

Milestone 2

Team Members: Rajat Aghi, Pal Doshi, Khushal Navani, Akash Udani

Data Collection and Cleaning Efforts:

- a. The Consumer Finance Protection Bureau (CFPB) updates the Consumer Complaints data set fortnightly. We took the latest available data set for the purpose of this milestone.
- b. Some of the variables in the new data set were blank (ex. 'State', 'Consumer.disputed.'). These rows were removed from the data set.
- c. Since the variable 'State', of type factor, has 68 levels, we created a new variable called region which mapped all the states into 4 regions.
- d. For research question 5
 - i. We had to clean the files which contained information about the various financial institutions and their quarterly total assets. The data wasn't present for each and every quarter of every year, there were some files missing and in such case we had to drop those corresponding quarter rows from the consumer complaint database.
 - ii. For the files from 2011 Q2 to 2012 Q2, the bank/company names weren't uniquely identified by any ID, but from 2012 Q3 onwards each file had an additional column to uniquely identify the same. We did a manual reverse matching of the names and assigned the unique ID to the old files.
 - iii. We combined the data so obtained from all the asset related files and compiled them into one single document.
 - iv. After the combined document was created, there was a lot of whitespace characters, and we trimmed the extra space and special characters in Excel.
 - v. We created a mapping between the name of the financial institutions from the assets file and financial institutions from the consumer complaint database as their names weren't similar and we couldn't directly match them unless we performed a manual mapping between the names.
 - vi. We eliminated those financial institutions whose data wasn't present in the consumer complaint file but was present in the assets file and vice versa. In simpler words we performed an inner join between both the files based on the company name.
 - vii. The assets data we obtained had assets listed per quarter from 2011 to 2016. On the other hand, the consumer complaints data set lists complaints received by the CFPB, date wise. We aggregated the consumer complaints data set as per quarter from 2011 to 2016.

Research Question 1:**Predicting the response to a particular consumer complaint****Predictors:** Product, Issue, state, Submitted.via., Timely.response., Consumer.disputed.**Outcome:** Company.response.to.consumer (8 levels)**Sampling rate:** (Training:Testing)=75:25**Linear Regressions**

In this question, the dependent variable is a nominal variable with 8 factors (Closed/Closed with explanation/Closed with monetary relief/ Closed with non monetary relief, Closed with relief/ Closed with no relief/Untimely response/ In progress) and all the independent variables are multi-level nominal variables. Since the output variable is not continuous, linear regression cannot be applied for this question.

(Multinomial) Logistic Regression

We could not apply logistic regression directly on this question since the outcome variable is not binary. There are 8 levels as mentioned above. We could apply multinomial logistic regression, after some amount of sub-sampling for reducing the number of levels in the predictor variables. The results after applying this technique are as shown below

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

Coefficients:
(Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection
Closed -4.766778 0.2449441 -1.2288831 5.9648433 7.1022204
Closed with explanation -9.086534 0.4436236 -0.6846374 6.8025906 6.7241304
Closed with monetary relief -5.981112 -0.4980601 -1.0098254 2.8647036 4.8866994
Closed with non-monetary relief -4.909174 0.8172455 -0.6411912 7.3045106 7.4861419
Closed without relief -1.324661 0.1424075 0.2071223 -0.3901148 -0.8722197

ProductMoney transfers ProductMortgage ProductOther financial service ProductPayday loan
Closed 6.005648 1.2660529 8.1180782 -1.706981
Closed with explanation 6.166882 1.9030886 -1.0032365 4.886197
Closed with monetary relief 5.593056 -0.7634347 -2.7626063 4.850806
Closed with non-monetary relief 5.213512 1.1019531 -0.7407153 -1.123693
Closed without relief -1.635031 2.4224885 -0.6807093 -1.862240

ProductPrepaid card ProductStudent loan CompanyCapital one CompanyCitibank CompanyEquifax
Closed -0.6250472 0.8568133 0.167880880 -1.106342800 0.2396908
Closed with explanation -1.3491738 0.7440262 0.331774290 -0.068463411 1.5382878
Closed with monetary relief 6.6592375 -2.6528108 0.294639867 0.457168367 -0.9851929
Closed with non-monetary relief -0.9542155 0.9683947 -0.007346563 0.009169237 2.8255498
Closed without relief -0.4822357 1.7868019 0.572433652 -0.421519512 0.2543941

CompanyExperian CompanyNavient Solutions, LLC. CompanyOcwen
Closed 1.4082802 -9.628595 1.1022738
Closed with explanation 1.3178897 -9.497241 -0.2752694
Closed with monetary relief 3.0486056 -6.522671 -1.6257477
Closed with non-monetary relief 2.8963710 -10.884399 -1.5353074
Closed without relief 0.2375914 -7.297269 -0.7459325

CompanyTransunion Intermediate Holdings, Inc. Companywells Fargo & CompanyRegionNorth East RegionSouth
Closed 0.1613209 -0.2355539 0.1184613 0.09949657
Closed with explanation 2.2873939 0.8135603 -0.1870855 -0.08075334
Closed with monetary relief 1.3815501 0.3605799 -0.2035885 -0.25961077
Closed with non-monetary relief 2.8823071 0.3146305 -0.2966082 -0.24778105
Closed without relief 0.4562535 0.7083963 -0.2978879 -0.07739006

Regionwest Submitted.viaFax Submitted.viaPhone Submitted.viaPostal mail Submitted.viaReferral
Closed 0.10111851 4.0263939 3.5168248 4.2985302 3.4791876
Closed with explanation -0.05298872 4.7405852 4.3893795 5.1801961 3.7240983
Closed with monetary relief -0.28166774 4.8752257 4.2722604 5.0170714 3.9558208
Closed with non-monetary relief -0.13393673 3.3641569 2.3793904 3.0581803 1.8070692
Closed without relief 0.03793787 -0.2572413 0.2714878 0.5017654 0.2550014

Submitted.viaWeb Timely.response.Yes Consumer.disputed.Yes
Closed 3.45315412 0.65064712 0.4502022
Closed with explanation 4.01715455 4.00637915 0.6262966
Closed with monetary relief 4.87616448 2.10497433 -0.3302438
Closed with non-monetary relief 2.43325624 1.29376836 -0.1759732
Closed without relief 0.05455458 -0.06048239 0.6446952
```

```

Std. Errors:
(Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection
closed 2.9120232 0.3069203 0.1406019 3.393298 2.695875
closed with explanation 3.8649430 0.2938918 0.1344431 3.385098 2.697176
closed with monetary relief 3.7171257 0.2698520 0.0985518 3.385207 2.696928
closed with non-monetary relief 2.6440468 0.3464840 0.1762276 3.385724 2.698302
closed without relief 0.7469791 0.3304056 0.1251105 6.992239 7.891383

ProductMoney transfers ProductMortgage ProductOther financial service ProductPayday loan
closed 6.928620 0.1185080 27.77011 41.22059
closed with explanation 6.927986 0.1211402 83.18979 11.93245
closed with monetary relief 6.919856 0.1189353 83.72385 11.90660
closed with non-monetary relief 6.986054 0.1623125 72.97187 42.89155
closed without relief 25.688934 0.1217107 81.52517 39.54927

ProductPrepaid card ProductStudent loan CompanyCapital One CompanyCitibank CompanyEquifax
closed 64.52182 0.4673517 0.1543105 0.14594772 5.027296
closed with explanation 62.49251 0.4573885 0.1537563 0.12020670 5.018578
closed with monetary relief 22.01656 0.7868518 0.1173250 0.09895921 5.124799
closed with non-monetary relief 76.65346 0.5105354 0.2032929 0.15415855 5.018363
closed without relief 58.16638 0.3943078 0.1279830 0.11251436 10.685605

CompanyExperian companyNavient Solutions, LLC. CompanyOcwen
closed 3.361602 19.935744 0.1362160
closed with explanation 3.353930 25.569414 0.1504989
closed with monetary relief 3.332439 12.508209 0.2901295
closed with non-monetary relief 3.352827 53.017274 0.2954318
closed without relief 6.607551 5.792419 0.1494099

companyTransunion Intermediate Holdings, Inc. Companywells Fargo & Company RegionNorth East Regionsouth
closed 3.648719 0.1212277 0.1344709 0.1254731
closed with explanation 3.628943 0.1101557 0.1303190 0.1198543
closed with monetary relief 3.630046 0.1068130 0.1170714 0.1100736
closed with non-monetary relief 3.627969 0.1526662 0.1539824 0.1395529
closed without relief 6.855478 0.1053405 0.1264113 0.1162372

Regionwest Submitted.viaFax Submitted.viaPhone Submitted.viaPostal mail Submitted.viaReferral
closed 0.1308491 2.9263773 2.9038319 2.9071500 2.9015785
closed with explanation 0.1254597 3.7465798 3.7300511 3.7324330 3.7287314
closed with monetary relief 0.1160116 3.7189725 3.6989701 3.7018893 3.6977697
closed with non-monetary relief 0.1474471 2.6320510 2.6054078 2.6073106 2.6015627
closed without relief 0.1203690 0.8387948 0.7234472 0.7408659 0.7158259

Submitted.viaweb Timely.response.Yes Consumer.disputed.Yes
closed 2.9014739 0.2360110 0.09807583
closed with explanation 3.7285197 1.0138685 0.09488889
closed with monetary relief 3.6974833 0.3738399 0.10218668
closed with non-monetary relief 2.6007155 0.4698686 0.12363891
closed without relief 0.7152649 0.1936061 0.09168980

