Consumer Complaint Analysis

Financial Products & Services

OPIM 5671 – Data Mining & Business Intelligence

Team Members

Lily Abarbanel
Vijay Garad
Tianshu Ding
Dongting Li

Introduction

The Consumer Financial Protection Board (CFPB) provides customers with information and resources in order to protect and manage their finances. The CFPB owns a database called the consumer complaint database which is essentially a collection of customer complaints in regards to the financial industry. It is important for specific financial institutions and companies to take these complaints into consideration because if they do not, their customers may want to use a different company. Furthermore, if many complaints against a single financial institution are made without the financial institution spending time to solve the problem at hand, this may permanently damage that specific financial institution. If companies take a complaint, talk about the issue the customer had, and come up with a successful way to solve the issue, then the customer may have a really positive experience with that specific financial institution and will recommend that company to people he/she knows.

Each time a customer makes a complaint against a company or financial institution, said company has to pay a price. Whether that is the time that they spend listening to the customer, talking to the customer, brainstorming ways to solve the customer's issue, or spending actual money to provide monetary relief, a price is being paid by the company. In order to help mitigate the price being paid by the company, our ultimate goal is to predict consumer complaints using predictive modeling/data mining/text mining, and also predict the number of consumer complaints each week.

By having this information companies will be able to better prepare for customer complaints, and find ways to mitigate the number of requests as a whole. Ultimately, when customers are satisfied, the company will do better as a whole and be more successful.

Literature

In the middle of July 2020, the CFPB released updated consumer complaint data that shows how much the coronavirus has affected financial complaints. One statistic that really speaks volumes as to how much the pandemic has affected financial complaints is that "Comparing the weekly average complaint volume before and after the coronavirus emergency declaration, prepaid card complaints saw the greatest percent increase at 105 percent, and student loan complaints saw the greatest percent decrease at 24 percent" (CFPB July 2020)

Based on the CFPB article published in mid-July 2020, it is evident that the pandemic had a significant impact on customer complaints, and an article based in California definitely supports the CFPB's findings. In "Financial Complaints Soared During Pandemic", Reports Say, written by Jacqueline Sergeant,

"The California Department of Business Oversight said it had experienced an increase of more than 40% in consumer contacts. The department said that from March 1 through the end of June, consumer complaints increased more than 37% to an average of 588 per month".

The findings in California certainly are backed by the data we are seeing from the Consumer Complaint Database (Figure 4).

Since coronavirus is a good example of an unpredicted shock that has a direct effect on consumer complaints, companies may want to keep this in mind for the future: maybe they should have a specific team of people who are monitoring consumer complaints when unpredicted shocks occur, so that they could better satisfy their customers, and more quickly solve the problems that the customers are having.

Lastly, there was a study done by Xin Xu, that looks at the performance of time series, multiple linear regression and BP Neural Network models in regards to predicting customer complaints. They found that the neural network model is the algorithm that is best at predicting consumer complaints: "In can be concluded form the above table that the prediction model constructed by the neural network algorithm is much higher than the other two algorithms" (Xu 2019). The table referenced is found below (Table 1).

Algorithm	The relative error is less than 10%	The relative error i s less than 20%	The relative error i s less than 30%	Relative error is l ess than 40%
Multiple linear regression	24.83%	45.15%	62.75%	74.27%
ARIMA time series	22.48%	43.56%	60.21%	71.12%
Neural network algorithm	31.38%	58.47%	76.52%	90.29%

Table 1: Xin Xu's Conclusion

Data

Data Description

The data used for the predictive modeling and time series applications, which will be discussed later, comes from the consumer complaints database. This dataset has 18 attributes that are defined in Table 2.

Attribute	Definition
Date received	Date when complaint reached financial institution
Product	Product name for which complaint is logged
Sub-product	Sub product category of product
Issue	High level issue text
Sub-issue	sub details on issue
Consumer complaint narrative	consumer complaint narrative
Company public response	company public response on the complaint
Company	company
State	State
Zip code	Zip Code
Tags	tagging of any special customer
Consumer consent provided?	whether consumer consent was provided or not
Submitted via	how the complaint was submitted (web, phone, etc
Date sent to company	Date when the complaint was sent
Company response to consume	company response to consumer
Timely response	Boolean attribue that shows whether the complain
Consumer disputed?	Boolean if consumer disputed or not on resolution
Complaint ID	Unique identifier of each complaint

Table 2: Attribute Definitions

The data set has a total of 930,764 complaints, so in order to narrow it down, the data set will be filtered to the Capital One Bank.

Preliminary Data Analysis

In order to gauge which attributes will have a positive or negative impact on consumer complaints, preliminary exploratory data analysis was done. The first attribute that was looked at is the State. In Figure 1, it is evident that the number or complaints per 1,000 population by state varies.

