Analysis of CFPB's Consumer Complaints Dataset

DATA COLLECTION AND CLEANING

- The CFPB updates the consumer complaints data set fortnightly. We took the latest available data set for this milestone.

- We repeated our data cleaning steps for this new data set.

For our research question where we predict number of complaints based on financial assets of an institution:

- The financial assets data was split in multiple files, a separate file was provided for each quarter of every year from 2011-2015. We collated all data into one document.

- We mapped consumer complaints data from the consumer complaints data set to the financial data. Only institutions common to both data sets were considered.

- Since the consumer complaints data is date wise, we aggregated the data to represent quarter wise values for each year.

RESEARCH QUESTIONS

1.

Predicting a Company's Response to a Consumer Complaint

Model Description

DATA SET DETAILS:

- **Predictors:** Product, Issue, State, Submitted Via, Timely Response, Consumer Disputed
- **Outcome:** Company Response to Consumer
- **Sampling Rate:** 75 (Training): 25 (Testing)

TECHNIQUES APPLIED:

- Multinomial Logistic Regression
- Naïve Bayes
- Decision Trees*

*Two predictors, 'Issue' and 'State' were not considered as their high number of levels was too much to handle and no model was being created by the C50 library.

Problems Faced

- Too many categories in the attributes: Company, State, Issue
- Unable to apply bagging and random forests since we had more categories than R could handle.
- The data was biased; a few categories constituted more portion of the entire dataset compared to other categories

Results

Metric	Naïve Bayes		
	Standard	Laplace = 2	Laplace =3
Accuracy (Direct)	0.5856	0.635	0.639
Accuracy (Sub-Sampling)	0.4733	0.4884	0.4869
Metric	**Decision Trees**		
	Standard; size=60	Boosting Trials = 5; size=16	Boosting Trials = 10; size=7
Accuracy (Training Set)	0.763	0.758	0.762
Accuracy (Testing Set)	0.761	0.758	0.761

Note: For multinomial logistic regression, the predicted output variable is in the form of probability of classification for each record in the testing data set. Hence, we could not directly calculate it's accuracy as the model did not predict the class for each record in the testing data set.

2.

Likelihood of a Non-Disputed Complaint Feedback

Model Description

DATA SET DETAILS:

- **Predictors:** Product, Issue, State, Submitted Via, Timely Response, Company Response to Consumer
- **Outcome:** Consumer Disputed
- **Sampling Rate:** 75 (Training): 25 (Testing)

TECHNIQUES APPLIED:

- Logistic Regression
- Naïve Bayes
- Decision Trees*

*Two predictors, 'Issue' and 'State' were not considered as their high number of levels was too much to handle and no model was being created by the C50 library.

Results

Metric	Naïve Bayes			Decision Trees, size=1
	Standard	Laplace = 2	Laplace =3	
Accuracy (Direct)	0.5856	0.635	0.639	0.792 (training) 0.7912 (testing)

Logistic Regression:

Coefficients: (3 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.71E+00	8.10E-01	-2.111	0.03473	*
ProductConsumer Loan	6.99E-02	4.53E-02	1.541	0.12334	
ProductCredit card	9.77E+09	1.09E+10	0.899	0.36868	
ProductCredit reporting	-4.28E-01	3.60E-02	-11.893	< 2e-16	***
ProductDebt collection	-3.07E-01	3.62E-02	-8.489	< 2e-16	***
ProductMoney transfers	9.77E+09	1.09E+10	0.899	0.36868	
ProductMortgage	9.77E+09	1.09E+10	0.899	0.36868	
ProductOther financial service	9.77E+09	1.09E+10	0.899	0.36868	
ProductPayday loan	-3.63E-01	1.46E-01	-2.479	0.01317	*
ProductPrepaid card	9.77E+09	1.09E+10	0.899	0.36868	
ProductStudent loan	3.78E-03	9.02E-02	0.042	0.96659	
ProductVirtual currency	9.77E+09	1.09E+10	0.899	0.36868	
IssueAccount terms and changes	-2.32E-02	1.37E-01	-0.17	0.86532	

Note: Boosting was abandoned after the first trial as there were very few classifiers. The results of the first trial were same as the one without boosting (accuracy: 0.7912).