Residual Deviance: 28420.67
AIC: 28710.67

```

In this case, **'Email'** is set as the base reference medium against which all other mediums are compared.

Similarly, **'Bank account or service'** is set as the base reference against which all other Products are compared.

'Midwest' region is set as the base region while **'Bank of America'** is set as the base Company for all corresponding comparisons.

The base reference value is **'No'** for both Timely response and Consumer Disputed variables.

To give a brief idea about how we interpret these results,

The log odds for the company's response being 'Closed' versus 'Closed with Relief' increase by 0.24 as we move from Product 'Bank account or service' to 'Consumer Loan'.

The log odds for the company's response being 'Closed with explanation' versus 'Closed with relief' will decrease by 0.068 if the institution is 'Citibank' as compared to 'Bank of America'.

Actual values for the training sample

	Product	Company	Region	Submitted.via	Company.response.to.consumer	Timely.response:	Consumer.disputed:
400952	Mortgage	Citibank	South	Web	Closed with relief	Yes	Yes
317869	Mortgage	Ocwen	South	Referral	Closed	Yes	No
4390	Credit reporting	Experian	South	Postal mail	Closed with explanation	Yes	No
115888	Mortgage	Bank of America	West	Referral	Closed without relief	Yes	Yes
401151	Credit card	Capital One	North East	Referral	Closed	Yes	No
118851	Bank account or service	Bank of America	West	Web	Closed without relief	Yes	Yes
255169	Mortgage	Wells Fargo & Company	North East	Referral	Closed with relief	Yes	No
47304	Money transfers	Bank of America	South	Web	Closed with monetary relief	Yes	No
38831	Credit card	Bank of America	Midwest	Web	Closed	Yes	No
1211	Mortgage	Ocwen	Midwest	Referral	Closed with explanation	Yes	No

Fitted values for the training sample

```
> head(pp <- fitted(mlr),n=10)
```

	Closed with relief	Closed closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed without relief
400952	9.134769e-02	0.09568915	0.34098577	0.101349899	0.055695084
317869	9.796397e-02	0.61123384	0.11858247	0.007504402	0.008126483
4390	5.413502e-05	0.11429071	0.18558145	0.048344771	0.651711153
115888	6.594657e-02	0.21470716	0.20218526	0.018050653	0.023868316
401151	4.210703e-01	0.08677473	0.06323097	0.181945809	0.026824556
118851	2.114438e-01	0.18910513	0.12957585	0.311710380	0.047555631
255169	8.009949e-02	0.13365828	0.25898651	0.047299956	0.040241900
47304	1.009125e-03	0.23304965	0.15327164	0.568246896	0.044374560
38831	3.569056e-01	0.05381818	0.06216755	0.353425510	0.057635060
1211	1.010777e-01	0.57093378	0.13264179	0.010038129	0.010742410

Actual values for the testing sample

	Product	Company	Region	Submitted.via	Company.response.to.consumer	Timely.response:	Consumer.disputed:
401426	Credit card	Capital One	Midwest	Postal mail	Closed with relief	Yes	No
18442	Credit reporting	Experian	Midwest	Web	Closed with non-monetary relief	Yes	No
543462	Credit card	Citibank	West	Web	Closed with relief	Yes	No
116505	Bank account or service	Wells Fargo & Company	South	Referral	Closed without relief	Yes	No
2600	Mortgage	Bank of America	South	Web	Closed with explanation	Yes	No
20231	Credit reporting	TransUnion Intermediate Holdings, Inc.	Midwest	Web	Closed with non-monetary relief	Yes	No
262086	Credit card	Capital One	North East	Web	Closed with relief	Yes	No
705	Credit reporting	Experian	South	Web	Closed with non-monetary relief	Yes	Yes
116008	Bank account or service	Wells Fargo & Company	Midwest	Referral	Closed without relief	Yes	No
114757	Credit card	Citibank	South	Postal mail	Closed with relief	Yes	No
55808	Mortgage	Ocwen	South	Web	Closed	Yes	Yes
400487	Credit card	Bank of America	South	Web	Closed with relief	Yes	Yes
121389	Mortgage	Bank of America	North East	Referral	Closed with relief	Yes	No
119818	Credit card	Capital One	South	Web	Closed with relief	Yes	No
29487	Bank account or service	Wells Fargo & Company	South	Web	Closed with monetary relief	Yes	No

Fitted values for the testing sample

