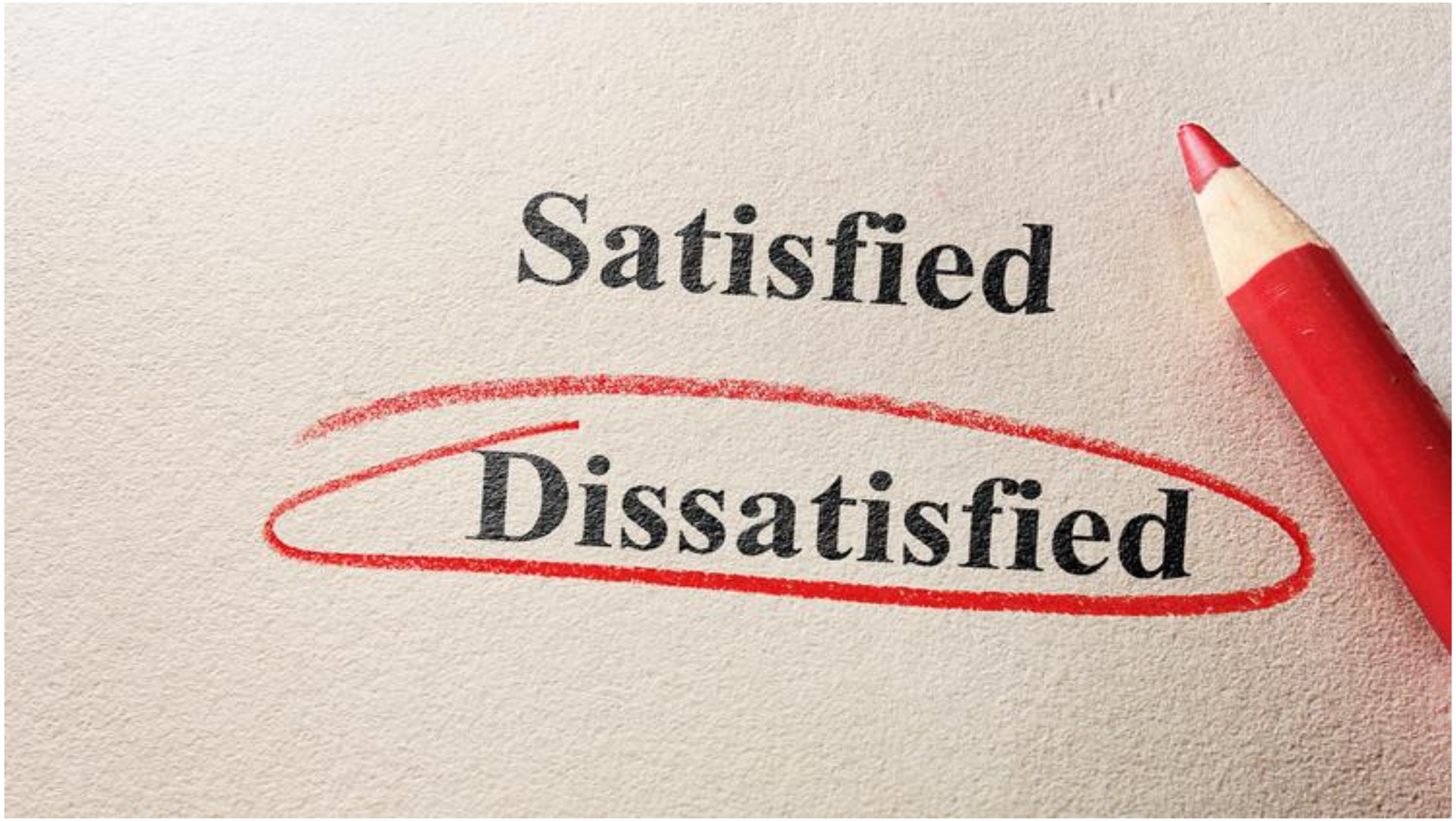
Consumer Complaints Classification

Kevin Okiah

6/10/2019

Motivation: Customer is …

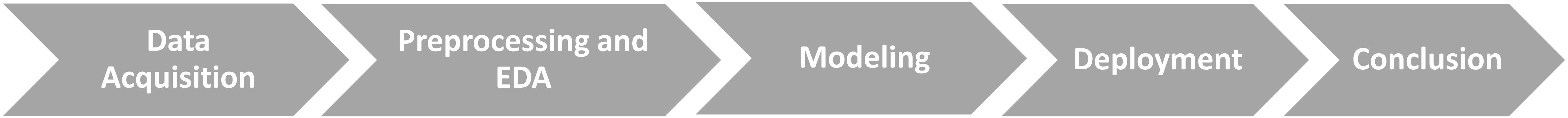


The Objective is to *build a classification model* to accurately route future complaints to the right department

A customer service research unit within Wells Fargo approaches the NLP data science team with a request for help processing an industry-wide data source of customer complaints. Their goal is to rout the complaints to the proper research team based on product groupings. The business unit receives complaints in the form of free form text and wants to route the complaints to one of seven different departments (product_group name in the data in parenthesis):

- Bank account or service (bank_service)
- Credit card (credit_card)
- Credit Reporting (credit_reporting)
- Debt collection (debt_collection)
- Lines and loans (loan)
- Money transfers (money_transfers)
- Mortgage (mortgage)

NLP Pipeline Approach



Dataset: 268,362 Customer Complaints Obtained from Customer Service Research Unit

Data Features Definition

#	Feature	What is it?	Type	Example
1	complaint_id	a message identifier	numeric	1999158, 2657456, 1414106
2	product_group	verified correct complaint department	categorical	bank_service, credit_card, credit_reporting
3	text	complaint text	text	I opened a Bank of the the West account. The a...
4	category_id	Class_label for the product group to be used for predictions	categorical	0, 1, 2, ... 6

Data Source: Wells Fargo

Dataset: Customer Complaints Distribution

Value	Count	Frequency (%)
credit_reporting	81229	30.3%
debt_collection	61473	22.9%
mortgage	40269	15.0%
loan	31036	11.6%
credit_card	29550	11.0%
bank_service	20071	7.5%
money_transfers	4734	1.8%

Random Under Sampling (RUS) used to offset class imbalance for training set.

Advantages of RUS

- Very easy to program
- Reasonable performance
- Allowed working with limited computing resources.

Disadvantage of RUS

- Can remove important instances or feature

Data is highly imbalanced with Minority Class money_transfers making 1.8% of the Customer Complaints compared to the Majority Class credit_reporting with 30.3%. Random Under Sampling during Classification model training explored to balance the classes and prevent overfitting on the Majority class.