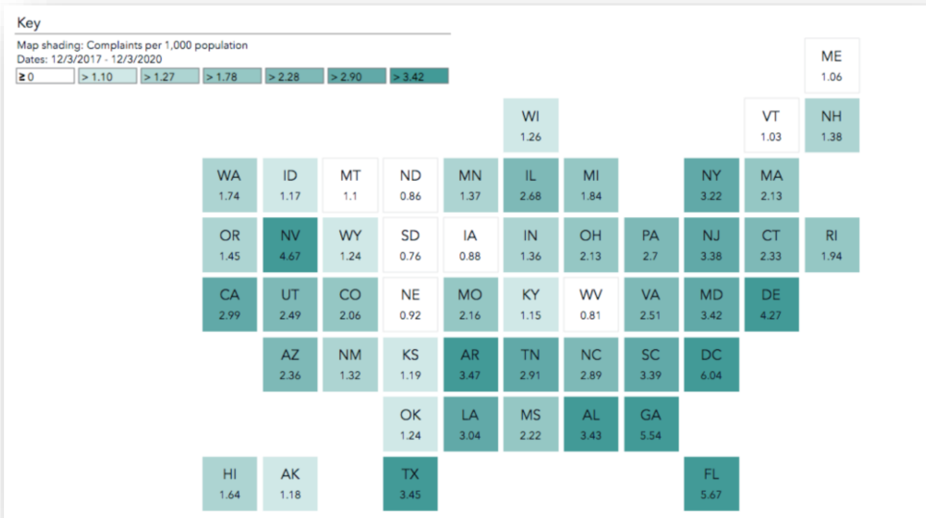


Figure 1: Number of Complaints per 1,000 Population by State

The darker blue states indicate that the state has a higher number of complaints. Based on the chart, some states that have a higher number of complaints are Florida, Texas, Alabama, Georgia, etc. This will be important to note going forward with our predictive modeling, because it seems as though larger states have a higher number of complaints.

Another attribute that is worth looking into is the impact sub-product and product have on number of complaints. Looking at table 2, product and sub-product are defined as the name for which the complaint is logged, and the sub product category of the product, respectively. Looking at figure 2, we may see that credit reporting, credit repair services, and debt collection are among some of the products that yield the highest number of complaints. Credit reporting alone has more than double the number of complaints as the second highest complaint product (debt collection). Keeping the state, and product/sub-product charts in mind will help predictive modeling ability later in the project.

Figure 2: Number of Complaints by Product/Sub-Product

Another attribute to consider is the company. Maybe there is a better chance a complaint is going to happen based on the company. Figure 3 shows the number of complaints by company.

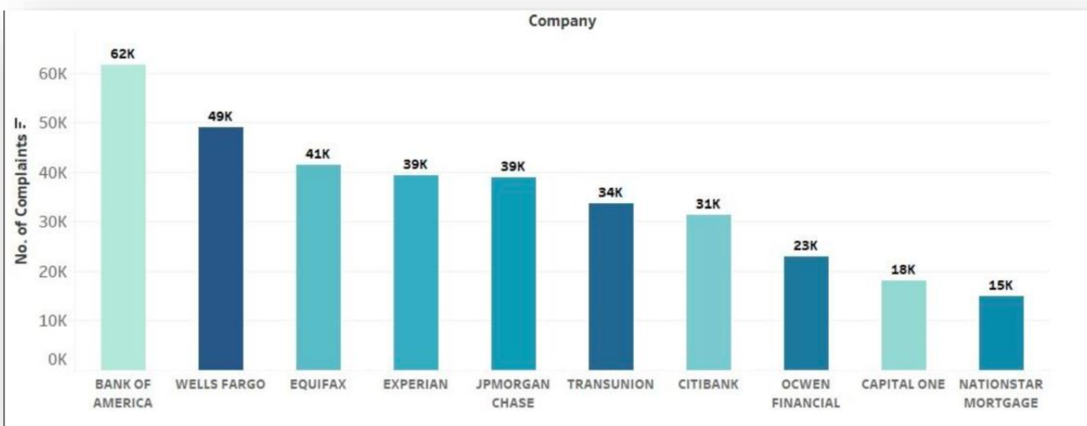


Figure 3: Number of Complaints by Company

As we can see by the chart, Bank of America has the highest number of complaints; however, this may be due to the fact that Bank of America is a huge bank and therefore has a lot of customers. We will specifically be looking at Capital One in order to account for the dataset being large.

Also, in our data set, it is evident that the coronavirus pandemic has significantly impacted the number of complaints for financial institutions. Looking at Figure 4, we are able to see that the number of consumer complaints have been consistently increasing since February of 2020 (when the pandemic first started and businesses started shutting down).

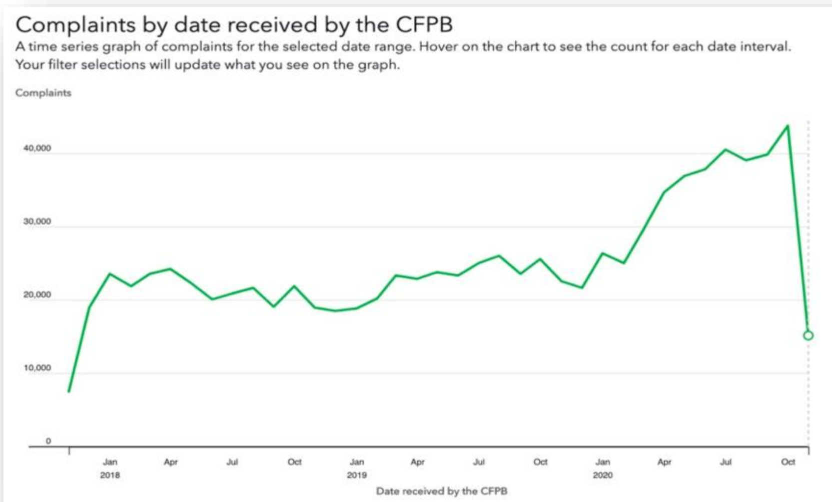


Figure 4: Number of Consumer Complaints over the Past 3 Years

Data Preprocessing

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. We have used SEMMA process, i.e., Sampling, Exploring, Modifying, Modeling, and Assessing, to preprocess our Capital One Bank's Customer Complaints data.
We have used tableau's data prep functionality to clean & validate the Bank's data.