```
> predict(nlr,newdata=test_sample,"probs")
      Closed with relief      Closed with explanation      Closed with monetary relief      Closed with non-monetary relief      Closed without relief
401426      2.031054e-01      8.436240e-02      1.577328e-01      3.109005e-01      6.082264e-02      1.830762e-01
18442      8.892578e-05      7.298058e-02      1.032885e-01      8.942343e-02      7.341984e-01      2.017995e-05
543462      3.631378e-01      2.003935e-02      5.601906e-02      4.285830e-01      5.176294e-02      8.045789e-02
116505      2.513739e-01      1.160407e-01      1.347864e-01      3.011456e-01      4.405570e-02      1.525978e-01
2600      1.040428e-01      2.100562e-01      2.223243e-01      1.016752e-01      7.495032e-02      2.869511e-01
20231      8.598096e-05      2.027843e-02      2.633159e-01      1.632424e-02      6.999712e-01      2.428055e-05
262086      3.296231e-01      6.618356e-02      6.635377e-02      3.575240e-01      3.927762e-02      1.410379e-01
705      1.065003e-04      1.514473e-01      2.134528e-01      5.937504e-02      5.755757e-01      4.262073e-05
116008      2.255932e-01      9.427689e-02      1.311362e-01      3.503724e-01      5.065449e-02      1.479668e-01
114757      2.821753e-01      3.620504e-02      1.354653e-01      3.919736e-01      6.705415e-02      8.712667e-02
55808      6.122945e-02      5.838692e-01      1.858593e-01      8.462276e-03      7.967499e-03      1.526123e-01
400487      3.584283e-01      9.365029e-02      1.077306e-01      1.967779e-01      3.788799e-02      2.055249e-01
121389      1.245810e-01      2.630981e-01      1.785560e-01      5.129617e-02      4.569390e-02      3.367748e-01
119818      3.220606e-01      6.345032e-02      7.210495e-02      3.302898e-01      4.029679e-02      1.717975e-01
29487      1.666579e-01      7.495656e-02      1.197911e-01      5.011668e-01      5.463329e-02      8.279439e-02
```

As we can see from the above results for both the training and the testing samples, for some of the records the probabilities for the actual response are higher as compared to other responses. However for many other records, we get incorrect predictions in terms of probabilities of the response. Hence, we can infer that this technique is not a very good estimator for predicting the outcome variable for this research question.

Naive Bayes

We applied the Naive Bayes technique for this question directly onto the complete dataset and also on the sub-sampled dataset. Following are the results for the first method:

(a)

Direct method:

The training and testing dataset were created in the proportion 75:25.

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	1026	2518	79	466	8	12	2	17
Closed with explanation	1898	82739	3240	7070	478	3033	519	178
Closed with monetary relief	430	19933	8211	2964	784	1095	108	37
Closed with non-monetary relief	532	24698	264	11415	0	0	248	5
Closed with relief	9	78	23	16	11	30	1	2

Closed without relief	102	1770	167	259	43	295	12	24
In progress	2	1	4	2	0	0	4	0
Untimely response	104	571	15	50	0	0	8	750

Accuracy=0.5856

Sub-sampling method:

For sub-sampling, we found that the class “Closed with relief” had the lowest count (say low) in the dataset. So, we randomly chose ‘low’ number of rows with each class and created the training and testing dataset in the proportion 75:25. Following are the results of this method:

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	644	144	70	139	19	13	2	17
Closed with explanation	73	186	32	12	52	69	519	178
Closed with monetary rel	95	154	622	113	292	144	108	37
Closed with non-monetary	117	228	57	855	0	0	248	5
Closed with relief	83	146	398	58	664	331	1	2
Closed without relief	280	476	152	103	336	790	12	24
In progress	0	0	0	0	0	0	4	0
Untimely response	0	0	0	0	0	0	8	750

Accuracy=0.4733

(b)

Direct Method, Laplace=2

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	982	1192	30	138	2	5	12	15
Closed with explanation	2254	92031	3741	7838	522	3260	558	139
Closed with monetary rel	450	18138	7922	2750	770	1061	94	22
Closed with non-monetary	209	19401	223	11358	0	0	227	5
Closed with relief	2	2	0	0	1	1	0	3
Closed without relief	28	397	46	66	29	138	3	10
In progress	1	8	8	1	0	0	5	0
Untimely response	177	1139	33	91	0	0	3	819

Accuracy=0.635

Sub-sampling Method, Laplace=3

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	70	150	65	107	19	13	0	0

	8							
Closed with explanation	72	201	35	13	48	70	0	0
Closed with monetary rel	96	157	633	115	295	144	0	0
Closed with non-monetary	66	220	62	889	0	0	0	0
Closed with relief	77	140	390	55	660	330	0	0
Closed without relief	27 3	466	146	101	341	790	0	0
In progress	0	0	0	0	0	0	0	0
Untimely response	0	0	0	0	0	0	0	0

Accuracy=0.4884

Direct Method, Laplace=3

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	91 1	1056	31	128	2	4	12	19
Closed with explanation	23 11	93065	3874	8025	540	3313	562	147
Closed with monetary rel	45 7	17517	7793	2672	759	1023	93	23
Closed with non-monetary	21 3	19180	222	11265	0	0	225	5
Closed with	1	2	1	0	1	1	0	3

relief								
Closed without relief	26	333	43	57	22	124	2	10
In progress	1	4	6	0	0	0	5	0
Untimely response	183	1151	33	95	0	0	3	806

Accuracy=0.639

Sub-sampling Method, Laplace=3

Predicted\Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	709	154	66	115	19	13	0	0
Closed with explanation	72	195	36	13	47	69	0	0
Closed with monetary rel	97	158	635	115	298	144	0	0
Closed with non-monetary	62	218	60	882	0	0	0	0
Closed with relief	76	140	388	54	657	330	0	0
Closed without relief	276	469	146	101	342	791	0	0
In progress	0	0	0	0	0	0	0	0
Untimely response	0	0	0	0	0	0	0	0

Accuracy=0.4869

NOTE:

Since the accuracy is deteriorating, in the other questions we decided to only apply the direct method.

Decision Trees and Random Forests

Here we have removed two attributes: Issue and State as they were increasing the number of levels to be handled by the decision tree and no model was being created.

(a)

Distribution of training data

Class	Proportion
Closed	0.023194012
Closed with explanation	0.746802005
Closed with monetary relief	0.068000626
Closed with non-monetary relief	0.127945574
Closed with relief	0.00779211
Closed without relief	0.026261709
In progress	0
Untimely response	3.96E-06

Distribution of testing data

Class	Proportion
Closed	0.023061283
Closed with explanation	0.747479251
Closed with monetary relief	0.069278971
Closed with non-monetary relief	0.12564505
Closed with relief	0.007681149

Closed without relief	0.026854296
In progress	0
Untimely response	0

(b)

NOTE: Since our dataset has >5000 classes, the tree created is massive and RStudio does not allow us to scroll to the top. Hence, it is impossible to figure out the root of the tree and explain a branch. Hence, explanation of certain if-then rules are not mentioned in any of the research questions.

Confusion matrix (Training Test)

Predicted/ Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	3200	8106	4	394	0	0	0	0
Closed with explanation	1539	368229	756	6309	10	3	0	0
Closed with monetary rel	181	32374	1249	505	4	1	0	0
Closed with non-monetary	129	51672	357	12403	2	0	0	0
Closed with relief	23	3841	30	14	19	5	0	0
Closed without relief	35	13144	16	8	19	30	0	0

In progress	0	0	0	0	0	0	0	0
Untimely response	0	2	0	0	0	0	0	0

Size=60

Error rate= 23.7%

Success rate= 76.3%

Confusion matrix (Testing Data)

Predicted/Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	1037	574	68	45	10	17	0	0
Closed with explanation	2726	122712	10998	17092	1264	4483	0	0
Closed with monetary rel	3	255	407	122	8	6	0	0
Closed with non-monetary	113	2181	174	3875	5	1	0	0
Closed with relief	0	3	6	0	4	8	0	0
Closed without relief	0	4	0	0	1	2	0	0
In progress	0	0	0	0	0	0	0	0
Untimely response	0	0	0	0	0	0	0	0

Accuracy=0.7612 Error rate=23.9%

Success rate=76.1%

(c)

Boosting, trials=10

Confusion matrix (Training data)

Predicted/Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	0	11369	1	334	0	0	0	0
Closed with explanation	0	375676	69	1101	0	0	0	0
Closed with monetary relief	0	33892	374	48	0	0	0	0
Closed with non-monetary	0	57813	274	6476	0	0	0	0
Closed with relief	0	3932	0	0	0	0	0	0
Closed without relief	0	13252	0	0	0	0	0	0
In progress	0	0	0	0	0	0	0	0
Untimely response	0	2	0	0	0	0	0	0

Statistics

Trial	Size	Errors
0	60	119483(23.7%)
1	13	128522(25.5%)
2	14	132673(26.3%)

3	24	137941(27.3%)
4	25	151721(30.1%)
5	16	127984(25.4%)
6	34	167825(33.3%)
7	14	129139(25.6%)
8	26	139444(27.6%)
9	7	122209(24.2%)
boost		122087(24.2%)

Error rate=24.2%

Success Rate=75.8%

Confusion Matrix (Testing Data)

Predicted/Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	0	0	0	0	0	0	0	0
Closed with explanation	3785	125333	11522	18991	1292	4517	0	0
Closed with monetary relief	0	24	120	99	0	0	0	0
Closed with non-monetary	94	372	11	2044	0	0	0	0
Closed with relief	0	0	0	0	0	0	0	0
Closed without relief	0	0	0	0	0	0	0	0
In progress	0	0	0	0	0	0	0	0

Untimely response	0	0	0	0	0	0	0	0
--------------------------	---	---	---	---	---	---	---	---

Accuracy=0.758

Error rate=24.2%

Success rate=75.8%

Boosting, trials=5

Confusion matrix (Training data)

Predicted/Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	3159	8181	10	354	0	0	0	0
Closed with explanation	1610	372086	687	2457	0	6	0	0
Closed with monetary relief	189	33030	930	162	0	3	0	0
Closed with non-monetary	124	55985	298	8154	0	2	0	0
Closed with relief	23	3880	13	0	0	16	0	0
Closed without relief	35	13185	14	0	0	18	0	0
In progress	0	0	0	0	0	0	0	0
Untimely response	0	2	0	0	0	0	0	0

Statistics:

Trial	Size	Errors
0	60	119483(23.7%)
1	13	128522(25.5%)
2	14	132673(26.3%)
3	25	158897(31.5%)
4	16	122281(24.2%)
boost		120266(23.8%) <<

Error rate=23.83%

Success rate= 76.17%

Confusion Matrix (Testing Data)

Predicted/Actual	Closed	Closed with explanation	Closed with monetary relief	Closed with non-monetary relief	Closed with relief	Closed without relief	In progress	Untimely response
Closed	1030	586	66	46	10	17	0	0
Closed with explanation	2753	124074	11195	18442	1273	4483	0	0
Closed with monetary relief	1	204	330	104	6	10	0	0
Closed with non-monetary	95	862	58	2542	0	0	0	0
Closed with relief	0	0	0	0	0	0	0	0
Closed without relief	0	3	4	0	3	7	0	0

In progress	0	0	0	0	0	0	0	0
Untimely response	0	0	0	0	0	0	0	0

Accuracy=0.7609

Error rate=23.91%

Success rate=76.09%

NOTE:

We tried boosting with trials=15. But the accuracy remained the same as boosting with trials=10 i.e. ~0.75.

(d)

We were unable to create a model by bagging or random forest. We got the error stating that the model cannot handle more than 58 categories. Even after sub-sampling and creating a dataset with just 34 classes, the model did not work.

Comparative Analysis:

For multinomial logistic regression, since the predicted output variable is in the form of probability of classifying the dependent variable for each record in the testing data set, into a specific class (probability the company's response, in this case), it is not possible to calculate accuracy of the model like other classification techniques where the class of categorical dependent variable is directly predicted by the model.

Naive Bayes	Naive Bayes, Laplace=2	Naive Bayes, Laplace=3	Decision Tree size=60	Decision Tree size=16, Boosting trials=5	Decision Tree size=7, Boosting trials=10
<u>Accuracy (Direct):</u> 0.5856 <u>Accuracy(Sub-sampling):</u> 0.4733	<u>Accuracy (Direct):</u> 0.635 <u>Accuracy (Sub-sampling):</u> 0.4884	<u>Accuracy (Direct):</u> 0.639 <u>Accuracy (Sub-sampling):</u> 0.4869	<u>Size:</u> 60 <u>Error Rate(trainin g):</u> 23.7% <u>Error Rate(testing):</u> 23.9%	<u>Error Rate(trainin g):</u> 24.2% <u>Error Rate(testing):</u> 24.2%	<u>Error Rate(trainin g):</u> 23.83% <u>Error Rate(testing):</u> 23.91%

For this research question, Decision Tree works the best since it has the lowest error rate compared to the other models we built.

Research Question 2:**Predict if a consumer would dispute the company's feedback or not**

Predictors: Product, Issue, state, Submitted.via, Company.response.to.consumer, Timely.response.

Outcome: Consumer.disputed.

Sampling rate: (Training:Testing)=75:25

Linear Regression:

In this question, the dependent variable is a nominal variable with 2 factors (Yes/No) and all the independent variables are multi-level nominal variables. Since the output variable is not continuous, linear regression cannot be applied for this question.

Logistic Regression:

Since the dependent variable is a nominal variable with only 2 levels, logistic regression can be applied in this case. The dependent and independent variables for our model are as follows:

Following is a snapshot of our model:

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	

* full model with coefficients of all factors for each of the variables can be found [here](#).

- The intercept value for our model is: **-1.711**
- The coefficients for every factor of each of the independent variables can be found in the excel file [here](#). The following are the variables which were statistically significant (<0.05):

Independent Variable	Factor	pvalue
Product	Credit Reporting	<2e-16
Product	Debt Collection	<2e-16
Product	Payday Loan	0.013
Issue	Can't Contact Lender	0.003
Issue	Can't Repay My Loan	0.00046
Issue	Communication Tactics	<2e-16
Issue	Credit Monitoring or Identity Protection	0.0044
Issue	Credit Reporting Company's Investigation	<2e-16
Issue	Deposit and Withdrawals	0.009
Issue	Disclosure Verification of Debt	1.49e-12
Issue	False Statements or Representation	0.0007
Issue	Improper Use of My Credit Report	<2e-16
Issue	Making/Receiving Payments, Sending Money	0.029
Issue	Problems caused by my funds being low	2.74E-09
Issue	Problems when you are unable to pay	2.61E-07
Issue	Repaying your loan	0.01
Company.response.to.consumer	Closed with monetary relief	< 2e-16
Company.response.to.consumer	Closed with non-monetary relief	< 2e-16
Company.response.to.consumer	Closed with relief	< 2e-16

sumer		
Company.response.to.con sumer	Closed without relief	3.95E-12
Timely.response.	Yes	3.49E-10

C. The log odds and odd ratios of the outcome variable (whether a complaint is disputed by a consumer or not) for each of the statistically significant independent variables are as follows:

Independent Variable	Factor	Log Odds	Odd Ratios
Product	Credit Reporting	-0.4276	6.520722
Product	Debt Collection	-0.307	7.356506
Product	Payday Loan	-0.363	6.955864
Issue	Can't Contact Lender	-0.4981	6.076842
Issue	Can't Repay My Loan	-0.3385	7.128388
Issue	Communication Tactics	-0.3522	7.031395
Issue	Credit Monitoring or Identity Protection	-0.1782	8.367751
Issue	Credit Reporting Company's Investigation	0.5141	16.72133
Issue	Deposit and Withdrawals	-0.0668	9.353823
Issue	Disclosure Verification of Debt	0.2591	12.95763
Issue	False Statements or Representation	0.1442	11.55115
Issue	Improper Use of My Credit Report	0.5572	17.45777

Issue	Making/Receiving Payments, Sending Money	-0.085	9.184113
Issue	Problems caused by my funds being low	-0.2065	8.134263
Issue	Problems when you are unable to pay	-0.2806	7.553304
Issue	Repaying your loan	-0.2561	7.740646
Company.response.to.consumer	Closed with monetary relief	-0.9175	3.995166
Company.response.to.consumer	Closed with non-monetary relief	-0.6594	5.171615
Company.response.to.consumer	Closed with relief	-0.6915	5.008243
Company.response.to.consumer	Closed without relief	0.2091	12.32568
Timely.response.	Yes	0.1665	11.81164

D. The following are the top 5 predictive features (Variable-Factor pair) based on coefficient values of the IV's:

Independent Variable	Factor
Issue	Improper use of my credit report
Issue	Credit reporting company's investigation
Issue	Disclosure verification of debt
Company.response.to.consumer	Closed without relief
Timely.response	Yes

E. The model trained on the training data set was then applied to the testing data set. The output of the model for testing data in the case of logistic regression is the probability of a customer disputing a company's response for each of the records in the testing data set. We appended the probability output of the model to the corresponding rows in the testing data set and extract the resulting data frame as a csv file. The file can be found [here](#).

Naive Bayes**(a)**

Following are the results we achieved for our Naive Bayes model:

Confusion Matrix

Predicted/Actual	No	Yes
No	131847	34604
Yes	1288	465

Accuracy=0.7866

(b)

Confusion Matrix, Laplace=2

Predicted/Actual	No	Yes
No	132502	34716
Yes	633	353

Accuracy= 0.7898

Confusion Matrix, Laplace=3

Predicted/Actual	No	Yes
No	132491	34717
Yes	644	354

Accuracy=0.7898

Decision Trees and Random Forests

Just like in Question 1, we have removed two attributes: Issue and State as they were increasing the number of levels to be handled by the decision tree and no model was being created.

(a)

Distribution of training data

Class	Proportion
No	0.7916225
Yes	0.2083775

Distribution of testing data

Class	Proportion
No	0.7911643
Yes	0.2088357

(b)

Confusion matrix (Training Set)

Predicted/Actual	No	Yes
No	399463	105150
Yes	0	0

Size=1

Error rate=20.8%

Success rate= 79.2%

Confusion matrix (Testing Set)

Predicted/Actual	No	Yes
No	133077	35127
Yes	0	0

Accuracy=0.7912

Error rate=20.9%

Success rate=79.1%

(c)

Boosting, trials=5

Boosting was abandoned after the first trial due to very few classifiers. The results for the first trial were exactly the same as mentioned in (b).

(d)

We were unable to apply bagging and random forests technique on this dataset due to the same reason as mentioned in Question 1.

Comparative Analysis:

Logistic Regression	Naive Bayes	Naive Bayes, Laplace = 2	Naive Bayes, Laplace = 3	Decision Trees
-	Accuracy: 0.7866	Accuracy: 0.7898	Accuracy: 0.7898	Accuracy: 0.7912

For logistic regression, since the predicted output variable is in the form of probability of classifying the dependent variable for each record in the testing data set, into a specific class (probability that a company's response to a complaint will be disputed by the consumer, in this case), it is not possible to calculate accuracy of the model like other classification techniques where the class of categorical dependent variable is directly predicted by the model.

Research Question 3:**Predict the medium through which a complaint will be received**

We are trying to predict the medium via which complaints may be received by the companies. We have a total of 6 mediums (Fax, Postal Mail, Web, Phone, Email and Referral) in our dataset. Our dependent variable is nominal with 6 levels, whereas all our independent variables are also multiclass nominal variables.

Predictors: Product, Company, Region(derived from the State), Timely.response.

Outcome: Submitted.via

Sampling rate: (Training:Testing)=75:25

Linear Regression

We cannot use Linear Regression for answering this question because Linear regression requires the dependent variable to be continuous. In our case, the dependent variable is nominal with 6 levels, hence Linear Regression technique cannot be applied.

(Multinomial) Logistic Regression

Logistic Regression can be used only when we have binary outcome variables. Since our outcome variable is nominal with multiple levels, we use a different version of Logistic Regression called as Multinomial Logistic Regression.

Multinomial Logistic Regression is the linear regression analysis to conduct when the dependent variable is nominal with more than two levels.

Additional Data Cleaning

Since 'Company' is one of the predictors in our model, and since it contains more than 4000 unique values, we wanted to reduce the number of levels for this predictor. Using all the levels for this predictor variable was causing memory issues in R while trying to build the model. Hence we decided to reduce the levels by taking the top 10 companies based on the number of complaints received.

We then took equal number of records for each Medium and created a new sample, so that each Medium has equal weightage in the new sample. We then created training and testing samples from this new dataset, and used the training sample for building our model.

Below is the output table for our multinomial logistic regression model:

Coefficients					
	Fax	Phone	Postal mail	Referral	Web
(Intercept)	3.2667390	3.8195074	1.9445406	3.0280617	0.6240287

ProductConsumer Loan	0.6125018	0.1783681	0.7817142	-0.1412845	0.726971 2
ProductCredit card	-0.1877197	-0.589320 0	0.2484799	-0.6213384	0.340721 8
ProductCredit reporting	1.0766695	-1.937273 4	1.8990043	-0.4222901	1.672868 4
ProductDebt collection	5.960214	5.305073	6.374081	5.392382	6.686203
ProductMoney transfers	0.6220733	0.3283795	0.8601145	0.2468963	1.356348 8
ProductMortgage	-0.2738661	-1.524377 2	-0.7262500	-0.6737521	-0.452725 1
ProductOther financial service	1.3162354	0.2109142	1.7614861	1.6633983	-3.968042 3
ProductPayday loan	-1.1207386	4.8119479	-0.9066705	-1.6140049	-1.023950 8
ProductPrepaid card	0.9503677	1.0405603	1.5176877	0.5111405	-3.487960 8
ProductStudent loan	-1.3799662	-1.685851 7	-1.2108084	-1.1055878	0.448669 8
ProductVirtual currency	0	0	0	0	0
CompanyCapital One	-0.6002952	-0.419474 8	-0.5387811	-0.8719237	-0.507219 5
CompanyCitibank	-0.1237264 5	-0.157029 58	0.07141042	-0.26092836	-0.205618 22
CompanyEquifax	0.2346229	0.5393751	0.6039668	0.5644665	-0.231360 1
CompanyExperian	-0.3474079 2	-0.263413 92	0.21474530	0.07542162	-0.650993 76
CompanyJPMorgan Chase & Co.	-0.0453020 7	-0.113383 44	0.02741001	-0.05017897	-0.203767 51

CompanyNavient Solutions, LLC.	4.478214	2.150386	2.638092	-3.361793	-4.944054
CompanyOcwen	0.7069878	0.5923445	0.9317224	0.4298421	0.7282926
CompanyTransUnion Intermediate Holdings, Inc.	-0.49047975	-0.09434054	-0.02084690	-0.00783501	-1.27497601
CompanyWells Fargo & Company	0.19020170	0.29353658	0.15097901	0.12582857	0.09705934
RegionNorth East	-1.2775899	-0.8570490	-0.9950595	-0.6840810	-0.8768950
RegionSouth	-1.1521000	-0.7835652	-0.5055920	-0.6256602	-0.7413581
RegionWest	-1.3586371	-1.1471116	-1.0610296	-0.8507107	-0.8659320
Timely.response.Yes	1.458256	1.703406	1.833005	1.837840	3.718782
Std. Errors					
	Fax	Phone	Postal mail	Referral	Web
(Intercept)	0.7215132	0.7212928	0.7548711	0.7261596	0.8649107
ProductConsumer Loan	1.057950	1.052692	1.062596	1.059836	1.058437
ProductCredit card	0.4407680	0.4366835	0.4436194	0.4390206	0.4397546
ProductCredit reporting	2.310675	2.338246	2.307941	2.324340	2.306659
ProductDebt collection	11.55864	11.55839	11.55876	11.55861	11.55842
ProductMoney transfers	4.124769	4.107419	4.135348	4.118190	4.109298
ProductMortgage	0.3465066	0.3440285	0.3511046	0.3446121	0.3471586

ProductOther financial service	8.393350	8.392931	8.393720	8.374529	13.365147
ProductPayday loan	17.467841	7.914007	17.476459	16.757198	17.376699
ProductPrepaid card	11.01237	10.98102	11.01181	11.01082	14.86806
ProductStudent loan	1.097851	1.069163	1.121191	1.068470	1.056621
ProductVirtual currency	Nan	Nan	.Nan	Nan	4.72004e-15
CompanyCapital One	0.4928257	0.4879023	0.4964469	0.4925736	0.4900811
CompanyCitibank	0.4019138	0.4005005	0.4062808	0.4011602	0.4018015
CompanyEquifax	2.372865	2.401294	2.369901	2.386617	2.368864
CompanyExperian	2.337812	2.366475	2.334760	2.351485	2.333689
CompanyJPMorgan Chase & Co.	0.3380256	0.3371619	0.3435607	0.3368056	0.3387164
CompanyNavient Solutions, LLC.	9.307108	9.324852	9.357150	14.120347	17.581180
CompanyOcwen	0.4420942	0.4430681	0.4481420	0.4420866	0.4432083
CompanyTransUnion Intermediate Holdings, Inc.	2.343762	2.372677	2.340664	2.357616	2.339906
CompanyWells Fargo & Company	0.3241183	0.3230622	0.3300894	0.3230854	0.3248179
RegionNorth East	0.4393021	0.4397420	0.4414903	0.4397955	0.4401134
RegionSouth	0.4200655	0.4207815	0.4216151	0.4208033	0.4209333

RegionWest	0.4250356	0.4258559	0.4271047	0.4257128	0.425809 3
Timely.response.Yes	0.5523404	0.5534213	0.5922948	0.5587640	0.728896 5
Residual Deviance: 56674.8					
AIC: 56914.8					

In this case, **'Email'** is set as the base reference medium against which all other mediums are compared.

Similarly, **'Bank account or service'** is set as the base reference against which all other Products are compared.

'Midwest' region is set as the base region while **'Bank of America'** is set as the base Company for all corresponding comparisons.

To give a brief idea about how we interpret these results,

The log odds of a complaint being received by 'Fax' versus by 'Email' will increase by 1.057 while moving from Product 'Bank Account or service' to Product 'Consumer Loan'.

The log odds of a complaint being received by 'Web' versus by 'Email' will increase by 0.49 if the institution is 'Capital One' as compared to 'Bank of America'.

The log odds of a complaint being received by 'Postal Mail' versus by 'Email' will increase by 0.42 if the complaint originates from the 'South' region as compared to the 'Midwest' region.

The exponentiated regression coefficients are relative risk ratios for a unit change in the predictor variable. We can exponentiate the coefficients from our model to see these risk ratios as follows.