3.
Predicting Medium of Receiving a Complaint

Model Description

DATA SET DETAILS:

- **Predictors:** Product, Company*, Region (derived from State), Timely Response
- **Outcome:** Submitted Via
- **Sampling Rate:** 75 (Training): 25 (Testing)

TECHNIQUES APPLIED:

- Multinomial Logistic Regression
- Naïve Bayes
- Decision Trees

*The variable 'Company' has more than 4,000 levels and considering all was causing memory issues so we considered only the top 10 companies

Results

Metric	Naïve Bayes		
	Standard	Laplace = 2	Laplace =3
Accuracy (Direct)	0.5993	0.6272	0.6305

Metric	Decision Trees		
	Standard, size=9	Boosting Trials = 5, size=1	Boosting Trials = 10
Accuracy (Training Set)	0.6710	0.670	-
Accuracy (Testing Set)	0.6688	0.668	-

Logistic Regression:

```
> exp(coef(mlr))
            (Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection ProductMoney transfers
Fax           26.225677            1.8450416          0.8288470              2.9348885              387.6933              1.862786
Phone         45.581748            1.1952652          0.5547044              0.1440963              201.3557              1.388716
Postal mail    6.990420            2.1852150          1.2820750              6.6792403              586.4462              2.363431
Referral      20.657154            0.8682422          0.5372249              0.6555439              219.7261              1.280046
Web            1.866432            2.0688050          1.4059620              5.3274269              801.2740              3.881994
            ProductMortgage ProductOther financial service ProductPayday loan ProductPrepaid card ProductStudent loan
Fax               0.7604339                     3.72935525          0.3260389          2.58666068          0.2515871
Phone             0.2177566                     1.23480645        122.9709136          2.83080270          0.1852866
Postal mail       0.4837196                     5.82108162          0.4038667          4.56166518          0.2979563
Referral          0.5097922                     5.27721420          0.1990887          1.66719154          0.3310162
Web               0.6358929                     0.01891042          0.3591731          0.03056313          1.5662273
            ProductVirtual currency CompanyCapital One CompanyCitibank CompanyEquifax CompanyExperian CompanyJPMorgan Chase & Co.
Fax                               1          0.5486497       0.8836215      1.2644318       0.7065171                 0.9557087
Phone                             1          0.6573920       0.8546788      1.7149349       0.7684238                 0.8928083
Postal mail                       1          0.5834590       1.0740219      1.8293612       1.2395461                 1.0277891
Referral                          1          0.4181464       0.7703361      1.7585093       1.0783387                 0.9510592
Web                               1          0.6021676       0.8141438      0.7934537       0.5215272                 0.8156520
            CompanyNavient Solutions, LLC. CompanyOcwen CompanyTransUnion Intermediate Holdings, Inc. CompanyWells Fargo & Company
Fax                          88.077243121     2.027874                                    0.6123326                    1.209494
Phone                         8.588175008     1.808223                                    0.9099728                    1.341162
Postal mail                  13.986487572     2.538878                                    0.9793689                    1.162972
Referral                      0.034673046     1.537015                                    0.9921956                    1.134088
Web                           0.007125652     2.071541                                    0.2794377                    1.101926
            RegionNorth East RegionSouth RegionWest Timely.response.Yes
Fax                0.2787082   0.3159725  0.2570108              4.298457
Phone              0.4244127   0.4567746  0.3175527              5.492623
Postal mail        0.3697014   0.6031484  0.3460993              6.252651
Referral           0.5045537   0.5349082  0.4271113              6.282950
Web                0.4160728   0.4764664  0.4206593             41.214169
```

Note: For trials = 10, the boosting truncated forcibly at 4 trials since the last classifier was very inaccurate. The error rate was hence similar to the results for trials = 5.

