



Multi-class text classification of Chase complaints

Sunil Kafle

January 2022

1. Background

- Consumer complaints are an expression of dissatisfaction to their associated institution
- Total 490,530 complaints received by CFPB in the last year
- Complaints provides an opportunity to improve the business
- Finance sector highly consumer centric



2. Introduction


- Financial service sector is an important sector
- Filing consumer complaint in Consumer Finance Protection Bureau(CFPB)
- Complaints has been categorized in different classes
- Can we predict the complaint narrative to the specific class offered by the financial institution?

3. Table of contents

1. Background
2. Introduction
3. Executive summary
4. Data Exploration
5. Results
6. Conclusions
7. Recommendations
8. Reference
9. Appendix



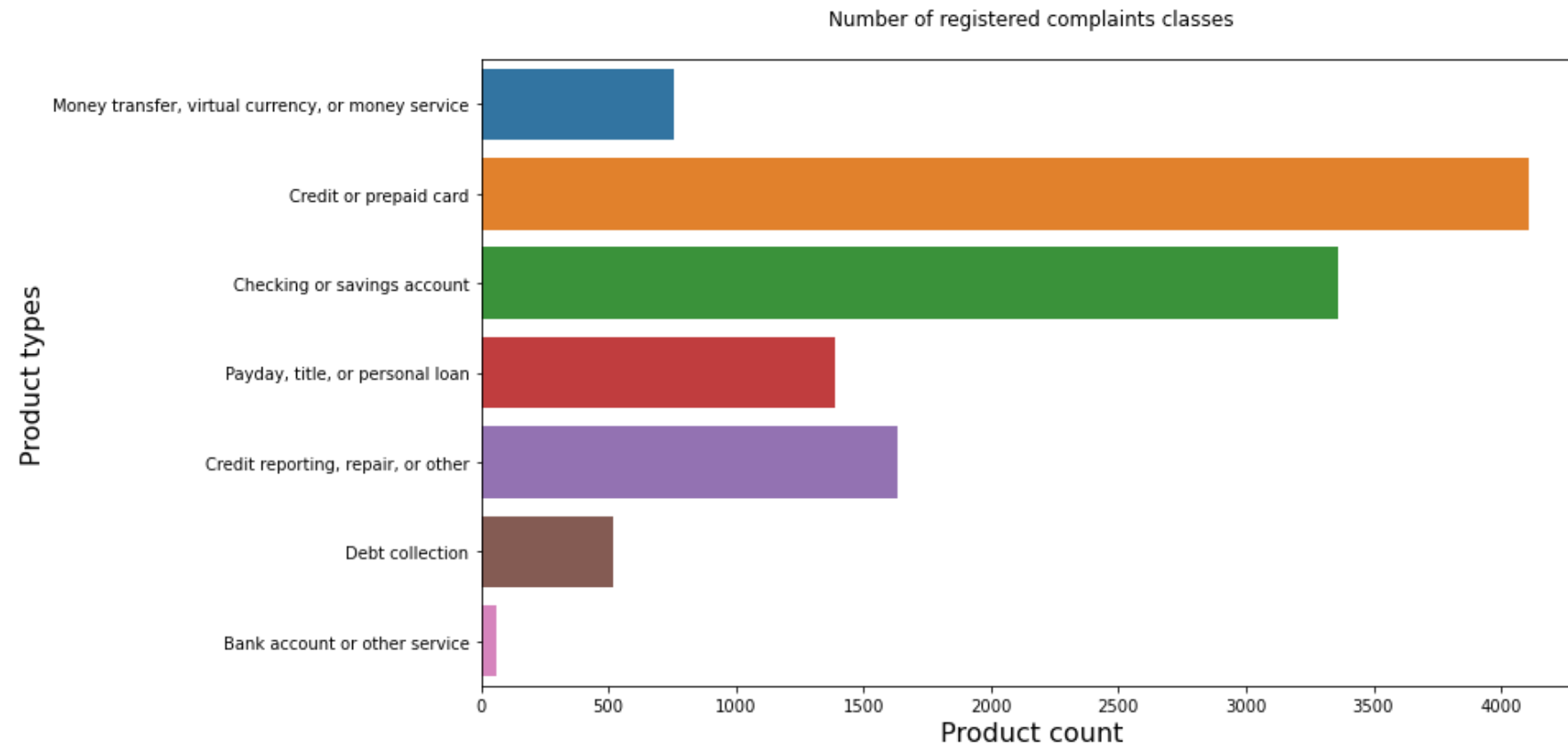
4. Executive summary

- Different classification algorithms have been used to classify text
 - Random Forest, Logistic Regression, Multinomial Naïve Bayes, Linear SVC and XGBoost
 - Similar classes has been merged to make the results more comparable
 - Linear SVC came out as the best model with a ROC_AUC score of 95 and an Accuracy of 83%.
- 
- A large yellow triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

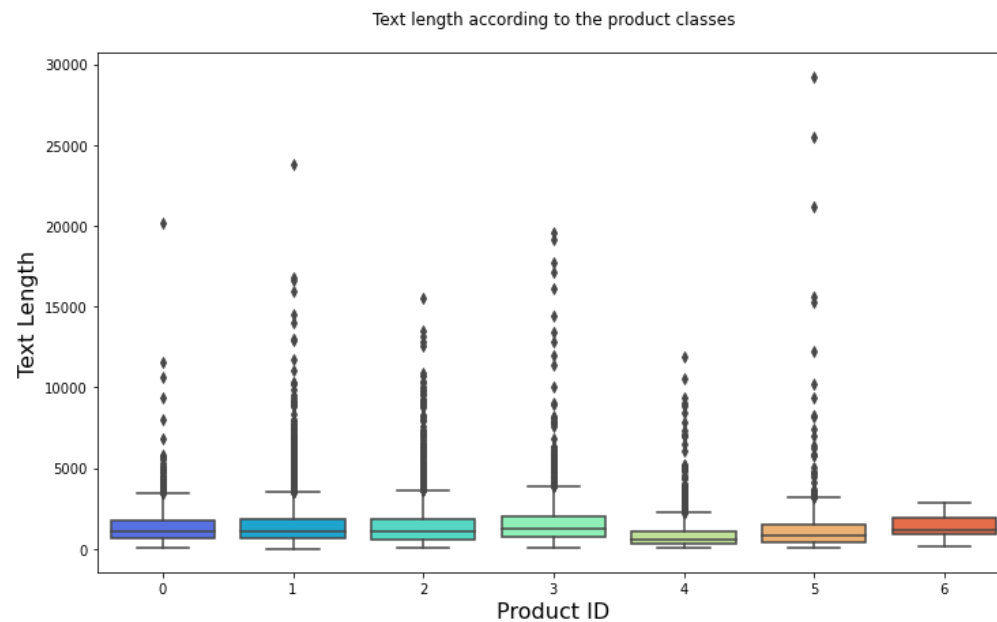
5. Data exploration

- The original dataset consist of 1,048,575 observations and 18 features.
- Chase consists of 13791 (25622 with missing) observations after removing all the observations with missing values
- The product feature consists of 15 classes