```
> 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 0.5486497 0.8836215 1.2644318 0.7065171 0.9557087
Phone 0.6573920 0.8546788 1.7149349 0.7684238 0.8928083
Postal mail 0.5834590 1.0740219 1.8293612 1.2395461 1.0277891
Referral 0.4181464 0.7703361 1.7585093 1.0783387 0.9510592
web 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
```

The relative risk ratio while moving from Product 'Bank Account or service' to 'PayDay Loan' is 122.97 for complaints received via 'Phone' versus 'Email'.

The relative risk ratio while moving from 'Midwest' region to 'West' region is 0.257 for complaints received via 'Fax' versus 'Email'.

To better understand the model, we have calculated the fitted values for some of our observations in the training dataset.

Original Dataset values are as seen below






50757	Credit reporting	TransUnion Intermediate Holdings, Inc.	West	Postal mail	Yes
209550	Mortgage	JPMorgan Chase & Co.	North East	Fax	Yes
8556	Credit reporting	Experian	North East	Web	Yes
11063	Mortgage	Bank of America	North East	Phone	Yes
39019	Credit reporting	Experian	West	Postal mail	Yes
6425	Credit reporting	Experian	West	Web	Yes
329514	Credit reporting	TransUnion Intermediate Holdings, Inc.	West	Fax	Yes
51463	Credit reporting	Equifax	South	Referral	Yes
7088	Mortgage	Ocwen	North East	Web	Yes
13276	Mortgage	Bank of America	West	Web	Yes

Fitted values are as seen below