Data Cleansing (Text Mining)

As a part of Data cleansing, we have dropped the irrelevant columns like Date received; Date sent to Company, Zip Code etc. as these columns do not have that much predictive power. Columns with too many null values were also dropped like Tags, consumer complaints narrative and company public response.

Data Cleansing (Time Series Forecasting)

For predicting number of complaints per day in the future, we have kept only relevant columns in the dataset like Date, Complaint ID. To create time series, we have aggregated Complaint ID data per day to crat Number of Complaints per day column like figure 5 & exported it to .csv file for forecasting the time series.

Capital One

Comaplaints per Day

December 1, 2011	16
December 2, 2011	16
December 3, 2011	6
December 4, 2011	4
December 5, 2011	34
December 6, 2011	27
December 7, 2011	15
December 8, 2011	11
December 9, 2011	4
December 10, 2011	3
December 11, 2011	1
December 12, 2011	8
December 13, 2011	7
December 14, 2011	5
December 16, 2011	7
December 17, 2011	4
December 18, 2011	2
December 19, 2011	8
December 20, 2011	11
December 21, 2011	6
December 22, 2011	5
December 23, 2011	3
December 25, 2011	3
December 26, 2011	2
December 27, 2011	5
December 28, 2011	13
December 29, 2011	9
December 30, 2011	9
December 31, 2011	2

Figure 5: Consumer Complaints per Day

Data Mining (Text)

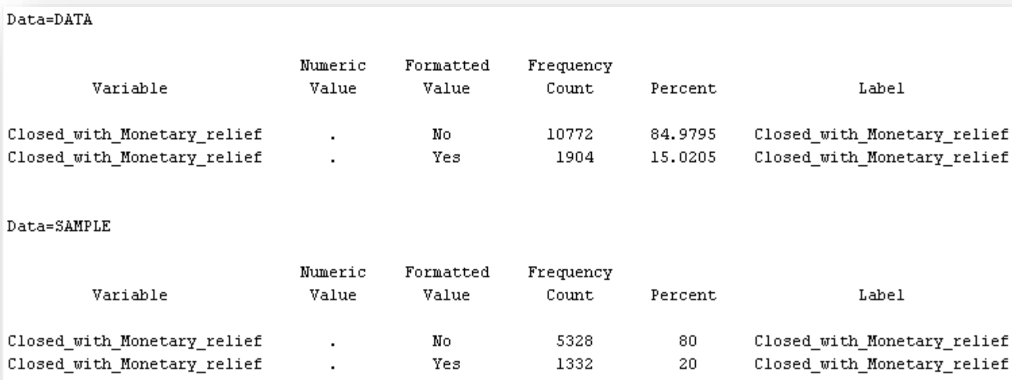
As a part of Initial Data setup, we have used File Import Node to import our newly created data file into the SAS Enterprise Miner. Then we have used Data partition node to divide our data into training & validation dataset into 70:30 ratio.

Our target variable, Closed with Monetary Relief, has low number of samples in our dataset which could negatively impact the predicting capacity of model building. Hence, to address this data imbalance issue, we have used data sampling node. We have used level based (70%) Stratified Sampling method with sample proportion as 20% to increase the number of target samples of our data. Because of this, we could increase target sample rate from 15% to 20%. Below figure shows the details of stratification strategy.

Columns: Label Mining

Name	Sample Role	Role	Level
Closed_with_M	Stratification	Target	Binary
Complaint_ID	Default	ID	Nominal
Consumer_dis	Default	Input	Nominal
Issue	Default	Text	Nominal
Product	Default	Input	Nominal
State	Default	Input	Nominal
Submitted_via	Default	Input	Nominal
Timely_respon	Default	Input	Binary
dataobs	Default	ID	Interval

Figure 6: Stratification on Target Variable



Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Closed_with_Monetary_relief	.	No	10772	84.9795	Closed_with_Monetary_relief
Closed_with_Monetary_relief	.	Yes	1904	15.0205	Closed_with_Monetary_relief

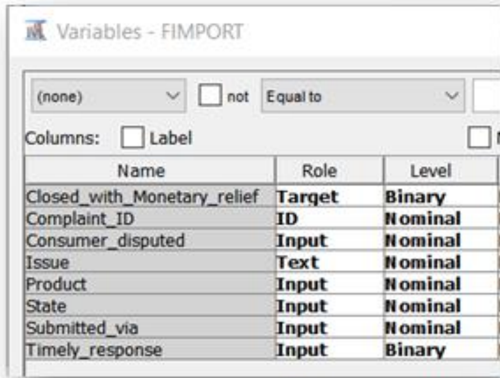
Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Closed_with_Monetary_relief	.	No	5328	80	Closed_with_Monetary_relief
Closed_with_Monetary_relief	.	Yes	1332	20	Closed_with_Monetary_relief

Figure 7: Target Variable Sampling after 20% level-based sampling

Our objective is to determine that whether the Capital One Bank Customer's Dispute closed with monetary relief or not?