Data Source: Wells Fargo

Preprocessing : 3.4% Duplicated Customer Complaints dropped

text
Categorical

Distinct count	259326
Unique (%)	96.6%
Missing (%)	0.0%
Missing (n)	0

There are many mistakes appear in my r...	441
Equifax mishandled my information whic...	101
I am filing this complaint because Experi...	86
Other values (259323)	267734

Dropping duplicated complaints

```
# Select all duplicate rows based on multiple column names in list
duplicateRowsDF = data[data.duplicated(['product_group', 'text'])]
#print("Duplicate Rows based on 2 columns are:", duplicateRowsDF, sep='\n')
print(duplicateRowsDF.shape)
print('sample duplicated complaint')
duplicateRowsDF[duplicateRowsDF.text == list(duplicateRowsDF.text)[0] ]
```

```
(8939, 4)
sample duplicated complaint
```

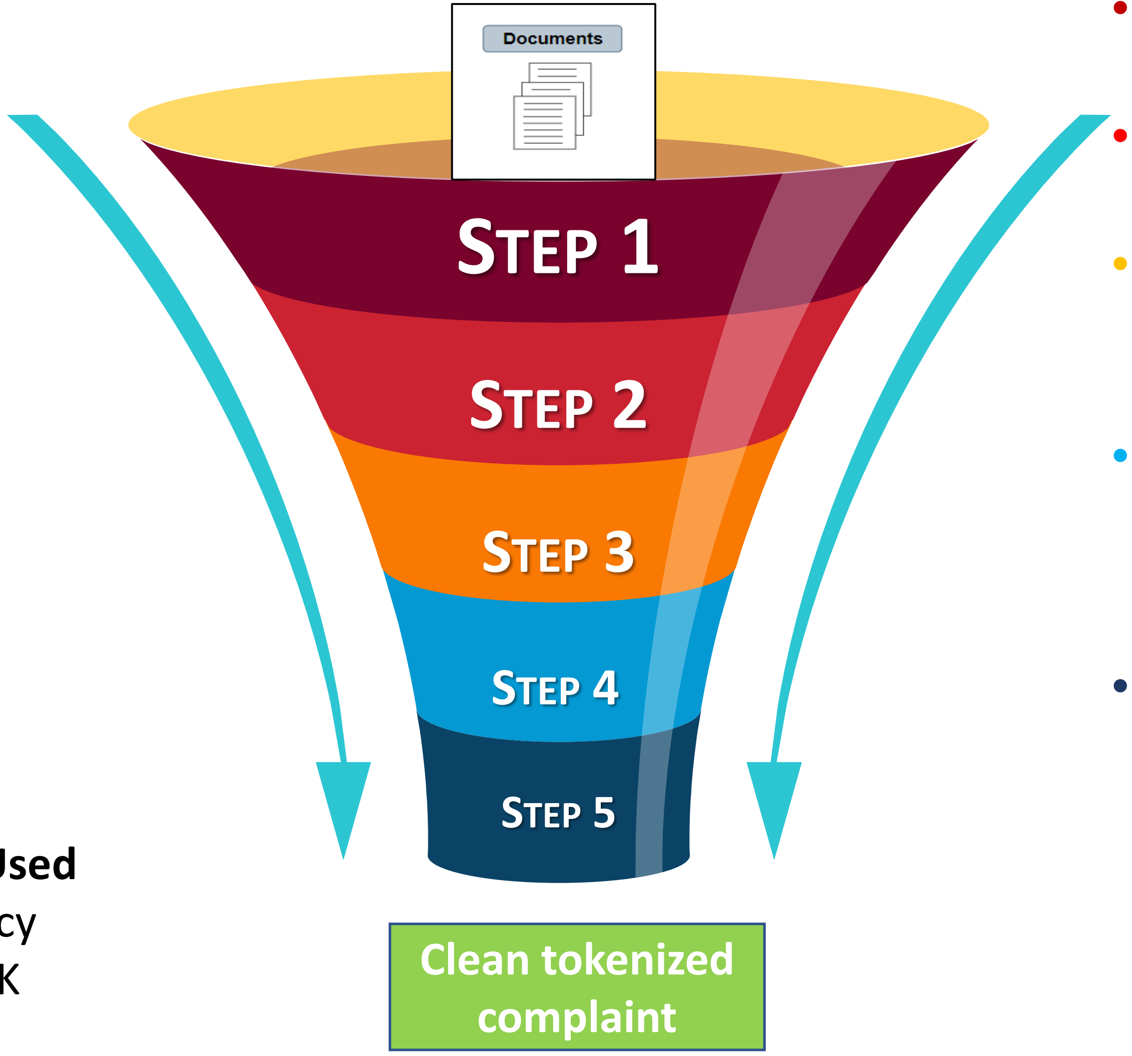
Sample Duplicated Complaint

	complaint_id	product_group	text	category_id
3899	1718737	bank_service	In 2013, Scottrade Bank illegally offset my Sc...	0
11607	1695177	bank_service	In 2013, Scottrade Bank illegally offset my Sc...	0
16785	1619071	bank_service	In 2013, Scottrade Bank illegally offset my Sc...	0

Duplicated Customer complaints are dropped leaving only one instance as long as they are filed on the same Product group

Data Acquisition → **Preprocessing and EDA** → Modeling → Deployment → Conclusion

Preprocessing: Spacy used to clean the and tokenize the complaints



- **STEP 1:**
  - Load complaint to Spacy Language model object
- **STEP 2:**
  - Spacy handles Expanding of Contractions
- **STEP 3:**
  - Spacy handles Tokenization of text and conversion to Lower fonts.
- **STEP 4:**
  - Remove Punctuations and stop words. *Numeric Masks "XXXX" were added to the Spacy Stop words vocab and removed from Complaints Tokens*
- **STEP 5:**
  - Lemmatization to base form

```
print('Raw text')
print("--------")
dataN.text[20000]
```

```
Raw text
--------

'Someone deposited counterfeit into my checking acct. and the bank honored them. The money was withdrawn the same day. They are holding me responsible for all monies withdrawn. They subsequently closed y account and say I owe them XXXX plus dollars!!'
```