```
> head(pp<-fitted(mlr),n=10)
      Email      Fax      Phone Postal mail Referral      web
50757 0.004053902 0.2110780 0.04226113 0.40115735 0.1461664 0.1952833
209550 0.009913195 0.2263548 0.20478697 0.07963975 0.3147428 0.1645624
8556 0.002882549 0.1877933 0.03391482 0.38564268 0.1334367 0.2563300
11063 0.009125537 0.2180264 0.21114894 0.07132974 0.3046443 0.1857251
39019 0.003084372 0.1852984 0.02715232 0.38629993 0.1208645 0.2773004
6425 0.003084372 0.1852984 0.02715232 0.38629993 0.1208645 0.2773004
329514 0.004053902 0.2110780 0.04226113 0.40115735 0.1461664 0.1952833
51463 0.001391744 0.1839650 0.03933079 0.44830912 0.1113831 0.2156203
7088 0.004887449 0.2367956 0.20448648 0.09699209 0.2507812 0.2060572
13276 0.010362893 0.2283144 0.17940680 0.07583030 0.2928527 0.2132329
```

We then used our model to predict the probabilities for our testing sample.

Actual values for the test sample

	Product 	Company 	Region 	Submitted.via 	Timely.response 
	All	All	All	All	All
12641	Mortgage	Wells Fargo & Company	West	Referral	Yes
150614	Bank account or service	Bank of America	West	Phone	Yes
102112	Consumer Loan	Wells Fargo & Company	South	Phone	Yes
20550	Mortgage	Citibank	West	Referral	Yes
60399	Mortgage	Bank of America	South	Referral	Yes
426632	Mortgage	Bank of America	Midwest	Fax	Yes
16410	Mortgage	Citibank	South	Referral	Yes
121851	Credit card	Citibank	West	Email	Yes
263938	Mortgage	Bank of America	Midwest	Fax	Yes
22397	Credit card	Bank of America	West	Phone	Yes
422897	Mortgage	JPMorgan Chase & Co.	North East	Fax	Yes
30209	Credit card	Citibank	West	Referral	Yes
339943	Credit reporting	Equifax	Midwest	Fax	Yes
48828	Bank account or service	Wells Fargo & Company	West	Referral	Yes

Predicted values for the test sample

```
> predict(mlr,newdata=test_sample,"probs")
      Email      Fax      Phone  Postal mail      Referral      web
12641 8.764306e-03 0.233546483 0.20349642 0.0745845135 0.2808875426 0.1987207361
150614 4.708179e-03 0.136409226 0.37431697 0.0712231380 0.2609925446 0.1523499419
102112 2.071590e-03 0.164665363 0.37977133 0.1387902493 0.1416135296 0.1730879359
20550 1.224809e-02 0.238444285 0.18122964 0.0962594354 0.2666348587 0.2051836918
60399 8.036027e-03 0.217666477 0.20011769 0.1024770482 0.2844119691 0.1872907901
426632 3.604063e-03 0.308953574 0.19648732 0.0761997350 0.2384622549 0.1762930498
16410 9.420411e-03 0.225468720 0.20050115 0.1290233232 0.2568364078 0.1787499902
121851 7.106204e-03 0.150788810 0.26784819 0.1480240832 0.1630231250 0.2632095892
263938 3.604063e-03 0.308953574 0.19648732 0.0761997350 0.2384622549 0.1762930498
22397 6.105565e-03 0.146619256 0.26926132 0.1184151888 0.1818264853 0.2777721819
422897 9.913195e-03 0.226354835 0.20478697 0.0796397472 0.3147428363 0.1645624207
30209 7.106204e-03 0.150788810 0.26784819 0.1480240832 0.1630231250 0.2632095892
339943 6.711193e-04 0.280754096 0.04152130 0.3584211898 0.1004107581 0.2182215349
48828 3.864195e-03 0.135410818 0.41202816 0.0679824046 0.2429298166 0.1377846050
```

When we compare the actual values for the testing sample versus the predicted values for the testing sample, we can see that our model gives a higher probability value for the medium via which the complaint was actually received. For example, as shown above for record 12641, the actual medium is 'Referral' and the predicted probability for this medium is calculated as 0.28 which is the highest among all other mediums. Hence our model has a good amount of predictive power for predicting the multiclass outcome variable.

Naive Bayes

(a)

Following are the results we achieved for our Naive Bayes model:

Predicted/Actual	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	0	1	1	6	3
Fax	0	7	15	10	27	112
Phone	1	47	606	169	389	2899
Postal Mail	3	126	150	1723	522	4767
Referral	48	1190	5208	3268	20659	28348
Web	35	1190	6417	6255	10258	83902

Accuracy=0.5993

(b)

Confusion Matrix, Laplace=2

Predicted/Actual	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	2	2	0	2	8
Fax	0	4	0	3	0	12
Phone	0	29	449	69	183	984
Postal Mail	0	46	35	670	181	1585
Referral	48	1160	5082	3159	20381	27075
Web	39	1319	6829	7525	11114	90367

Accuracy=0.6272

Confusion Matrix, Laplace=3

Predicted/Actual	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	2	2	0	2	11
Fax	0	3	0	2	0	11
Phone	0	27	430	62	163	935
Postal Mail	0	25	25	392	128	946
Referral	48	1149	5039	3119	20234	26721
Web	39	1354	6901	7851	11334	91407

Accuracy=0.6305

Decision Trees and Random Forests

(a)

Distribution of Training Data

Classes	Proportion
Email	0.00050732
Fax	0.014335739
Phone	0.070035849
Postal Mail	0.065057777
Referral	0.182652845
Web	0.667410471

Distribution of Testing Data

Classes	Proportion
Email	0.000499394
Fax	0.014738056
Phone	0.071734323
Postal Mail	0.065301658
Referral	0.182564029
Web	0.665162541

(b)

Confusion matrix (Training data)

Actual/Predicted	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	0	0	0	33	223
Fax	0	0	0	0	175	7059

Phone	0	0	0	0	1794	33547
Postal Mail	0	0	0	0	575	32254
Referral	0	0	0	0	8355	83814
Web	0	0	0	0	6608	330176

Size:9

Error rate: 32.9%

Success rate: 67.1%

Confusion matrix (Testing data)

Predicted/Actual	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	0	0	0	0	0
Fax	0	0	0	0	0	0
Phone	0	0	0	0	0	0
Postal Mail	0	0	0	0	0	0
Referral	11	52	591	201	2769	2153
Web	73	2427	11475	10783	27939	109730

Accuracy:0.6688

Error rate: 33.12%

Success rate: 66.88%

(c)

Boosting, trials=5

Confusion Matrix (Training data)

Actual/Predicted	Email	Fax	Phone	Postal Mail	Referral	Web
-------------------------	--------------	------------	--------------	--------------------	-----------------	------------

Email	0	0	0	0	0	0
Fax	0	0	0	0	0	0
Phone	0	0	0	0	0	0
Postal Mail	0	0	0	0	0	0
Referral	11	52	591	201	2769	2153
Web	73	2427	11475	10783	27939	109730

Error rate: 33%

Success rate: 67%

NOTE: The boosting stopped at trial=4 as the previous classifier had very less frequency.

Statistics

Trial	Size	Errors
0	9	166082(32.9%)
1	5	177745(35.2%)
2	3	168715(33.4%)
3	6	186238(36.9%)
4	1	167829(33.3%)
boost		166422(33.0%)

Confusion Matrix (Testing data)

Actual/Predicted	Email	Fax	Phone	Postal Mail	Referral	Web
Email	0	0	0	0	0	0
Fax	0	0	0	0	0	0

Phone	0	0	0	0	0	0
Postal Mail	0	0	0	0	0	0
Referral	6	32	412	120	1616	1101
Web	78	2447	11654	10864	29092	110782

Accuracy: 0.6682

Error rate: 33.2%

Success rate: 66.8%

Boosting, trials=10

The boosting truncated forcibly at 4 trails since the last classifier was very inaccurate. The error rate was hence similar to those mentioned above.

(d)

We were unable to apply bagging and random forests technique on this dataset due to the same reason as mentioned in Question 1.

Comparative Analyses

Multinomial Logistic Regression	Naive Bayes	Naive Bayes, Laplace = 2	Naive Bayes, Laplace = 3
-	Accuracy: 0.5993	Accuracy: 0.6272	Accuracy: 0.6305

For multinomial logistic regression, since the predicted output variable is in the form of probability of classifying the dependent variable for each record in the testing data set, into a specific class (probability that a complaint will be received via a particular medium, in this case), it is not possible to calculate accuracy of the model like other classification techniques where the class of categorical dependent variable is directly predicted by the model.

Research Question 4:**Predict the geographical region in United States where the complaint originated****Predictors:** Product, Submitted.via, Timely.response., Consumer.disputed., Company.response.to.consumer, Company**Outcome:** Region (4 levels)**Sampling rate:** (Training:Testing)=75:25**Linear Regression**

We cannot use Linear Regression for answering this question because Linear regression requires the dependent variable to be continuous. In our case, the dependent variable is nominal with 4 levels(Midwest/ Northeast/ South/ West), hence Linear Regression technique cannot be applied.

(Multinomial) Logistic Regression

We cannot use Logistic regression as the outcome variable isn't binary and it has 4 levels as defined above. Since our outcome variable is a multiclass nominal variable, we use multinomial logistic regression to predict the region from which complaints originate.

The results obtained on running the multinomial logistic regression are as follows