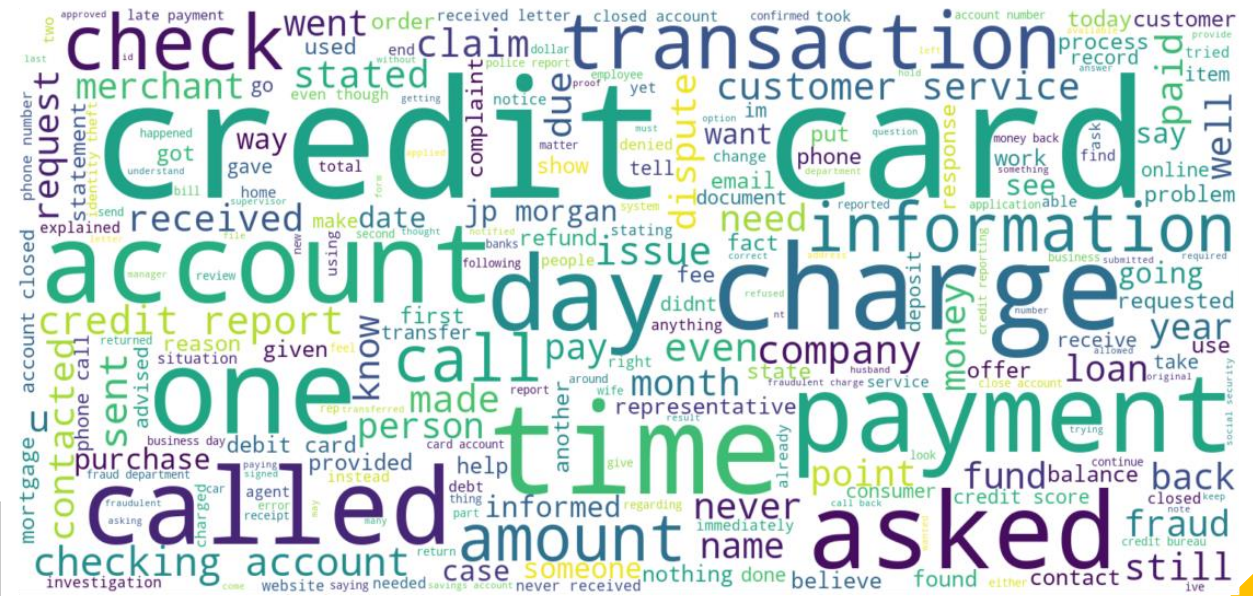
5. Data exploration



5. Data exploration



Top words in the whole dataset



6. Results

Best models:
Linear SVC &
Logistic Regression

1 = Checking or
savings account

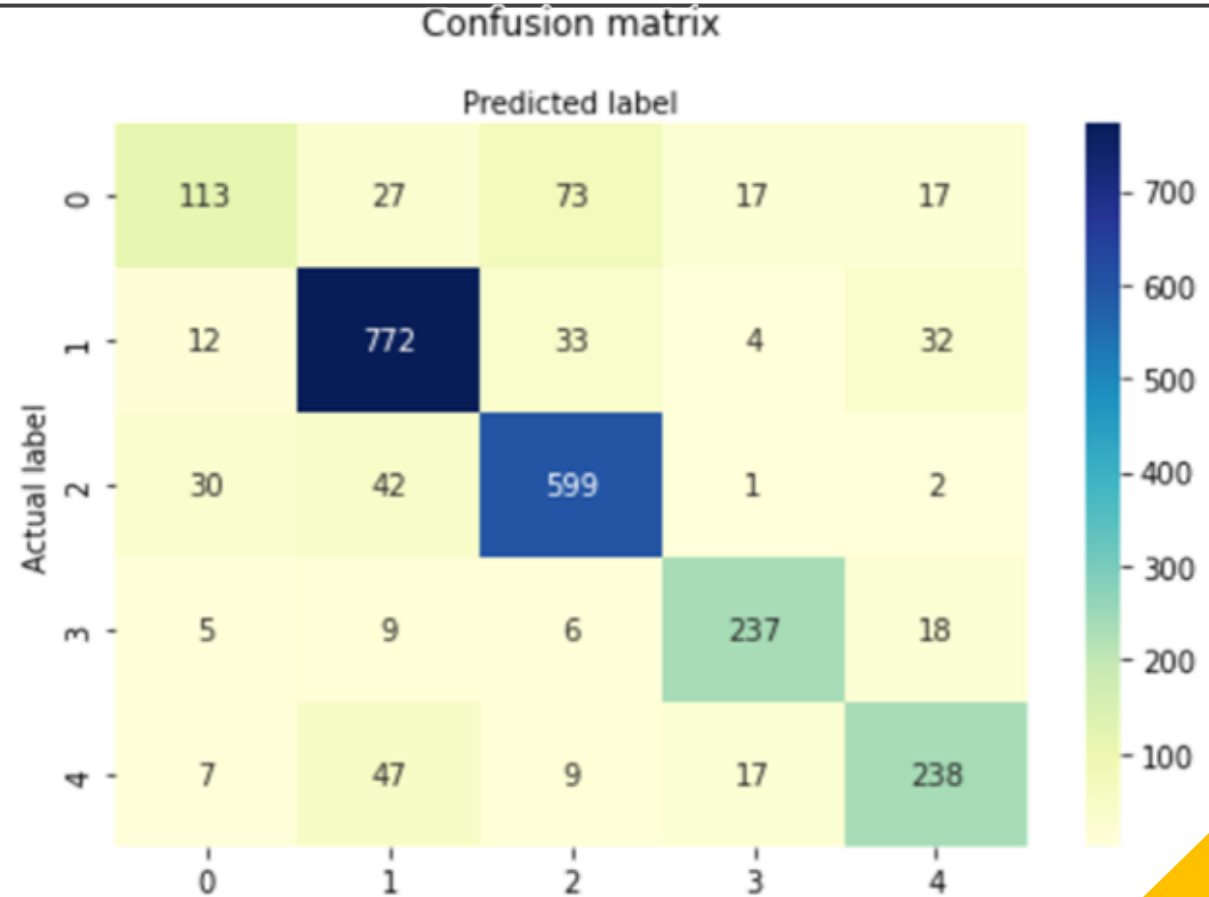
2 = Credit reporting,
repair, or other

3 = Payday, title, or