Sample Complaint cleaned and tokenized

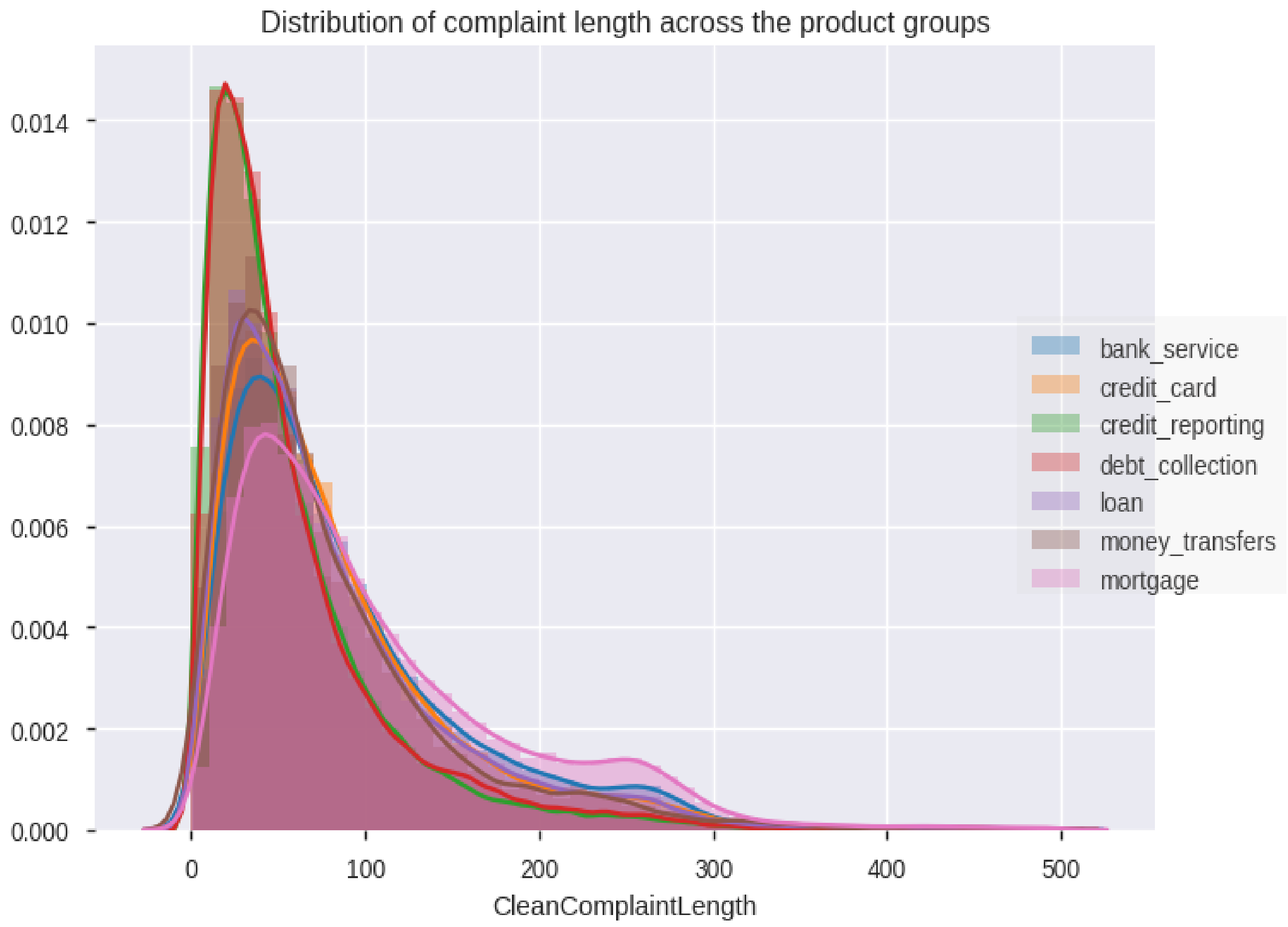
```
print('Cleaned Tokens')
print('--------------')
print(SpacyTextCleaner(dataN.text[20000], ['XX', 'XXXX']))
```

```
Cleaned Tokens
--------------
['deposit', 'counterfeit', 'check', 'acct', 'bank', 'honor', 'money', 'withdraw', 'day', 'hold', 'responsible', 'money', 'withdraw', 'subsequently', 'close', 'y', 'account', 'owe', 'plus', 'dollar']
```

Tools Used
- Spacy
- NLTK

Data Acquisition → **Preprocessing and EDA** → Modeling → Deployment → Conclusion

EDA : Engineered Feature Complaint Length



product_group	mean	median	count
credit_reporting	59.123880	41	73991
debt_collection	60.833467	41	59994
money_transfers	81.153667	58	4731
loan	83.483825	61	30974
credit_card	84.603347	64	29459
bank_service	91.377402	69	20037
mortgage	110.290280	82	40237

People Complaining mortgages write the longest complains compared to those complaining about credit reporting/ debt collection

Complaint length will potentially be a good discriminating feature for the classifiers

EDA : Top words used per Product Group

Credit Card

Bank Services

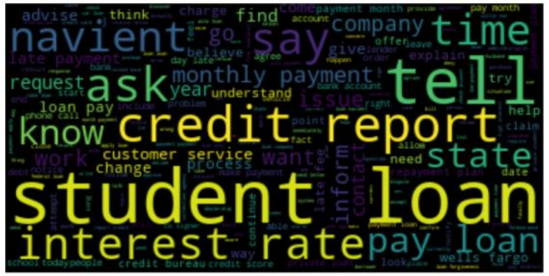
Money Transfers

Credit Reporting

Mortgages

Debt Collection

Loans



There is significant overlap of common words used in Customer Complaints for **Credit Cards, Credit Reporting , Loan** and **Debt Collection** Product Groups. Misclassification of Complaints to this Product Groups will potentially occur.

Data Acquisition | **Preprocessing and EDA** | Modeling | Deployment | Conclusion

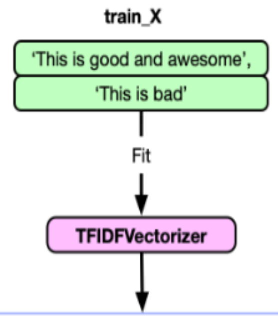
EDA: Topic Modeling - Complaints Tokens Vectorization Using TFIDF

- **Term Frequency:** How important is the word in the document?

$$TF(word\ in\ a\ document) = \frac{No\ of\ occurances\ of\ that\ word\ in\ document}{No\ of\ words\ in\ document}$$

- **Inverse Document Frequency:** How important the term is in the whole corpus?

$$IDF(word\ in\ a\ corpus) = -log(ratio\ of\ documents\ that\ include\ the\ word)$$



Term Frequency
No of time word appears/No of total terms in Document

This	is	good	bad	awesome	and
1/5	1/5	1/5	0	1/5	1/5
1/3	1/3	0	1/3	0	0

Inverse Document Frequency
-log(ratio of documents that include the word)