1. Target Variable: Closed with Monetary Relief.
2. Input Variables: Complaint_ID, Consumer_Disputed, Issue, Product, State, Submitted_via, Timely_response.



Variables - FIMPORT

(none) | not | Equal to

Columns: Label

Name	Role	Level
Closed_with_Monetary_relief	Target	Binary
Complaint_ID	ID	Nominal
Consumer_disputed	Input	Nominal
Issue	Text	Nominal
Product	Input	Nominal
State	Input	Nominal
Submitted_via	Input	Nominal
Timely_response	Input	Binary

Figure 8: Input, Text & Target Variable Dataset

Text Preprocessing

Test Parsing

The text parsing node generated the terms by document matrix. We could identify the most frequently occurring terms along with the number of documents the terms occurred in & weight of the particular term. With the help of various attributes like Start List, Stop List, Synonyms Lists, Multiterm Lists, we could be able to identify frequency, weight of the term.

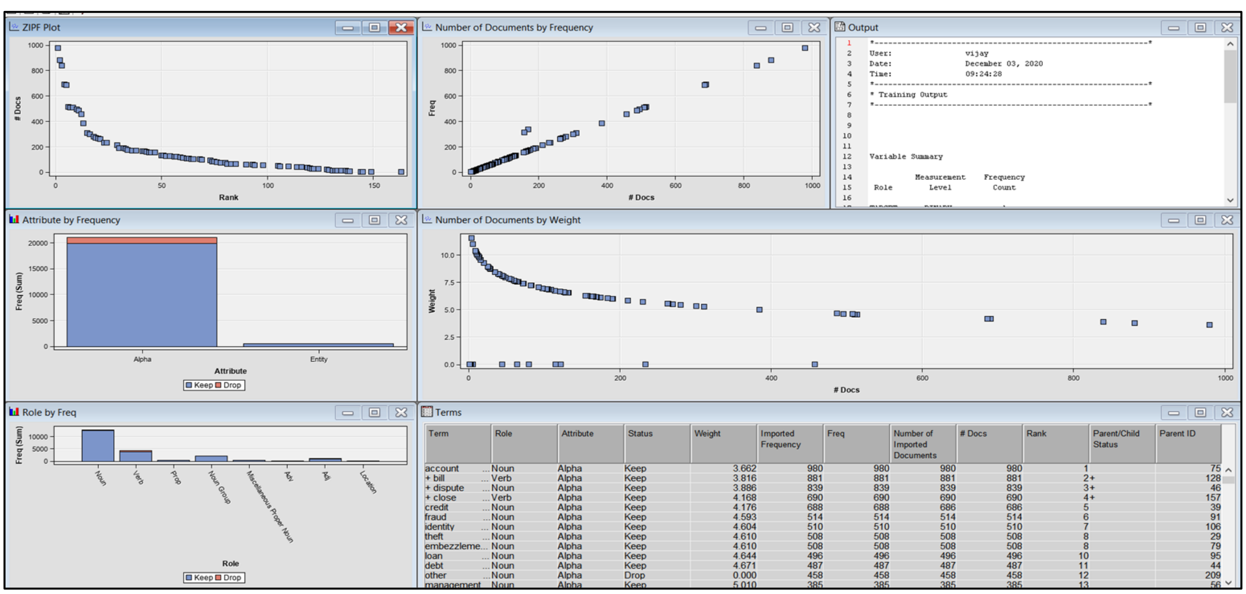


Figure 9: Text Parsing Node Results

Text Filtering

To reduce the number of parsed terms or documents that are analyzed, we have used Text Filter Node. For that we have used frequency **Weighting method** as **Log** and **Term Weighting method** as **Inverse Document Frequency**. We also kept Minimum number of documents the term needs to appear in to be **4**. With the term window of interactive filter viewer results from text filter node, we could get below results. We could get account, bill dispute as a most frequently used terms which makes sense considering our data is related to financial complaints.

Terms

TERM	FREQ	# DOCS	KEEP ▼	WEIGHT	ROLE
account	980	980	✓	3.661	Noun
⊞ bill	881	881	✓	3.815	Verb
⊞ dispute	839	839	✓	3.885	Noun
⊞ close	690	690	✓	4.167	Verb
credit	688	686	✓	4.176	Noun
fraud	514	514	✓	4.592	Noun
identity	510	510	✓	4.603	Noun
theft	508	508	✓	4.609	Noun
embezzlement	508	508	✓	4.609	Noun
loan	496	496	✓	4.644	Noun
debt	487	487	✓	4.67	Noun
management	385	385	✓	5.009	Noun
opening	385	385	✓	5.009	Noun
⊞ fee	311	311	✓	5.317	Noun

Figure 10: Tet Filtering Node Results.

We also analyzed strength of association of few terms to have better understanding of relationship between different text terms with the help of concept linkage viewer tab as shown in below figure.