personal loan

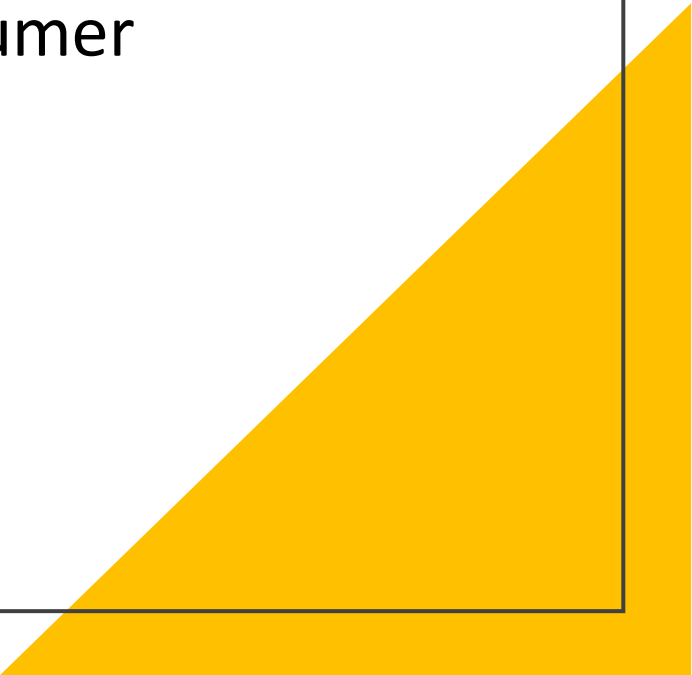
Model	Accuracy	ROC_AUC score	F1 score(1)	F1 score(2)	F1 score(3)
Logistic Regression(Base)	0.82	0.95	0.88	0.86	0.86
Logistic Regression (Tuned)	0.82	0.95	0.88	0.86	0.85
Linear SVC(Base)	0.83	0.95	0.88	0.86	0.86
Linear SVC (Tuned)	0.82	0.95	0.88	0.85	0.86