This	is	good	bad	awesome	and
log(2/2)	log(2/2)	log(2/1)	log(2/1)	log(2/1)	log(2/1)

x

Features

This	is	good	bad	awesome	and
0	0	1/5*log(2/1)	0	1/5*log(2/1)	1/5*log(2/1)
0	0	0	1/3*log(2/1)	0	0

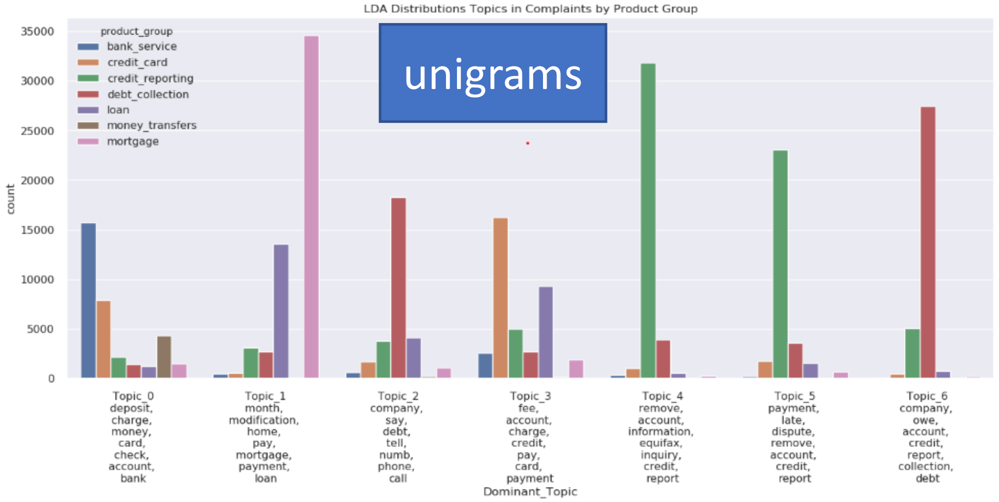
https://mlwhiz.com/blog/2019/02/08/deeplearning_nlp_conventional_methods/

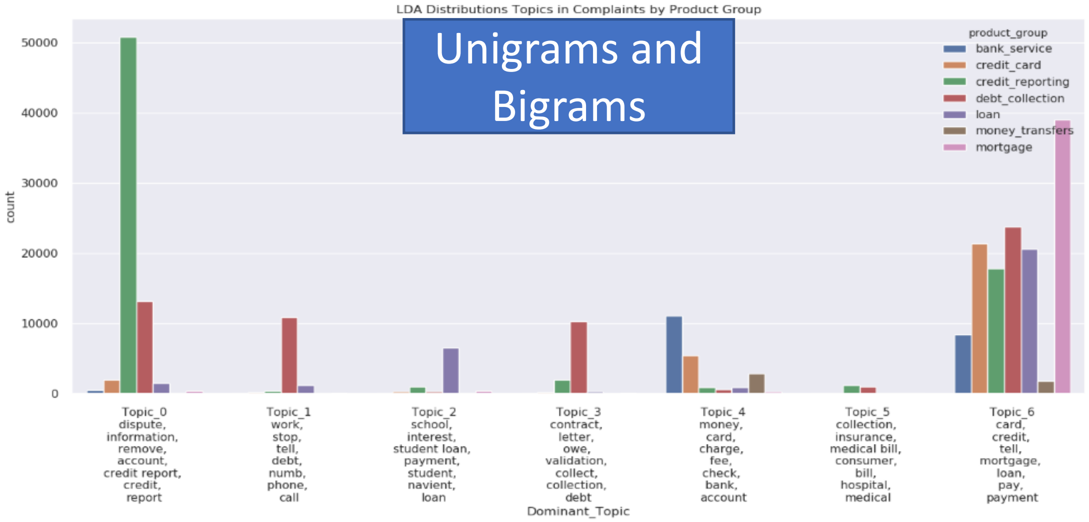
Sklearn TFIDF setup

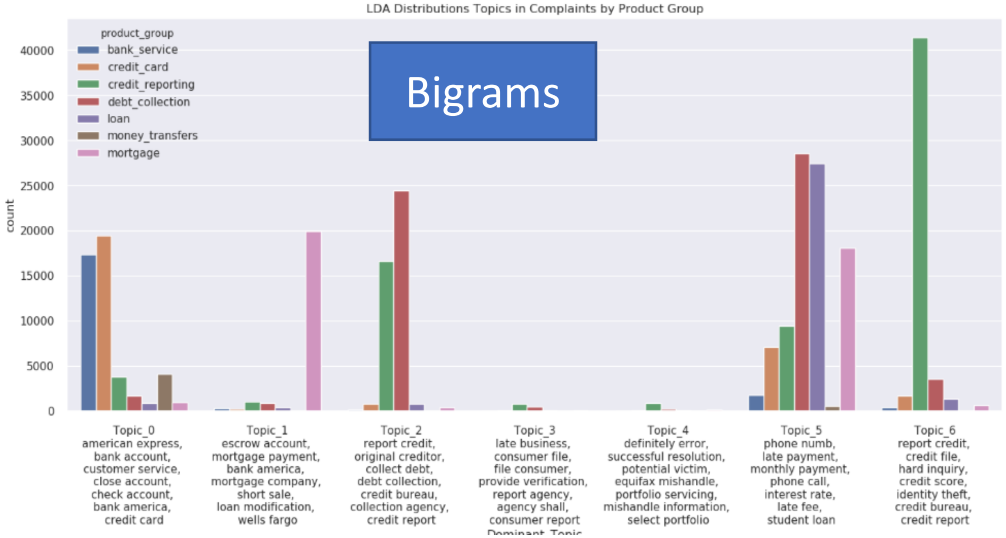
```
# TFIDF Vectorization
cv = TfidfVectorizer(max_df=0.95, min_df=5, ngram_range=(1,1))
```

- Filter unique Words and High frequency words
- N-gram Setup

EDA: Topic Modeling - Latent Dirichlet Allocation (LDA) with Unigrams gives Us a more realistic Topic distribution