```
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					

```

Std. Errors:
(Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection ProductMoney transfers ProductMortgage
North East 0.4505117 0.06994931 0.03722017 0.1356297 0.05787533 0.2106048 0.03242414
South 0.4998681 0.06773931 0.03916861 0.1322445 0.05758299 0.1996986 0.03344908
West 0.4688118 0.06936999 0.03806139 0.1370658 0.05696427 0.1861809 0.03245660
ProductOther financial service ProductPayday loan ProductPrepaid card ProductStudent loan CompanyCapital One CompanyCitibank CompanyEquifax
North East 0.3939048 0.5927392 0.2797135 0.08652407 0.03910731 0.03080396 0.1353655
South 0.3510574 0.8691344 0.2843805 0.10560018 0.04168326 0.03375883 0.1316009
West 0.3847756 0.6312553 0.2380966 0.09742496 0.04143372 0.03263651 0.1365862
CompanyExperian CompanyNavient Solutions, LLC. CompanyOwen CompanyTransunion Intermediate Holdings, Inc. CompanyWells Fargo & Company
North East 0.1354581 0.1861988 0.03367436 0.1357642 0.02734706
South 0.1318445 0.2162330 0.03537216 0.1320352 0.02820223
West 0.1366764 0.1841820 0.03363979 0.1370427 0.02683461
Submitted.viaFax Submitted.viaPhone Submitted.viaPostal mail Submitted.viaReferral Submitted.viaWeb Timely.response.Yes
North East 0.4395183 0.4365533 0.4363967 0.4357395 0.4355567 0.08626725
South 0.4889693 0.4865613 0.4861303 0.4857271 0.4855082 0.08772253
West 0.4592637 0.4568399 0.4565342 0.4559003 0.4557005 0.07966823
Company.response.to.consumerClosed with explanation Company.response.to.consumerClosed with monetary relief
North East 0.07803002 0.08448601
South 0.08181191 0.08867289
West 0.07790531 0.08485791
Company.response.to.consumerClosed with non-monetary relief Company.response.to.consumerClosed with relief
North East 0.08054779 0.1151047
South 0.08409056 0.1319560
West 0.08042115 0.1172073
Company.response.to.consumerClosed without relief Consumer.disputed.Yes
North East 0.08843006 0.02002058
South 0.09463735 0.02005134
West 0.08704359 0.01970112

Residual Deviance: 354383.3
AIC: 354569.3

```

In this case, **'Email'** is set as the base reference medium against which all other mediums are compared.

Similarly, **'Bank account or service'** is set as the base reference against which all other Products are compared.

'Midwest' region is set as the base region while **'Bank of America'** is set as the base Company for all corresponding comparisons.

To give a brief idea about how we interpret these results,

The log odds of a complaint originating from 'Northeast' in comparison to 'Midwest' will increase by 0.538 while moving from Product 'Bank Account or service' to Product 'Payday Loan'.

The log odds of a complaint originating from 'South' in comparison to 'Midwest' will decrease by 3.63 if the institution is 'Navient Solutions LLC' as compared to 'Bank of America'.

Actual values for the training dataset

	Product	Company	Region	Submitted.via	Company.response.to.consumer	Timely.response	Consumer.disputed
277347	Credit reporting	Experian	North East	Web	Closed with explanation	Yes	No
342369	Credit reporting	Experian	West	Web	Closed with explanation	Yes	Yes
639962	Credit reporting	Equifax	Midwest	Web	Closed with explanation	Yes	No
264430	Mortgage	Bank of America	West	Web	Closed with explanation	Yes	Yes
152034	Mortgage	Bank of America	West	Web	Closed with explanation	Yes	No
291393	Credit reporting	TransUnion Intermediate Holdings, Inc.	Midwest	Web	Closed with explanation	Yes	No
182758	Mortgage	Ocwen	North East	Referral	Closed with non-monetary relief	Yes	No
41362	Credit card	Citibank	South	Web	Closed with non-monetary relief	Yes	No
533562	Credit reporting	Equifax	Midwest	Web	Closed with non-monetary relief	Yes	No
232532	Credit card	Citibank	North East	Web	Closed with monetary relief	Yes	No
730379	Student loan	Navient Solutions, LLC.	Midwest	Web	Closed with explanation	Yes	No
548350	Bank account or service	Bank of America	Midwest	Phone	Closed with monetary relief	Yes	No
89140	Credit reporting	Equifax	South	Web	Closed with explanation	Yes	No
98407	Credit card	Bank of America	North East	Web	Closed with monetary relief	Yes	No
239512	Credit reporting	Equifax	South	Web	Closed with explanation	Yes	No

Fitted values for the training dataset

```
> head(pp <- fitted(mlr),n=20)
      Midwest North East      South      West
277347 0.2406253 0.21199944 0.29598717 0.2513881
342369 0.2353653 0.20461507 0.29049045 0.2695292
639962 0.2661401 0.19208671 0.31774890 0.2240242
264430 0.2281635 0.25126784 0.22297296 0.2975957
152034 0.2336466 0.26076455 0.22756621 0.2780226
291393 0.2968415 0.19998454 0.28521912 0.2179549
182758 0.2119959 0.32971595 0.20082857 0.2574595
41362  0.2936832 0.30917735 0.18406462 0.2130748
533562 0.2712856 0.19821861 0.30355030 0.2269455
232532 0.3006973 0.29904737 0.19206037 0.2081950
730379 0.9499905 0.02009654 0.01225636 0.0176566
548350 0.1902147 0.31208866 0.25526266 0.2424340
89140  0.2661401 0.19208671 0.31774890 0.2240242
98407  0.2394393 0.25628318 0.22537647 0.2789010
239512 0.2661401 0.19208671 0.31774890 0.2240242
```

The exponentiated regression coefficients are relative risk ratios for a unit change in the predictor variable. We can exponentiate the coefficients from our model to see these risk ratios as follows.

```

> exp(coef(m1r))
(Intercept) ProductConsumer Loan ProductCredit card ProductCredit reporting ProductDebt collection ProductMoney transfers ProductMortgage
North East 3.022125 0.5534784 0.7523231 0.8809841 0.6724775 0.6870929 0.7502074
South 1.936893 0.7172109 0.6405252 1.0398251 0.8444801 0.8936534 0.6330216
West 3.031188 0.5833114 0.6965785 0.7441916 0.8460932 1.1595594 0.6748912
ProductOther financial service ProductPayday loan ProductPrepaid card ProductStudent loan CompanyCapital One CompanyCitibank CompanyEquifax
North East 0.4288011 1.7139758 0.4219583 0.6577066 0.9447332 0.9291499 0.5506950
South 0.7473439 0.3689984 0.5129768 0.3175198 0.7872414 0.6785707 0.7462489
West 0.4907500 1.1875018 0.9697718 0.3431452 0.5992691 0.5944098 0.6415243
CompanyExperian Companynavient Solutions, LLC, CompanyOcwen CompanyTransUnion Intermediate Holdings, Inc. Companywells Fargo & Company
North East 0.6722297 0.02162032 0.9933607 0.5140390 0.6005707 0.5140390 1.023736
South 0.7688501 0.02640838 0.9912194 0.6005707 0.5140390 0.6005707 1.140436
West 0.7962179 0.03072014 0.9706115 0.5595907 0.5595907 0.5595907 1.162859
Submitted.viaFax Submitted.viaPhone Submitted.viaPostal mail Submitted.viaReferral Submitted.viaWeb Timely.response.Yes
North East 0.2242626 0.3698295 0.3693340 0.4443994 0.3206917 1.742692
South 0.4015449 0.5163800 0.8208753 0.5920424 0.5654615 1.497918
West 0.3630094 0.3869709 0.5030289 0.5371786 0.5077071 1.359851
Company.response.to.consumerClosed with explanation Company.response.to.consumerClosed with monetary relief
North East 0.8808194 0.8423634
South 0.9378488 0.8957361
West 0.8424989 0.7990374
Company.response.to.consumerClosed with non-monetary relief Company.response.to.consumerClosed with relief
North East 0.8916975 0.9634280
South 0.8789476 0.6332920
West 0.8372967 0.9734308
Company.response.to.consumerClosed without relief Consumer.disputed.Yes
North East 0.7621277 0.9867376
South 0.6613994 1.0033622
West 1.1281116 1.0961243

```

The relative risk ratio while moving from Product 'Bank Account or service' to 'Debt Collection' is 0.846 for complaints received from region 'West' in comparison to those received from 'Midwest'.