6. Results


Confusion matrix of
Linear SVC



7. Limitations

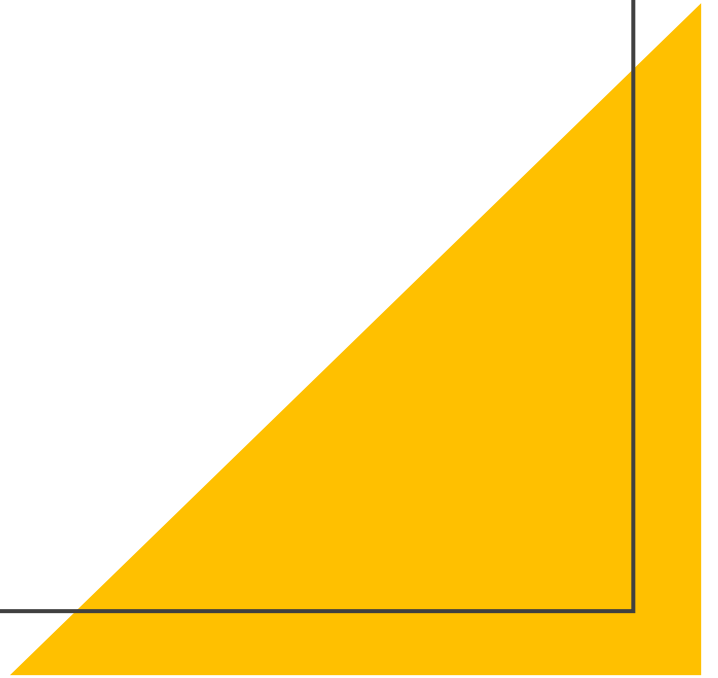
- After data cleaning only 11,831 observation was used in training the model
 - No better way to impute the empty entry in consumer narrative
 - Resource limitation for robust modeling
- 
- A large yellow right-angled triangle is positioned in the bottom right corner of the slide, extending from the bottom edge and the right edge towards the center.

8. Conclusions

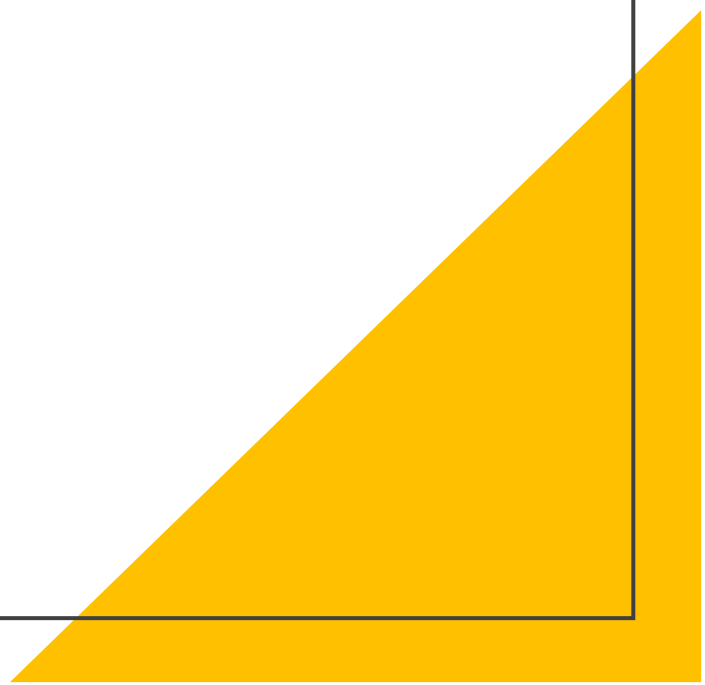
- Used five different models to find the best one
 - Linear SVC performed best compared to other
 - For better interpretation and more control Logistics
 - If concerns about overfitting use SVM
- 
- A large yellow right-angled triangle is positioned in the bottom right corner of the slide, pointing towards the top right.

9. Recommendations

- Training using a large dataset
- Try with the nested modeling technique
- Exploring deep learning models