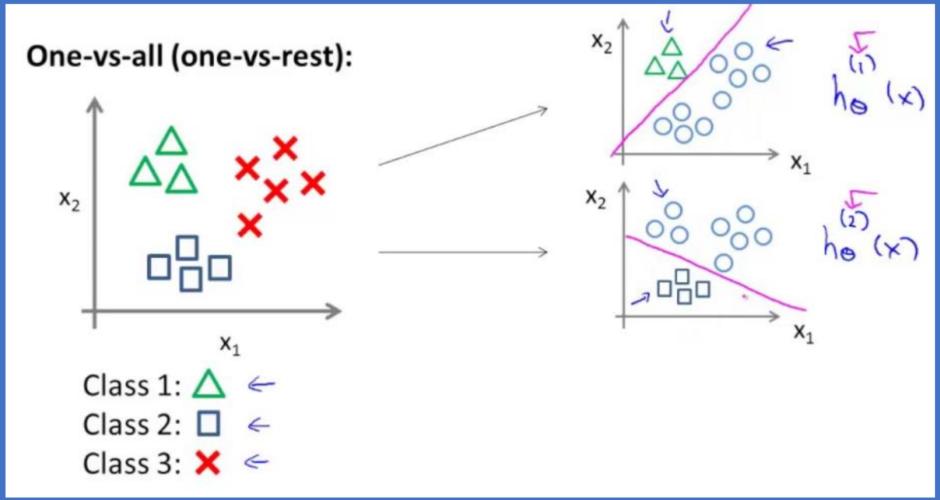
- LDA using **Unigrams** only gives us a more realistic Scenario for Classification. Each topic is to be ideally dominated by one Product Group.
- As we saw on the word clouds, we expect some overlap of topics between product categories (*Credit Cards, Credit Reports* and *Debt Collection).*
- Unigrams to be used for the classification model

Data Acquisition | **Preprocessing and EDA** | Modeling | Deployment | Conclusion

Modeling: 4 Classification Algorithms Targeted

- Logistic Regression
- Random Forest Classifier
- Linear Support Vector Classifier
- Multinomial Naïve Bayes

Modeling: Target Classification Algorithms



Andrew Ng

Logistic Regression

Random Forest Classifier

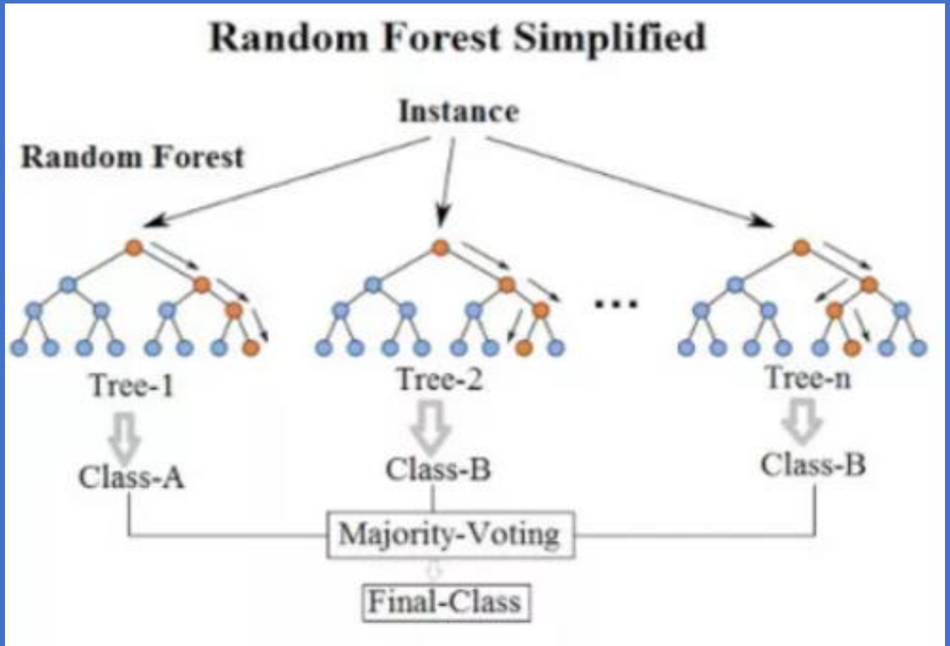
Linear Support Vector Classifier

Multinomial Naïve Bayes

Uses one vs all approach for a binary Classifier. Highest probability of the individual predicted class wins

Data Acquisition | Preprocessing and EDA | **Modeling** | Deployment | Conclusion

Modeling: Target Classification Algorithms



Logistic Regression

Random Forest Classifier

Linear Support Vector Classifier

Multinomial Naïve Bayes

An ensemble of decision trees from a random subset of the features

Data Acquisition | Preprocessing and EDA | **Modeling** | Deployment | Conclusion

Modeling: Target Classification Algorithms

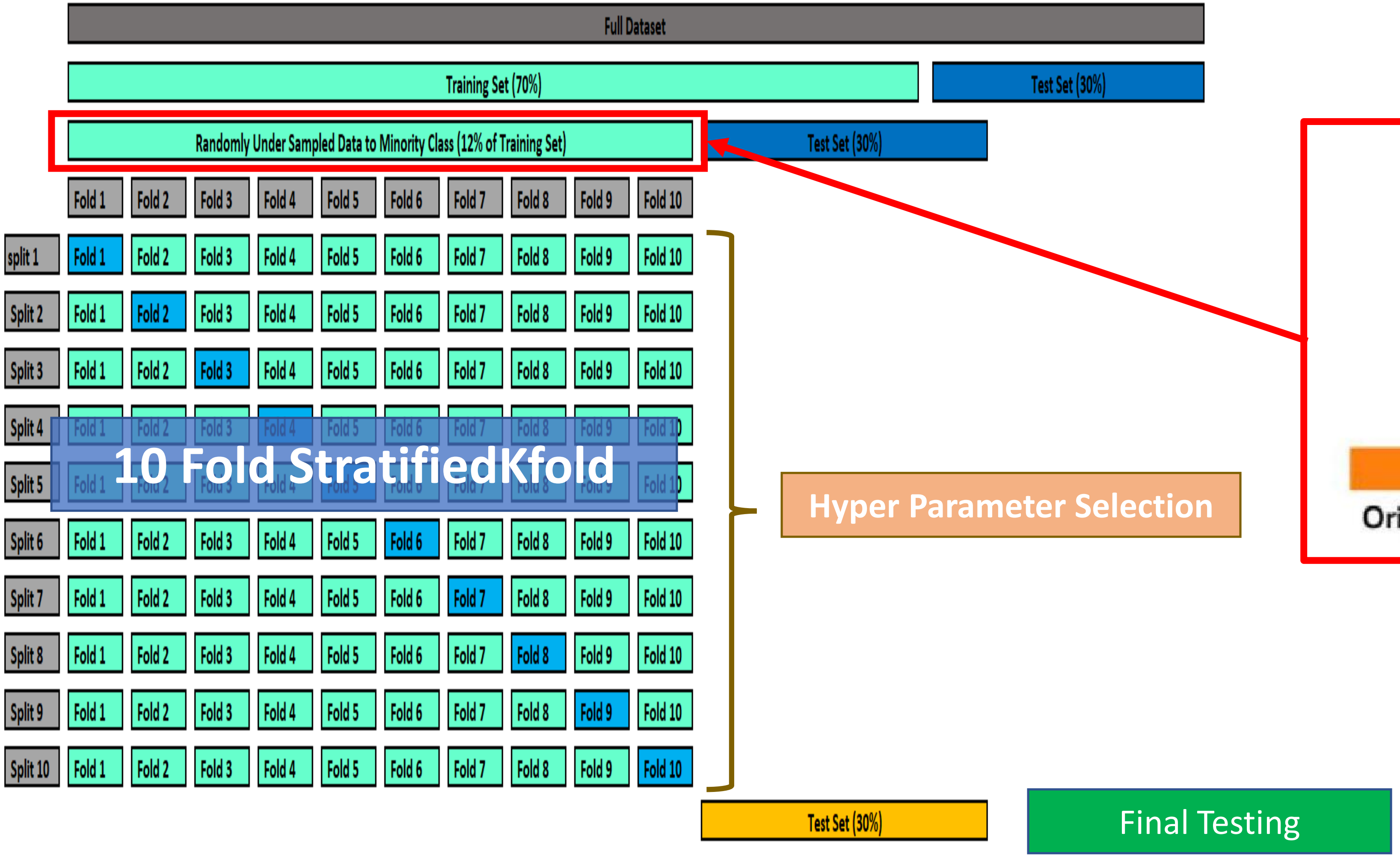
The probability of a document d being in class c is computed as