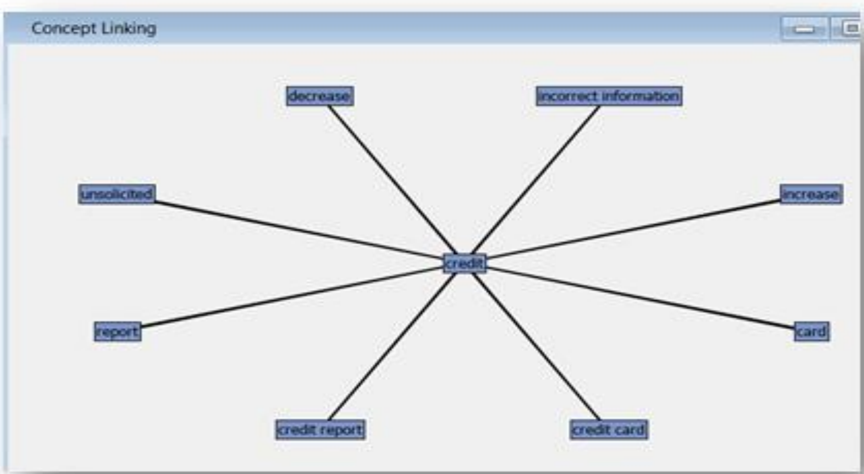


Figure 11: Concept linkage tab viewer results

Text Clustering

To reduce the dimensionality of dataset & to cluster the document into different set of clusters, we have used text cluster node with clustering algorithm to be Expectation-Maximization. SVD Resolution-High, Max SVD Dimensions-20, Number of Clusters-40.

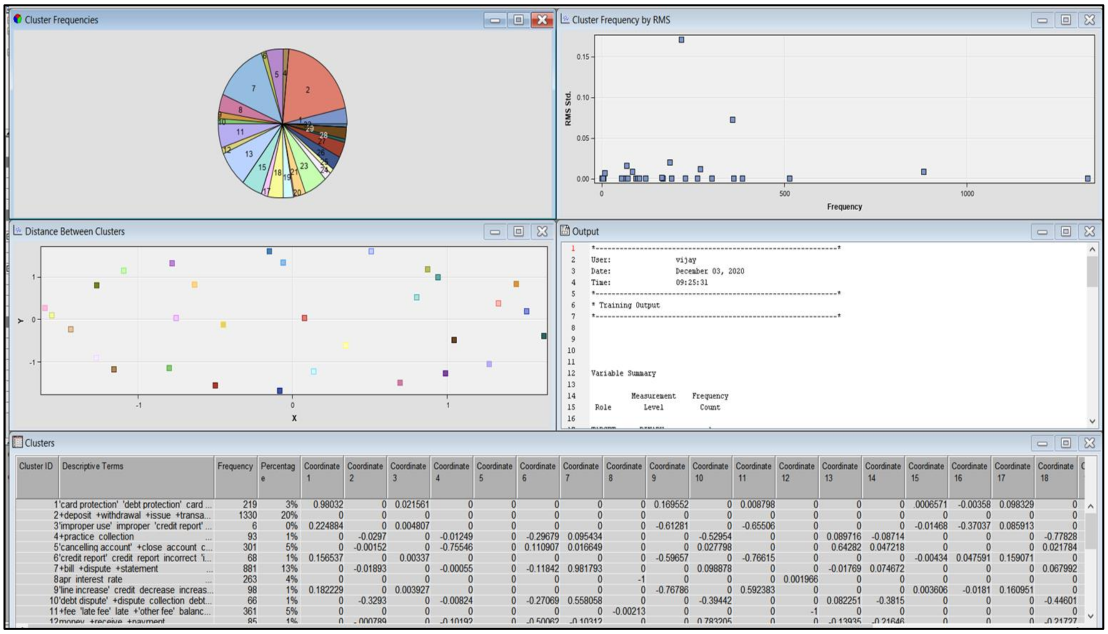


Figure 12: Text Cluster Node results

Text Topic Node

Text Topic node helped us to combine the terms into relevant topics for further analysis.

