of charged off accounts is higher than historical average

OUR REFLECTIONS

The real world is messy

Importance of distributing workload

Domain knowledge is key to make good feature selection

Thank you!