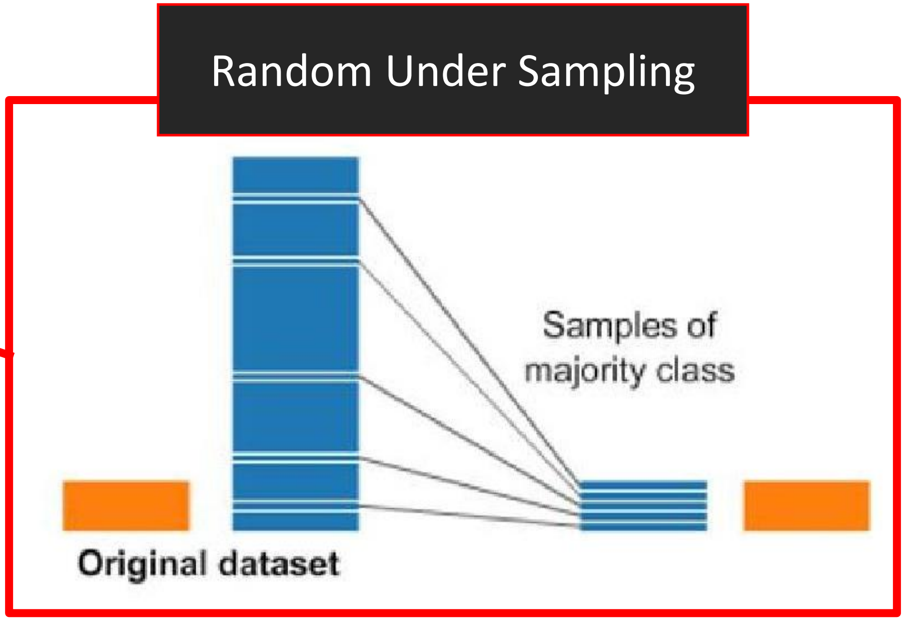
$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- Logistic Regression
- Random Forest Classifier
- Linear Support Vector Classifier
- **Multinomial Naïve Bayes**

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Modeling: Test and Training Set





- Data Split 70% Training 30% Test
- Training Set Randomly Under Sampled to Minority Class count. ~12% of Training Set used
- 10 StratifiedKfold used for Model Selection and Hyper parameter tuning

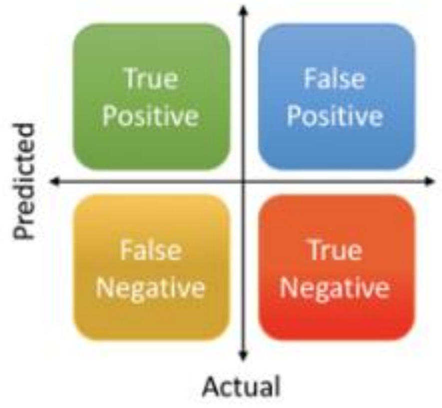
Modeling: Accuracy is the main Evaluation Metric

$$\text{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Main Evaluation Metric

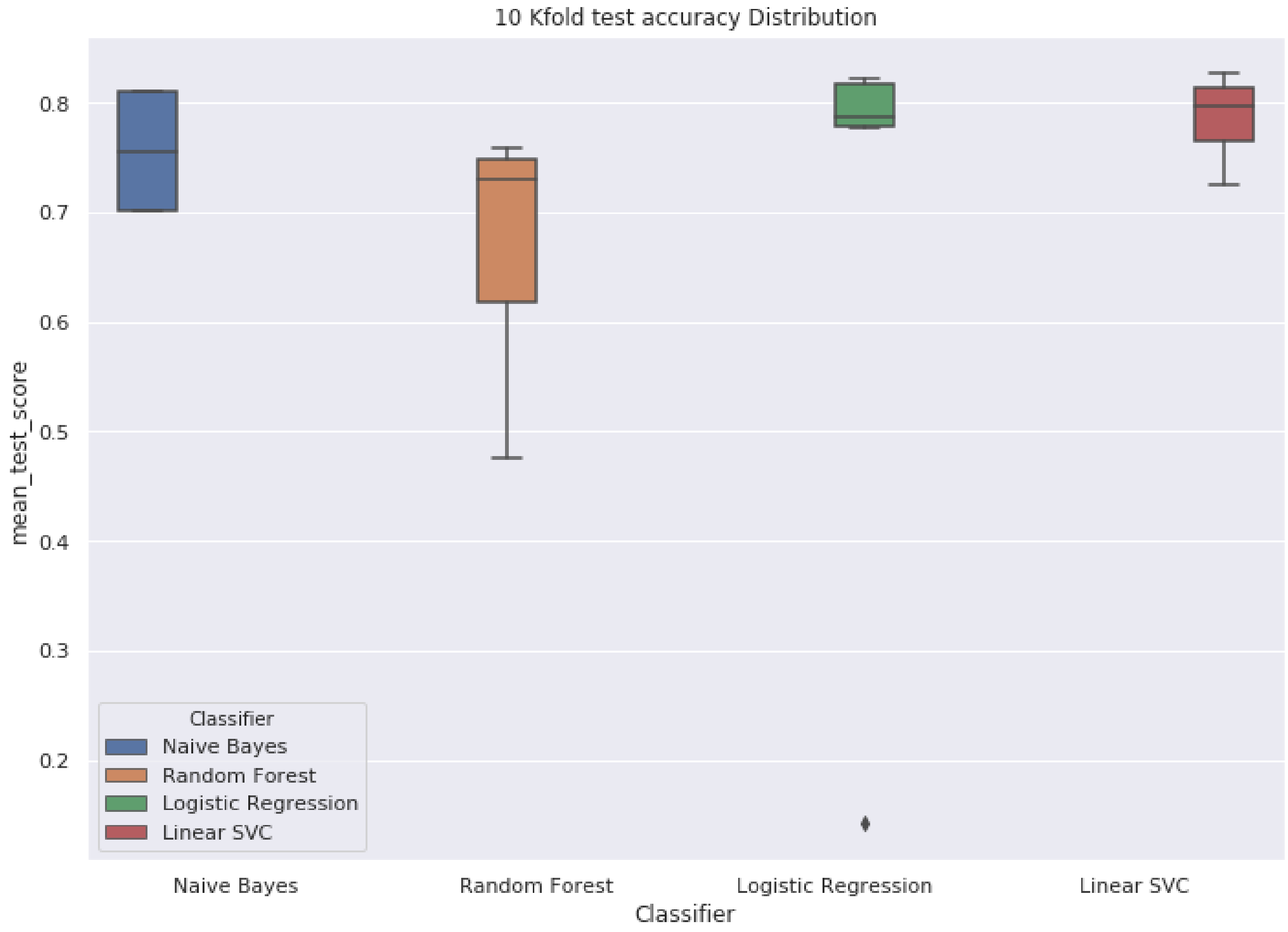
$$\boxed{\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}}$$