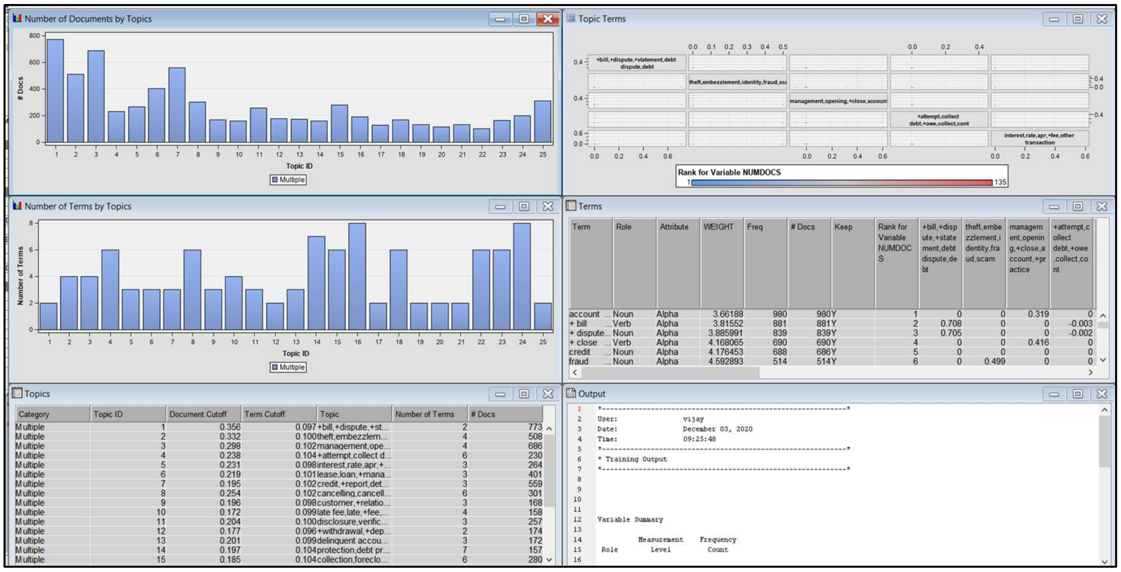


Figure 13: Text Topic Node Results

Model Building & Assessment

To predict consumer complaints which causes direct cost, we have built 5 classifier models on our parsed data. We compared those models with the help of model assessment node on the basis of misclassification rate & ROC on the validation dataset. From the model, we were able to find out that Neural Network Model Performs best in term of both ROC & Misclassification rate.

Five Classifier Models -

1. Decision Tree
2. Regression
3. Gradient Boosting
4. Neural Network
5. Text Rule Builder

Model Assessment (ROC & Misclassification Rate)

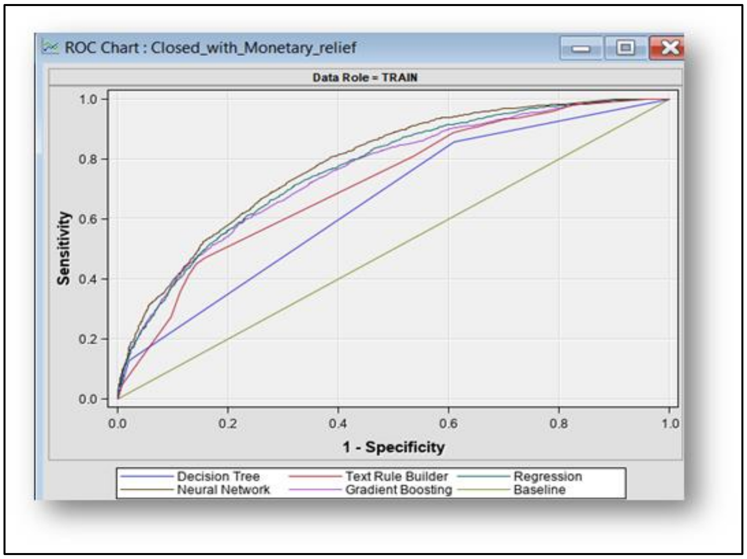


Figure 15: ROC curves for 5 models

Model Name	ROC Index	Misclassification Rate
Neural Network	**0.783**	**0.183 (Best Model)**
Regression	0.763	0.188
Gradient Boosting	0.754	0.189
Decision Tree	0.652	0.192
Text Rule Builder	0.715	0.233

Table 3: ROC & Misclassification Rate

Model Diagram

Figure : SAS Enterprise Miner Diagram

Model Assessment on Score Dataset

We imported our score dataset & generated the predictions based on our best model. (Neural Network).We found out that our misclassification rate is 0.11, which lowest one.

Misclassification Rate on Score Dataset -

(FN + FP) / Total Records

= (20 + 27) / 461 = 45/461 = 0.102

	BB	BC	BD	BG	CG	CH	CI
1	Product	State	Submitted_via	Complaint_ID	EM_PROBABILITY	EM_CLASSIFICATI	Closed_with_Monetary_relie
2	Debt collection	CA	Web	2271248	0.944874969	NO	No
3	Credit reporting	CA	Web	2271249	0.931954526	NO	No
4	Credit card	CA	Web	2271431	0.816697975	NO	No
5	Credit card	NY	Web	2271776	0.564104374	NO	No
6	Credit card	UT	Phone	2272457	0.945814854	NO	No
7	Credit card	RI	Web	2272750	0.657313733	NO	No
8	Credit reporting	CT	Web	2272956	0.922499615	NO	No
9	Bank account or service	MI	Web	2272966	0.884956469	NO	No
10	Mortgage	OH	Web	2273065	0.985625599	NO	No
11	Bank account or service	MO	Web	2273122	0.938875258	NO	No
12	Debt collection	NJ	Web	2274013	0.977827245	NO	No
13	Consumer Loan	NM	Web	2274017	0.992890407	NO	No
14	Credit card	NC	Web	2274380	0.922131631	NO	No
15	Debt collection	CA	Web	2274564	0.966477953	NO	No
16	Debt collection	CA	Web	2274948	0.910912583	NO	No
17	Credit card	MS	Web	2274988	0.65421836	NO	No
18	Bank account or service	NY	Web	2274998	0.817515204	NO	No
19	Bank account or service	CA	Phone	2275342	0.732152203	NO	No
20	Credit reporting	NV	Web	2275377	0.966501063	NO	No

Figure 16: Model Prediction Results on Score Dataset

Conclusion & Recommendations (Text Mining)

Based on the decisions of 5 classifier models, we recommend Capital One Bank

- To resolve the customer's complaint in timely manner as it is key determinant of your customer satisfaction.
- Text mining helped us understand that many issues with Capital One Bank's customer are related to Credit Reporting & Debt collection area. We recommend them to create robust system in those areas.

Objective - Analyze Bank's customer complaints data to save the money.