Thank you

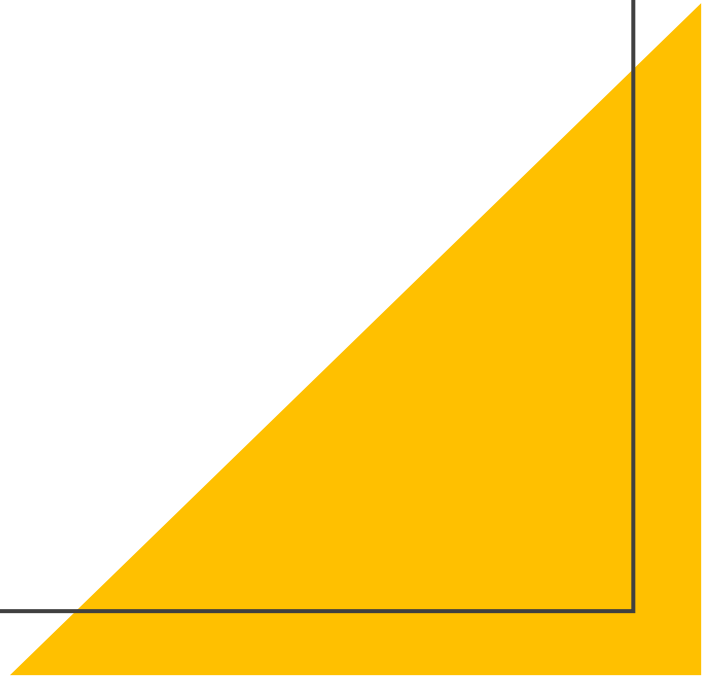


10. References

<http://norma.ncirl.ie/4227/>

<https://www.consumerfinance.gov/>

<https://esource.dbs.ie/handle/10788/4224>

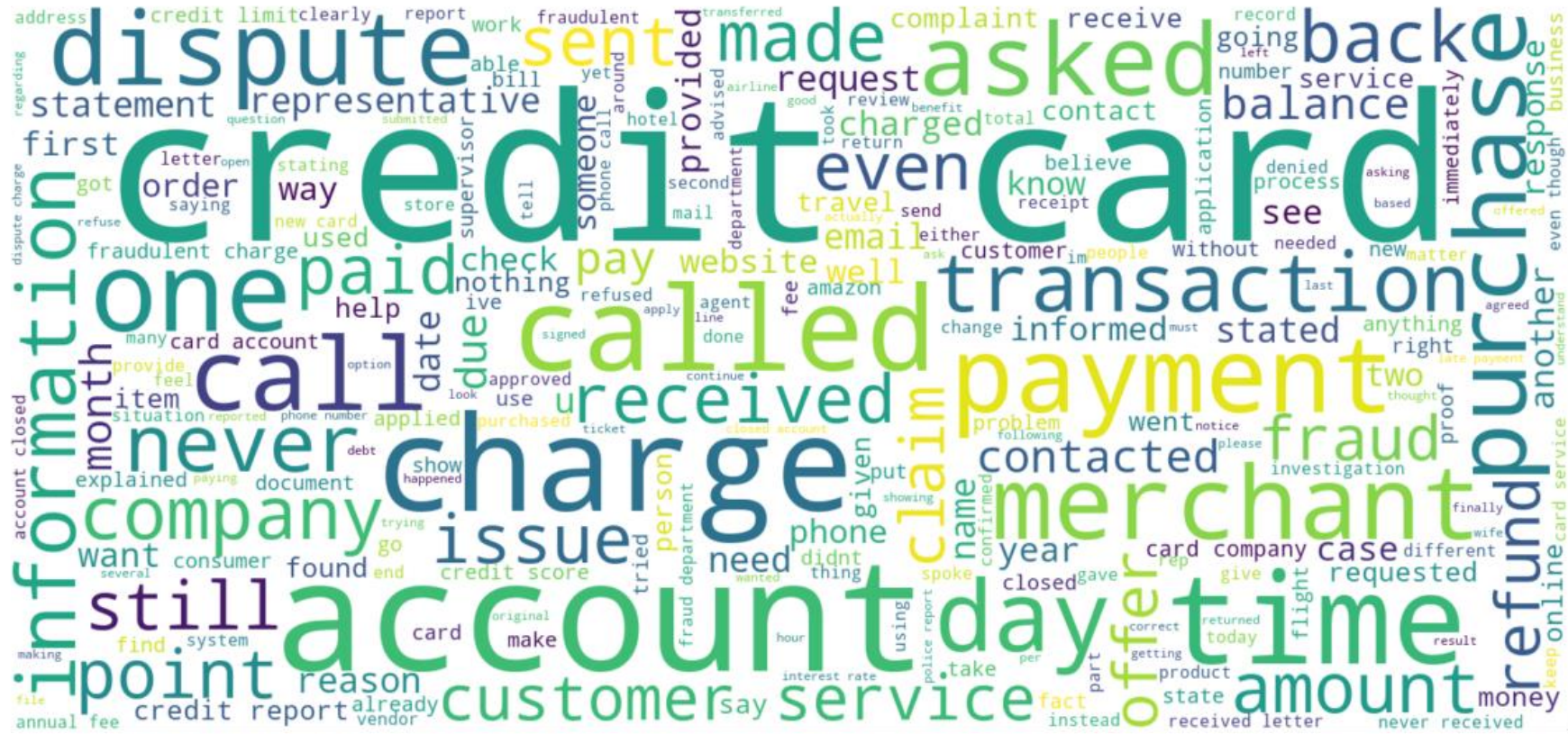


11. Appendix

Product(Complaint Channel)	Number of complaints