https://towardsdatascience.com/precision-vs-recall-386cf9f89488

Accuracy: The fraction of predictions our model got right.
Precision: Attempts to measure what proportion of positive identifications was actually correct.
Recall: Attempts to answer the question "What proportion of actual positives was identified correctly?".

Modeling: Model Selection - LinearSVC with a Training Accuracy of 82.6% is the best performing model



Classifier	max	mean	median	std
Linear SVC	0.826758	0.78663	0.796286	0.0367041
Logistic Regression	0.821352	0.732155	0.786325	0.207868
Naive Bayes	0.810018	0.755656	0.755656	0.0627713
Random Forest	0.759275	0.671603	0.7295	0.102001

- It is a Coin toss in terms for accuracy for LinearSVC 82.6% Training accuracy and Logistic Regressions 82.1% Training Accuracy as they are pretty close.
- LinearSVC is selected.

Modeling: Selected LinearSVC Classification model has a Test accuracy of 80.2%

Actual vs Predicted Confusion Matrix

Actual \ Predicted	bank_service	credit_card	credit_reporting	debt_collection	loan	money_transfers	mortgage
bank_service	4935	526	78	50	89	283	108
credit_card	652	7241	379	233	136	156	29
credit_reporting	443	A 1411	17232	B 1469	D 1088	48	517
debt_collection	479	842	B 1604	13585	C 1020	84	438
loan	242	364	479	479	7581	50	149
money_transfers	212	70	3	7	15	1138	9
mortgage	201	138	215	78	497	39	10706

Colour scale: 3000, 6000, 9000, 12000, 15000

```
%%time
# SVM
predictor = lambda x:best_svm_clf.predict(count_vect.transform([" ".join(x)]))
svm_predictions = [predictor(x) for x in X_test.Corpus]
print("svm Test Accuracy")
print('Test Accuracy Score :', accuracy_score(y_test,svm_predictions) )
```

Test Accuracy: 80.2%

```
svm Test Accuracy
Test Accuracy Score : 0.8020095853624064
CPU times: user 10.4 s, sys: 9.54 ms, total: 10.5 s
Wall time: 10.4 s
```

- The diagonal cells in the confusion matrix show the number of correct classifications by the trained classifier, while the off diagonal cells represent the misclassified predictions.
- **Our Classifier is working great as depicted by the majority predictions falling on the diagonal.(accuracy of 80.2%)**
- Significant misclassification is seen in A (*Credit Reporting vs Credit Cards*), B (*Credit Reporting and Debt collection*), C (*loan vs Debit Collection*) D (*loan vs Credit Reporting*). This high misclassification is attributed to this Product Group pairs sharing common features (words) as we saw in EDA.

Data Acquisition → Preprocessing and EDA → **Modeling** → Deployment → Conclusion

Modeling: LinearSVC Model displays mixed results for Individual classes predictions