Method - Text Mining

Models Built - Regression, Neural Network, Gradient Boosting, Decision Tree, Text Rule Builder

Model assessment - Neural (82%), Regression (81%), Gradient Boosting (80%), Decision Tree (78%).

Time Series Exploration

We used our dataset complaint for the Time Series Exploration task. Added Complaint Count as the dependent variable and added Day of Date received as the TimeID in the additional roles. We can see that the Interval automatically gets set to Week. In order to make equally spaced data, we used accumulation as sum for our models. Accumulation is the process of transforming transactional data into time series.

The purpose of exploring time series before modelling is to determine whether the series contains any noticeable trend or seasonality. From our trend component plot for the complaint count, we do not observe any trend. It looks like there has been a random increase and decrease in the complaints over the years.

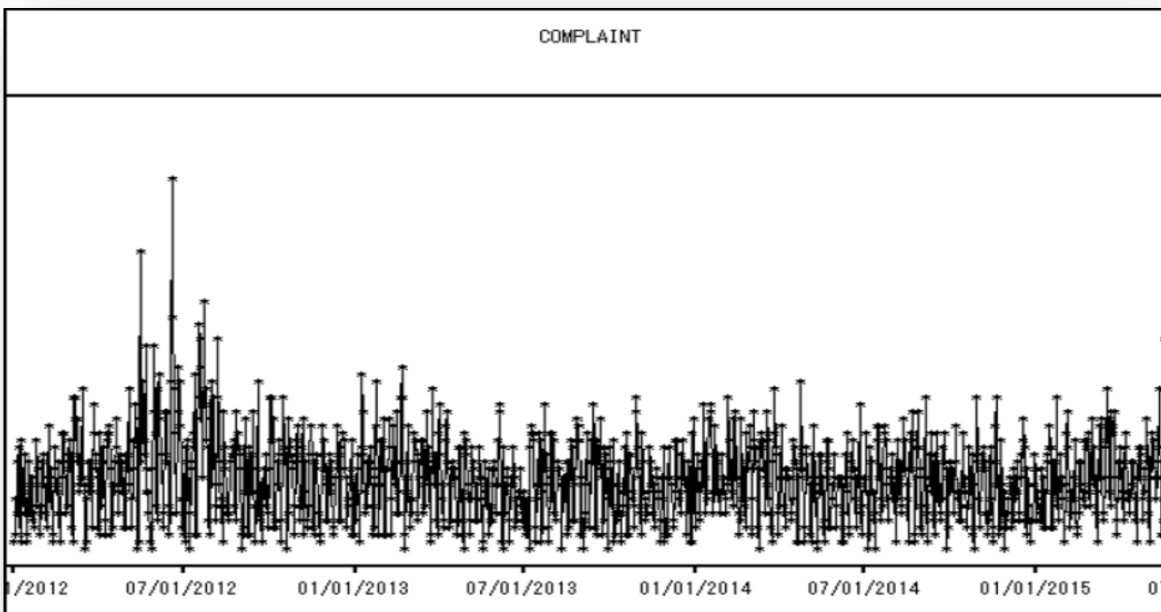
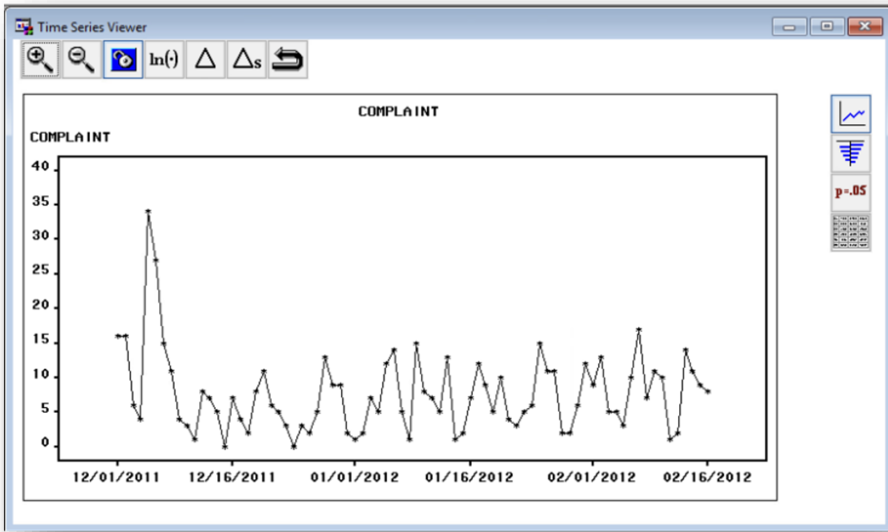