1. Credit card or prepaid card	4023
2. Checking or savings account	3364
3. Credit reporting, credit repair services, or other personal consumer reports	1631
4. Mortgage	1022
5. Money transfer, virtual currency, or money service	752
6. Debt collection	517
7. Vehicle loan or lease	316
8. Credit card	88
9. Bank account or service	58
10. Payday loan, title loan, or personal loan	38
11. Student loan	11
12. Credit reporting	5
13. Consumer Loan	3
14. Money transfers	2
15. Other financial service	1

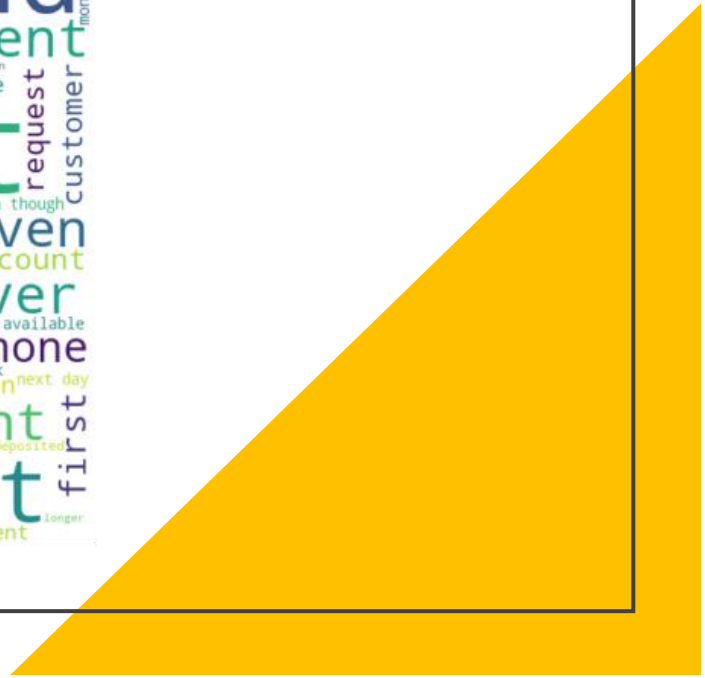
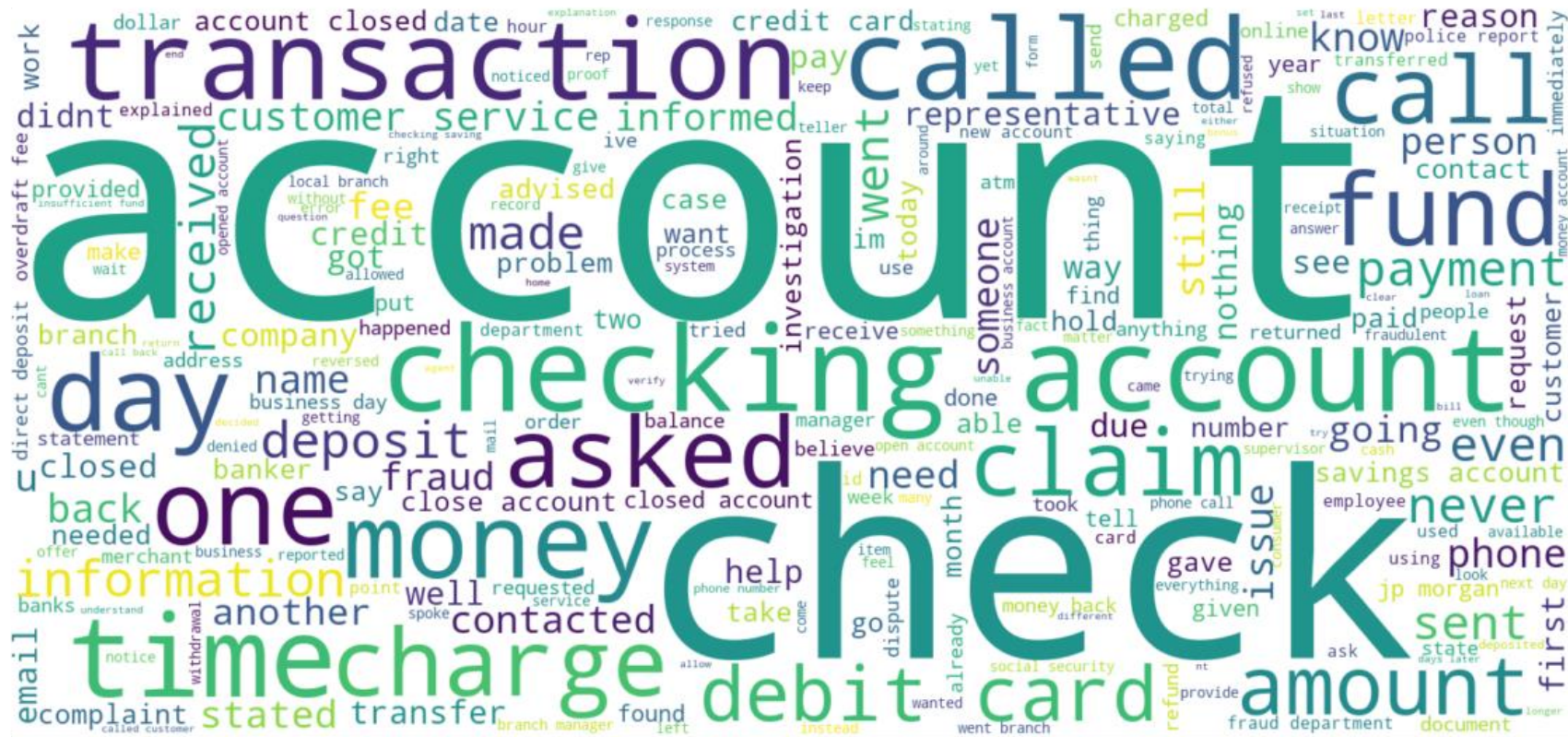
11. Appendix