4.
Predict Geographical Location Of a Complaint

Model Description

DATA SET DETAILS:

- **Predictors:** Product, Submitted Via, Timely Response, Consumer Disputed, Consumer Response to Consumer, Company*
- **Outcome:** Region (derived from State)
- **Sampling Rate:** 75 (Training): 25 (Testing)

TECHNIQUES APPLIED:

- Multinomial Logistic Regression
- Naïve Bayes
- Decision Trees

*The variable 'Company' has more than 4,000 levels and considering all was causing memory issues so we considered only the top 10 companies

Results

Metric	Naïve Bayes		
	Standard	Laplace = 2	Laplace =3
Accuracy (Direct)	0.3814	0.4083	0.4085

Metric	Decision Trees		
	Standard, size=324	Boosting Trials = 5	Boosting Trials = 10
Accuracy (Training Set)	0.438	-	-
Accuracy (Testing Set)	0.434	-	-

Logistic Regression:

```
call:
multinom(formula = Region ~ Product + Company + Submitted.via +
    Timely.response. + Company.response.to.consumer + Consumer.disputed.,
    data = train_sample)

Coefficients:
           (Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection ProductMoney transfers ProductMortgage
North East   1.1059602          -0.5915326         -0.2845894           -0.12671547           -0.3967867           -0.3752858      -0.2874055
South        0.6610853          -0.3323854         -0.4454668            0.03905251           -0.1690341           -0.1124373      -0.4572507
West         1.1089548          -0.5390341         -0.3615747           -0.29545674           -0.1671258            0.1480401      -0.3932038
           ProductOther financial service ProductPayday loan ProductPrepaid card ProductStudent loan CompanyCapital One CompanyCitibank CompanyEquifax
North East                     -0.8467621          0.5388157         -0.86284885          -0.4189964       -0.05685274     -0.07348518     -0.5965742
South                          -0.2912298         -0.9969629         -0.66752470          -1.1472150       -0.23922030     -0.38776663     -0.2926961
West                           -0.7118204          0.1718518         -0.03069452          -1.0696017       -0.51204449     -0.52018627     -0.4439083
           CompanyExperian CompanyNavient Solutions, LLC. CompanyOcwen CompanyTransunion Intermediate Holdings, Inc. CompanyWells Fargo & Company
North East      -0.3971552                    -3.834122 -0.006661388                                   -0.6654561                    0.02345911
South           -0.2628592                    -3.634074 -0.008819389                                   -0.5098750                    0.13141090
West            -0.2278824                    -3.482837 -0.029829015                                   -0.5805497                    0.15088121
           Submitted.viaFax Submitted.viaPhone Submitted.viaPostal mail Submitted.viaReferral Submitted.viaWeb Timely.response.Yes
North East       -1.494938         -0.9947133             -0.9960539          -0.8110315       -1.1372750           0.5554310
South            -0.912436         -0.6609123             -0.1973840          -0.5241770       -0.5701131           0.4040763
West             -1.013326         -0.9494058             -0.6871076          -0.6214247       -0.6778505           0.3073750
           Company.response.to.consumerClosed with explanation Company.response.to.consumerClosed with monetary relief
North East                                          -0.12690268                                                -0.1715438
South                                               -0.06416656                                                -0.1101095
West                                                -0.17138298                                                -0.2243475
           Company.response.to.consumerClosed with non-monetary relief Company.response.to.consumerClosed with relief
North East                                                  -0.1146284                                        -0.03725757
South                                                       -0.1290300                                        -0.45682372
West                                                        -0.1775768                                        -0.02692852
           Company.response.to.consumerClosed without relief Consumer.disputed.Yes
North East                                         -0.2716412             -0.01335115
South                                              -0.4103781              0.00335657
West                                               -0.1205451              0.09178060
```

Note: For Boosting Trials = 5,10 the tree was truncated at the first trial itself and hence results of boosting are same as the standard decision tree.