Actual values for test sample

	Product	Company	Region	Submitted.via	Company.response.to.consumer	Timely.response.	Consumer.disputed.
441832	Credit reporting	Experian	North East	Web	Closed with explanation	Yes	No
14405	Mortgage	Bank of America	South	Web	Closed with explanation	Yes	No
283790	Bank account or service	Wells Fargo & Company	North East	Web	Closed with explanation	Yes	Yes
373632	Debt collection	Capital One	Midwest	Web	Closed with explanation	Yes	No
284487	Mortgage	Wells Fargo & Company	Midwest	Referral	Closed with explanation	Yes	No
202888	Consumer Loan	Bank of America	South	Web	Closed with explanation	Yes	Yes
256152	Credit card	Citibank	North East	Web	Closed without relief	Yes	No
227023	Credit card	Citibank	South	Web	Closed with explanation	Yes	No
116685	Mortgage	Bank of America	South	Referral	Closed with non-monetary relief	No	No
171593	Mortgage	Ocwen	West	Web	Closed with explanation	Yes	No
36720	Credit reporting	TransUnion Intermediate Holdings, Inc.	West	Web	Closed with non-monetary relief	Yes	No
25812	Mortgage	Ocwen	South	Web	Closed with explanation	Yes	No
73939	Credit reporting	TransUnion Intermediate Holdings, Inc.	South	Web	Closed with explanation	Yes	No
142538	Bank account or service	Wells Fargo & Company	Midwest	Web	Closed with explanation	Yes	Yes
266731	Mortgage	Bank of America	West	Referral	Closed with explanation	Yes	No

Predicted values for test sample


```
> predict(mlr,newdata=test_sample,"probs")
      Midwest North East      South      West
441832 0.24062527 0.21199944 0.295987174 0.25138811
14405  0.23364665 0.26076455 0.227566205 0.27802259
283790 0.15359232 0.23081656 0.270413235 0.34517788
373632 0.25893319 0.24472721 0.264859014 0.23148059
284487 0.19192717 0.30387806 0.223205792 0.28098898
202888 0.24709568 0.20075904 0.273589502 0.27855578
256152 0.29847623 0.26856439 0.141192935 0.29176644
227023 0.29080807 0.30241565 0.194476629 0.21229965
116685 0.28930329 0.25991908 0.184583731 0.26619390
171593 0.23646056 0.26215292 0.228284650 0.27310187
36720  0.30191025 0.20591140 0.271870520 0.22030783
25812  0.23646056 0.26215292 0.228284650 0.27310187
73939  0.29684146 0.19998454 0.285219125 0.21795487
142538 0.15359232 0.23081656 0.270413235 0.34517788
266731 0.20723890 0.32051326 0.211333940 0.26091390
```

When we compare the actual values for the test sample versus the predicted values for the test sample, we can see that our model fares very poorly. Out of the 15 records compared above, we found only one record for which we get the predicted highest probability for a region matches the actual region from which the complaint originated. Hence, we can infer that this technique does not provide a good predictive model for answering this research question.

Naive Bayes

(a)

Following are the results of our Naive Bayes model:

Confusion Matrix

Predicted/Actual	MidWest	North East	South	West
MidWest	3118	1494	3138	2053
North East	6070	12049	12780	9040
South	13244	15652	41484	20893
West	6067	8495	14739	13426

Accuracy=0.3814

(b)

Confusion Matrix, Laplace=2

Predicted/Actual	MidWest	North East	South	West
MidWest	3224	755	1647	1048
North East	5631	12138	11860	8215
South	14012	16737	45172	21655
West	5632	8060	13462	14494

Accuracy=0.4083

Confusion Matrix, Laplace=3

Predicted/Actual	MidWest	North East	South	West
MidWest	3091	675	1502	950
North East	5571	12012	11727	8119
South	14184	16947	45471	21853
West	5653	8056	13441	14490

Accuracy=0.4085

Decision Trees

(a)

Proportion Distribution of Training data

Classes	Proportion
MidWest	0.1545652
North East	0.206802
South	0.3905046
West	0.2481282

Proportion Distribution of Testing data

Classes	Proportion
MidWest	0.1556223
North East	0.2061367
South	0.3921979
West	0.2460431

One of the subtrees:

SubTree [S35]

Company.response.to.consumer = Closed with non-monetary relief: Midwest (25/13)

Company.response.to.consumer in {Closed with explanation,Closed with relief,

: Closed without relief}:

...Submitted.via in {Email,Phone,Referral,Web}: North East (2565/1639)

Submitted.via = Postal mail: South (225/126)

Submitted.via = Fax:

...Company.response.to.consumer = Closed with relief: Midwest (0)

Company.response.to.consumer = Closed without relief: West (3/1)

Company.response.to.consumer = Closed with explanation:

...Consumer.disputed. = No: Midwest (57/30)

Consumer.disputed. = Yes: South (13/7)

Explanation:

Since we could not get to the root of the tree in Rstudio, we have taken a subtree and tried to explain the if-else rules in just that subtree:

- If the company manages to closed the complaint without monetary issue, the complaint originated from MidWest region.
- If the company closes the complaint with an explanation or with/without relief and
 - If the complaint was submitted via email/phone/referral/web, the complaint originated from North East.
 - The complaint is from South if it was submitted via Postal mail.
- If the complaint was submitted via fax and was closed with relief, it must have originated from MidWest.
- If the complaint was submitted via fax and was closed without relief, it must have originated from West.
- If the complaint was submitted via fax, was closed with explanation without being disputed by the consumer, it must have originated from MidWest.
- If the complaint was submitted via fax, was closed with explanation and disputed by the consumer, it must have originated from South.

(b)

Confusion Matrix (training data)

Predicted/Actual	MidWest	North East	South	West
MidWest	8479	3517	63269	4422
North East	1275	15661	85332	4350
South	3060	6269	185141	6857
West	1839	3744	105901	16440

Size=327**Error rate= 56.2%****Success rate=43.8%****Confusion Matrix (testing data)**

Predicted/Actual	MidWest	North East	South	West
MidWest	2798	483	1097	683
North East	1183	4922	2257	1315
South	21233	28606	61647	35126
West	1530	1414	2399	5159

Accuracy=0.4337**Error rate= 56.6%****Success rate=43.4%**

(c)

Boosting, trials=5, 10

The decision tree has no effect of boosting at all. It truncates at the first trial and hence gives the results mentioned in (b)

(d)

We were unable to apply bagging and random forests technique on this dataset due to the same reason as mentioned in Question 1.

Research Question 5:

Predict the number of complaints based on correlation between number of complaints and sum total of assets in the financial institutions

Linear Regression

In this question, the number of complaints is the dependent variable and the sum total of assets of an financial institution is the independent variable. The correlation is based on the assumption that as the number of assets possessed by a bank increase (assets could increase due to two possible factors, one could be as time goes by the assets possessed by people in the financial institution increase, or as the number of people engaging in an financial institution increases), it becomes difficult to manage, and this could result in increase in the number of complaints. We have time series data which has been grouped quarter wise.

Part A : Linear Regression Parameters

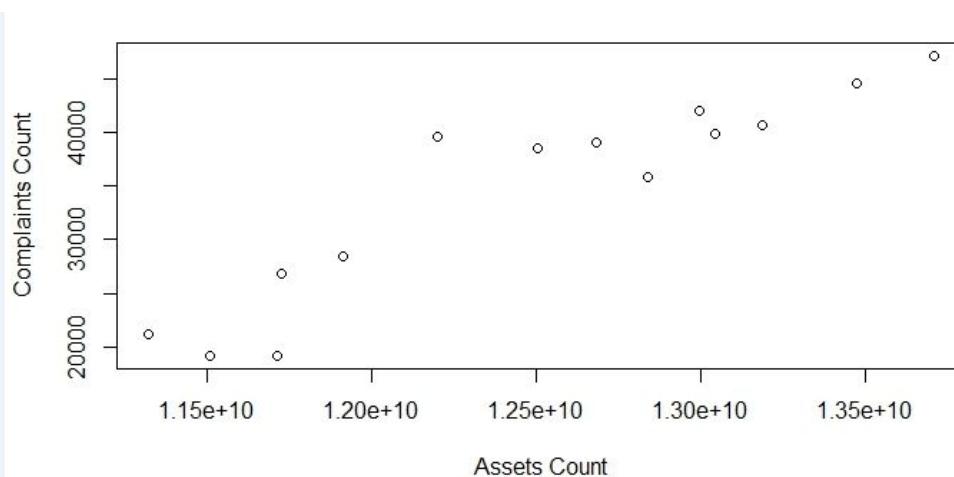
Predictor : no of complaints in a given quarter

Outcome: total assets possessed by a financial institution in that given quarter

Both the DV and IV are **continuous variables**. The scale of measurement for both the variables is **ratio**.

Since both the DV and IV are continuous variables, we can apply Linear regression model.

Plot of DV vs IV



The consumer complaint data we possess is from the year 2011 Q4 to 2017 Q1, but the assets data we possess is from 2011 Q2 to 2016 Q2. Due to the less temporal gap, we have very few data points.