Top words in the Credit or prepaid card class



11. Appendix

Top words in the Checking or savings account



11. Appendix

0 = Credit or prepaid card

1 = Checking or savings account

2 = Credit reporting, repair, or other

3 = Payday, title, or personal loan

4 = Bank account, debt, money or other service

Model	Accuracy	ROC_AUC score	F1 score(1)	F1 score(2)	F1 score(3)
Random Forest(Base)	0.78	0.92	0.84	0.81	0.82
Random Forest(Tuned)	0.74	0.94	0.79	0.81	0.77
Logistic Regression(Base)	0.82	0.95	0.88	0.86	0.86
Logistic Regression (Tuned)	0.82	0.95	0.88	0.86	0.85
Naive Bayes(Base)	0.71	0.93	0.78	0.81	0.62
Naive Bayes (Tuned)	0.71	0.93	0.78	0.81	0.62
Linear SVC(Base)	0.83	0.95	0.88	0.86	0.86
Linear SVC (Tuned)	0.82	0.95	0.88	0.85	0.86
XGBoost(Base)	0.80	0.94	0.85	0.83	0.84
XGBoost(Tuned)	0.77	0.91	0.83	0.80	0.81

1 1. Appendix

Confusion matrix
and classification
report of Logistic
Regression(base)

	precision	recall	f1-score	support
0	0.74	0.38	0.51	247
1	0.84	0.91	0.88	853
2	0.82	0.91	0.86	674
3	0.86	0.86	0.86	275
4	0.74	0.70	0.72	318
accuracy			0.82	2367
macro avg	0.80	0.75	0.77	2367
weighted avg	0.81	0.82	0.81	2367

roc_auc_score: 0.950938