5.
Effect Of Total Assets on #Complaints

Model Description

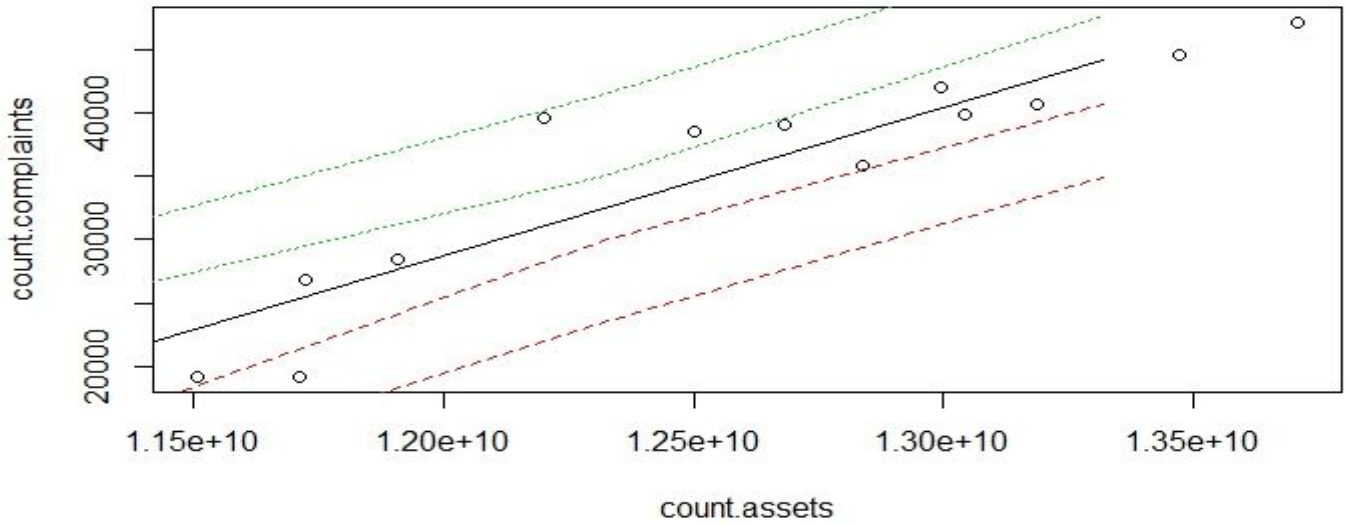
DATA SET DETAILS:

- **Predictors:** Assets of Financial Institutions
- **Outcome:** Number of Complaints
- **Sampling Rate:** 70 (Training): 30 (Testing)

TECHNIQUES APPLIED:

- Linear Regression

Results



- **Intercept:** -1.08e+05
- **Coefficient:** 1.163e-05
- **Correlation:** 0.3428
- **Mean Square Error:** 12071733

```
[1] "cor:  0.342840901065586  MSE:  6305398.16967114"
[1] "cor:  0.780532751861594  MSE:  8106072.34906073"
[1] "cor:  0.932646883668888  MSE:  5282470.45968027"
[1] "cor:  0.896841329846306  MSE:  5290419.98159268"
[1] "cor:  0.778974106322222  MSE:  5007942.84844342"
[1] "cor:  -0.783365868539241  MSE:  5355001.16582222"
[1] "cor:  0.105880522202444  MSE:  13920646.2223951"
[1] "cor:  0.473482025078609  MSE:  15247655.3431237"
[1] "cor:  0.713338117593158  MSE:  15152449.4886729"
[1] "cor:  0.78485209134768  MSE:  15734453.9531854"
[1] "cor:  0.906877937166014  MSE:  15157079.9942371"
[1] "cor:  0.904201443080769  MSE:  0"
[1] "cor:  0.898248041638909  MSE:  0"
[1] "cor:  -0.890404459009079  MSE:  0"
```

6.

Emotion Classification of Consumer Narratives

Model Description

DATA SET DETAILS:

- **Predictors:** Consumer Complaint Narrative
- **Outcome:** Emotion of The Narrative
- **Classes** : Anger, Sad, Dispute, Anticipation, Fear

API USED:

- Aylien API

METHODOLOGY:

- Used the Aylien API in Python to label 1,000 records in the data set.
- For the next milestone, we hope to use this labeling to predict the emotion of a narrative by building a classifier using supervised learning.

Thanks!

Any Questions?