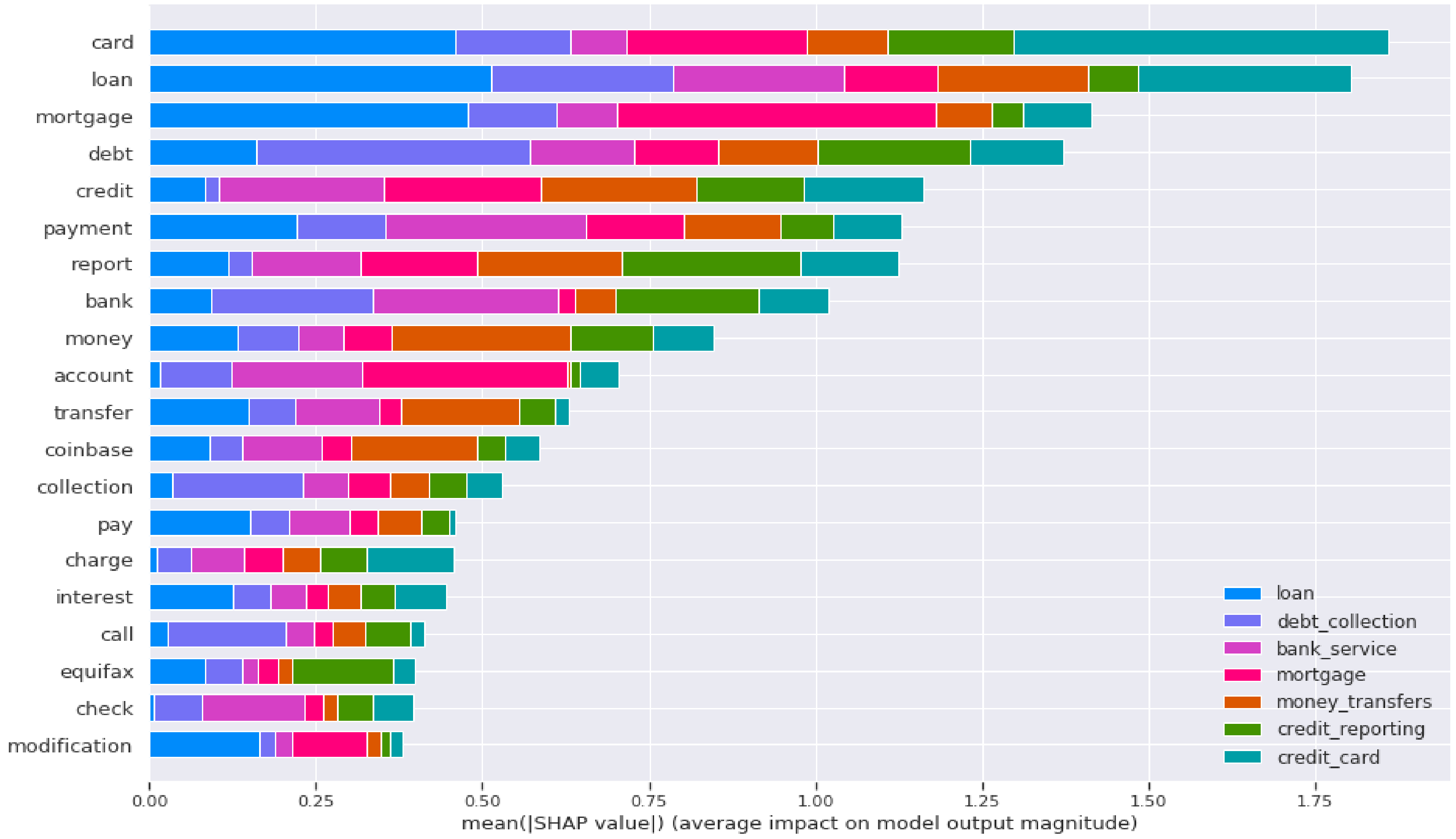
	precision	recall	f1-score	support	
bank_service	0.69	0.81	0.75	6069	Biased
credit_card	0.68	0.82	0.75	8826	
credit_reporting	0.86	0.78	0.82	22208	Conservative
debt_collection	0.85	0.75	0.80	18052	
loan	0.73	0.81	0.77	9344	Biased
money_transfers	0.63	0.78	0.70	1454	
mortgage	0.90	0.90	0.90	11874	Ideal
micro avg	0.80	0.80	0.80	77827	
macro avg	0.76	0.81	0.78	77827	
weighted avg	0.81	0.80	0.80	77827	

Biased (Low Precision, High Recall)
classifier returns many results, but most of its predicted labels are incorrect

Conservative (High Precision, Low Recall)
classifier returns very few results, but most of its predicted labels are correct when compared to the training labels.

Ideal Case (High Precision, High Recall)
classifier is returning accurate results

Modeling: Feature Importance



- Looking at the feature importance, we see top features show significant overlap in impact to classifications to Loan, debt collection, Credit_reporting and Credit card. This supports the significant miss classification that we saw in the confusion matrix.

Deployment: Classification model in action

```
(py36) kevimwe@kevimwe-System-Product-Name:~/Documents/KevinOKiahCustomerComplaintsClassification$ python ModelDeploy.py
Enter customer complaint and hit enter/return:
"Bank of America took money from my bank account for a 2014 debt without notifying me. The debt occurred as a result of the bank 's negligence in not having
a business continuity plan. They failed to close their XXXX office despite warnings of a severe snow storm. As a result, I was stranded for over 21 hours on
the interstate. As a result of that calamity, I was hospitalized at XXXX receiving XXXX. The Bank further damaged me by not approving my leave and I did n't
receive a paycheck while out on leave. This caused my account to become overdrawn. By the time I appealed and fought the decision not to pay me, my account h
ad been closed. At no time did the bank give me an opportunity to try and settle this debt. Instead, they waited 2 years to take money from my account withou
t warning. Bank of America has a history of abusing the elderly and XXXX veterans."

***** Customer Complaint entered ****
"Bank of America took money from my bank account for a 2014 debt without notifying me. The debt occurred as a result of the bank 's negligence in not having
a business continuity plan. They failed to close their XXXX office despite warnings of a severe snow storm. As a result, I was stranded for over 21 hours on
the interstate. As a result of that calamity, I was hospitalized at XXXX receiving XXXX. The Bank further damaged me by not approving my leave and I did n't
receive a paycheck while out on leave. This caused my account to become overdrawn. By the time I appealed and fought the decision not to pay me, my account h
ad been closed. At no time did the bank give me an opportunity to try and settle this debt. Instead, they waited 2 years to take money from my account withou
t warning. Bank of America has a history of abusing the elderly and XXXX veterans."

**** Prediction ****
Predicted Product Group: bank_service
```

Data Acquisition → Preprocessing and EDA → Modeling → **Deployment** → Conclusion

Conclusion

- Linear Support Vector Classification model is build to assign Customer Complaint to the correct Product Group with an accuracy of 80.2 +/- 3%.
- There is Significant of overlap in the Vocabulary used in complaints for Credit Cards, Credit Reporting and Debt Collection resulting in some misclassification of complaints going to this classes

Appendix

Additional Analysis that could be done but not covered in this summary

- Using NER to detect Unigrams, and Bigrams that can be used for classification of specific class.
- Explore Classification using Deep Learning approaches such as DNN, Bert http://jalammar.github.io/illustrated-bert/
- Engineer new features from the text data such as Complaint Length, number of punctuations etc. and see if they have any impact on the classification of the customer complaint. Masks used to hide confidential information.
- Beyond Classification of Customer Complaint, Sentiment Analysis can be done on the complaints to add a human feel to the complaint by rating the customer’s dissatisfaction level which could be used to prioritize routing customer complaint resolution.
- Evaluate performance of the classifier on the entire dataset without Random under sampling. Explore other sampling techniques, SMOTE and ROS.
- Evaluation of feature importance. What features drive the predictions.

References

- https://www.consumerfinance.gov/
- https://scikit-learn.org/stable/
- https://towardsdatascience.com/multi-class-text-classification-with-scikit-learn-12f1e60e0a9f
- https://towardsdatascience.com/https-medium-com-vishalmorde-humanizing-customer-complaints-using-nlp-algorithms-64a820cef373
- https://spacy.io/
- https://nlp.stanford.edu/IR-book/html/htmledition/naive-bayes-text-classification-1.html
- https://www.researchgate.net/post/Multiclass_Confusion_Matrix_Explanation