We perform linear regression and obtain the following result:

```
call:
lm(formula = countData$count.complaints ~ countData$count.assets)

Coefficients:
(Intercept)  countData$count.assets
-1.108e+05    1.164e-05
```

```

Call:
lm(formula = countData$count.complaints ~ countData$count.assets)

Residuals:
    Min       1Q   Median       3Q      Max
-6333.9 -1768.8  -396.1   1531.3   8491.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.108e+05  1.703e+04  -6.507 2.91e-05 ***
countData$count.assets  1.163e-05  1.362e-06   8.544 1.91e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3753 on 12 degrees of freedom
Multiple R-squared:  0.8588,    Adjusted R-squared:  0.8471
F-statistic: 73 on 1 and 12 DF, p-value: 1.905e-06

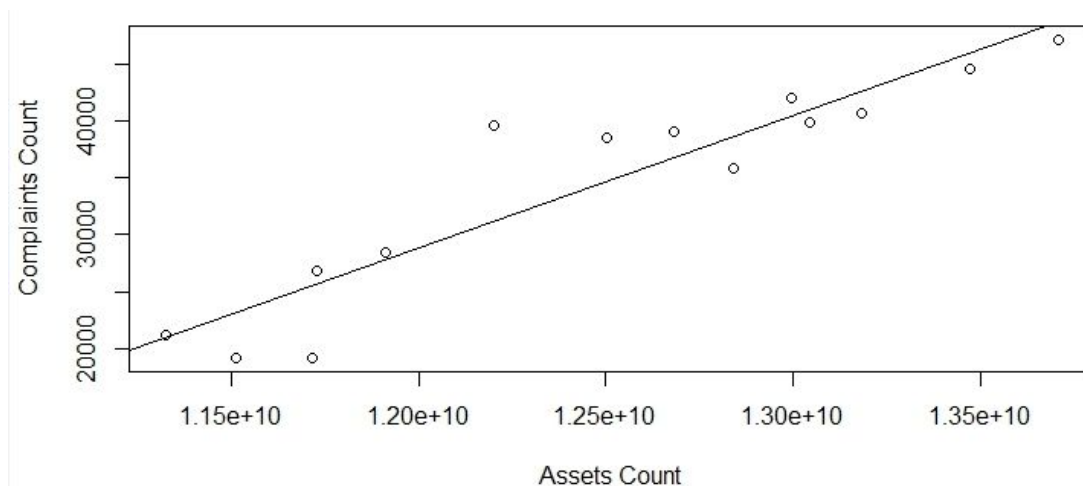
```

As we can see the p-value is $\ll 0.05$, so the null hypothesis holds true, but the relation between the DV and IV is marked by a very small coefficient factor i.e $1.163\text{E-}5$. The low number of records could be held responsible for such a result.

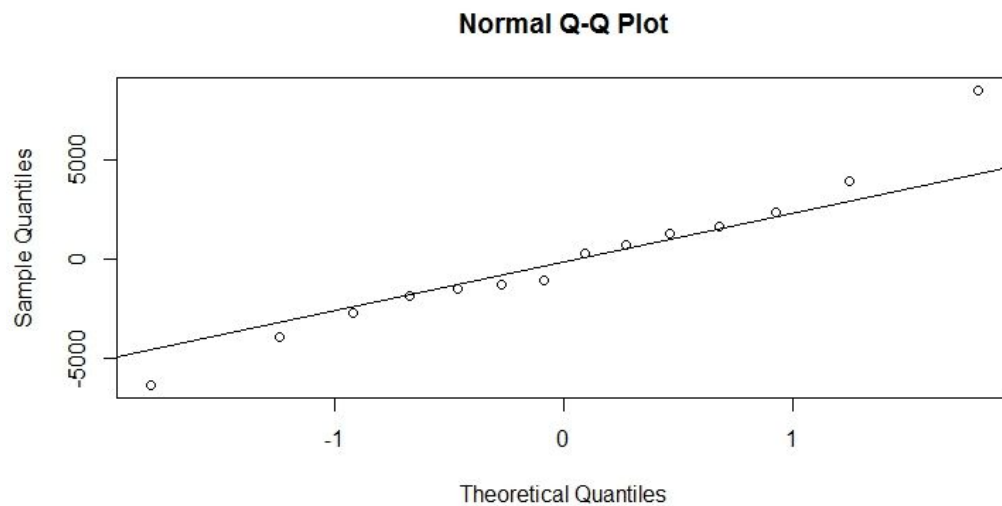
Intercept : **-1.108e+05**

Coefficient: **1.163e-05**

We plot the regression line through the X-Y plot



Linear regressions and predictions only work under the assumption that residuals (errors) follow a normal distribution. We check the same using QQ Plot

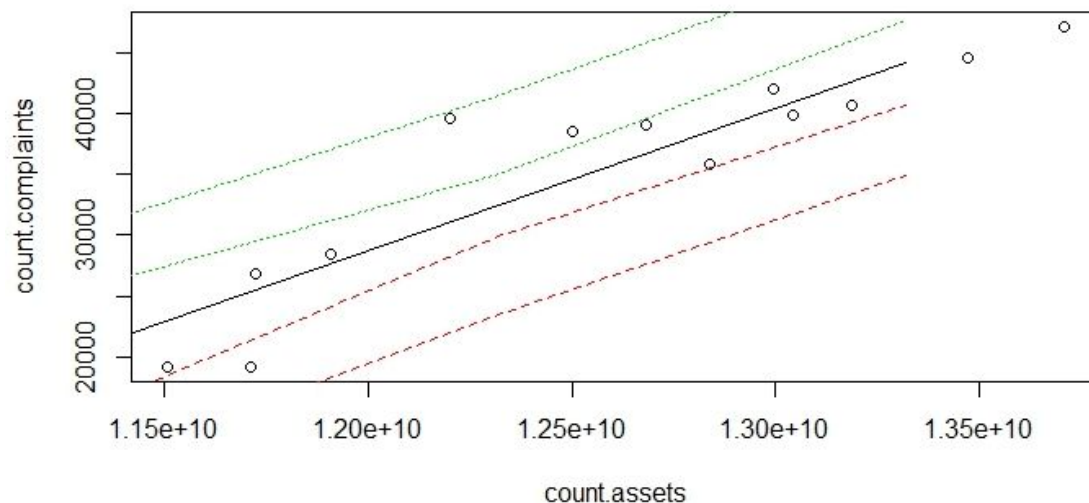


Most of the points lie closer to the line, so it appears to be normally distributed. Hence we can use Linear Regression,

This model cannot be assumed to be predictive, due to the extremely low value of the coefficient for the predictor variable. One unit increase in asset results in a very miniscule per unit increase (in order on E-05) in complaints.

All the other factors are categorical, hence they cannot be used to predict the number of complaints.

We divide the dataset into training and testing data with **10 points in training and 4 points assigned to testing dataset**. We obtain the following confidence and prediction bands after training our prediction model



Reporting Prediction accuracy

(1) The prediction accuracy using cor is **0.3428409**

(2) The prediction accuracy using mean square error is **12071733**

Part B : Multivariate Regression

We only have one predictor variable, the number of assets, so we cannot make use Multivariate regression for our prediction model and analysis.

Part C : Regularization

Whenever we use polynomial function or large set of features into fitting model, model will overfit on the data. If we are using linear function or fewer features set, then model will under fit on the data. In both cases, model will not be able to generalize for new data and prediction error will be large, so we require right fit on the data. So we use regularization to make sure that we our model doesn't suffer from overfit or underfit. Since we only have one independent variable, we cannot use regularization, we will only have a linear function.

Part D : Repeated Training

We are assigning 10 data points for training and 4 data points for testing. Since we have such limited data points, we will reiterate over the 14 points and split in repeated groups of 4 and 10 points such that we can cover all possible combinations. After running the code for 14 iterations we obtained the following output :

```
[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"
```

As we can see from the above result set so obtained, the correlation factor isn't remaining constant and it varies within a range. An interesting observation is that 11/14 records have a cor value between 0.7 to 0.9.

Logistic Regression and NB

Since the outcome variable is of type continuous and not binary, we cannot use logistic regression for our analysis. Since there are no classes present, we cannot use Naive Bayes classification.

Decision Trees and Random Forests

Since both the input and output are continuous variables, and it is difficult to establish clear boundaries, we cannot easily bin the data and hence we cannot perform decision trees and random forests. The outcome variable i.e. no of complaints comes within the range of 20,000 to 47,000 (per quarter) over the year 2011 to 2016, thereby establishing a clear trend that complaints are on an general increasing per year. It is difficult to provide allocative bin value thresholds for such a kind of data. For the predictor variable, the difference between the upper and lower bound is around 20% of the lower bound, in such a case defining such small boundaries would not provide to be useful. Since it is difficult to find ranges to bin the data and assign classes, we cannot do any form of classification using decision trees or random forests.

Comparative Analysis

Due to the continuous nature of our variables, we were able to perform only Linear regression, and that did not provide us with a good prediction model due to weak correlation values, a very low coefficient value and large mean square errors. Repeated training also did not provide any better prediction accuracy. This model is a very weak predictive model.

Research Question 6:

Predicting the emotion of the consumer based on their complaints

We have a field called as 'Consumer_Complaint_Narative' in our dataset, which contains people's narrations about the complaints. Since these are all complaints, we expect the narratives to have a negative sentiment. But, there are different emotions in these narratives: anger, anticipation, disgust, fear, joy, sadness, surprise, trust. We aim to predict the emotion of the consumer based on his/her narrative.

Predictors: Consumer_Complaint_Narrative

Outcome: emotion

As of now, we have used an API called the Aylien API to classify and thus label about 1000 rows of our dataset where each observation is the narrative of the consumer's complaint. We wrote python script to use this API.

In the next milestone we hope to make use of the emotion lexicon we have to create a classifier of our own and also, try out some of the supervised learning techniques on this dataset.

NOTE:

We have conducted detailed comparative analysis of the methods we used in each question and are at the end of each question above.