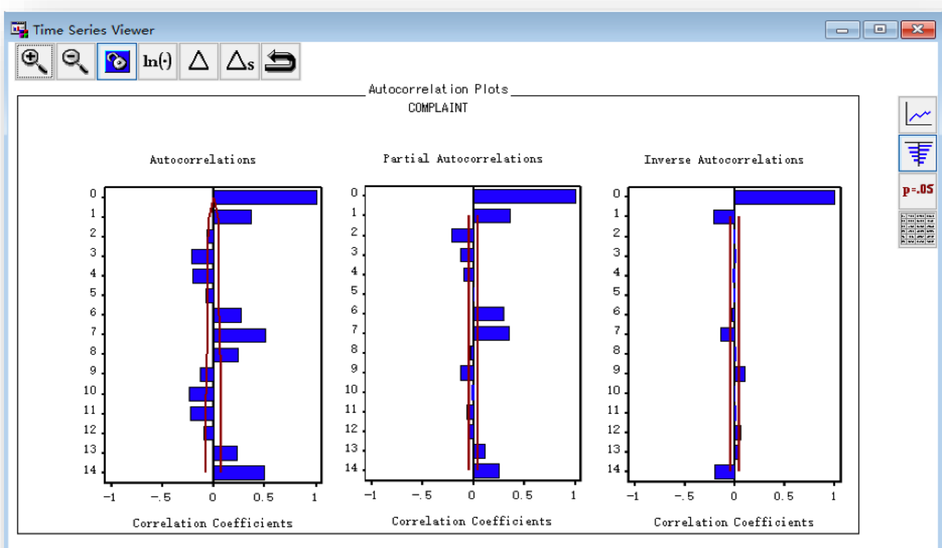
Figure 17: Complaints per Day Time Series

Then from the seasonality plot for the complaints count, we can observe that there is seasonality in our dataset. Ups & Downs in the data every 7 days. Hence, we could say that time series have seasonal factor in it.



<u>Figure 18: Seasonality Time Series (Zoom in)</u>

We also performed a correlation analysis for the complaints. We observed from the White Noise Prob plot that there is no white noise implying that they have only signal. Hence, we reject our null hypothesis here. From the autocorrelation plots of ACF, we can see that the correlation dies out slowly with every successive lag.



<u>Figure 19: Auto-Correlation Plot</u>

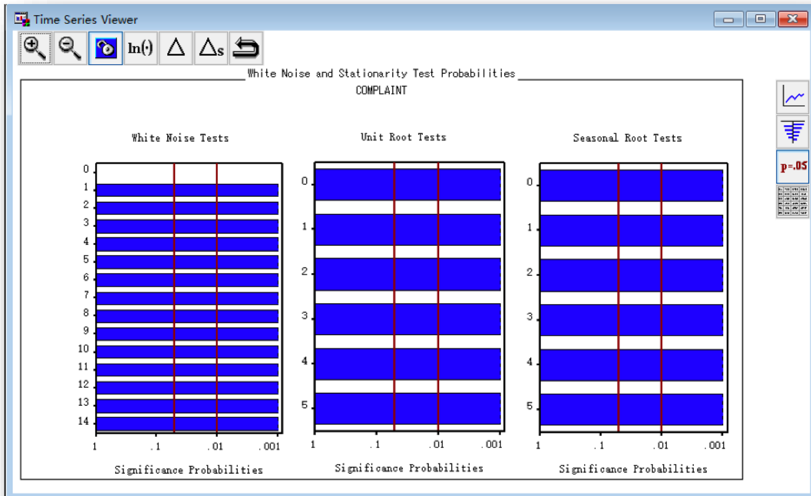


Figure 20: White Noise Plot

Model Fitting

To fit the model, we use hold out 315 days of data. Then we built 4 models: Simple Exponential Smoothing, Double Exponential Smoothing, Additive Seasonal Exponential Smoothing models and ARIMA model with linear trend and seasonal dummies.

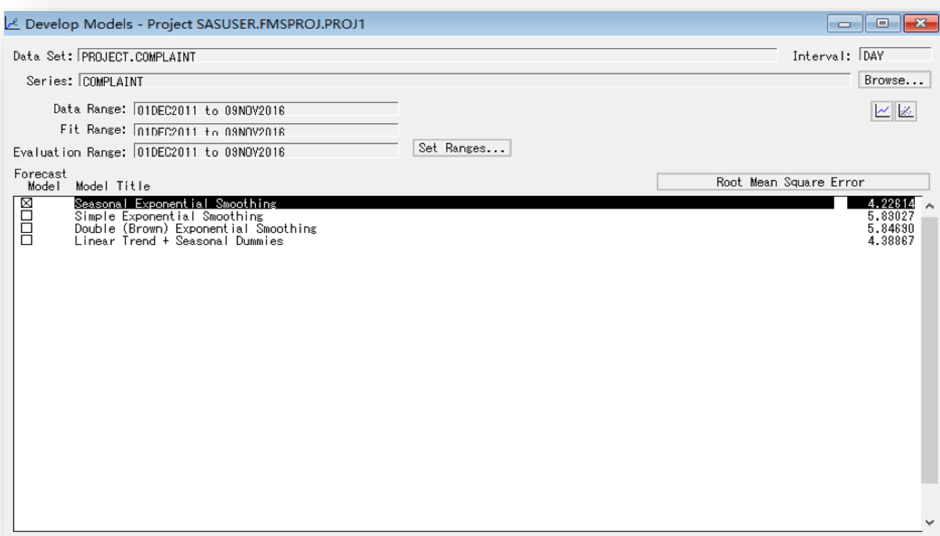


Figure 21: TSFS Model Building

Results

Single Exponential Smoothing Model

We analyzed the results of our four models. For the Single exponential smoothing model, we can see from the White Noise Prob plot for the prediction errors for complaints that there is white noise in our model. Hence, we fail to reject the null hypothesis. Since it is white noise, it consists of only noise with independent and identically distributed random variables with a mean of 0 and the same finite variance. We observed some white noise in our fit statistics model.

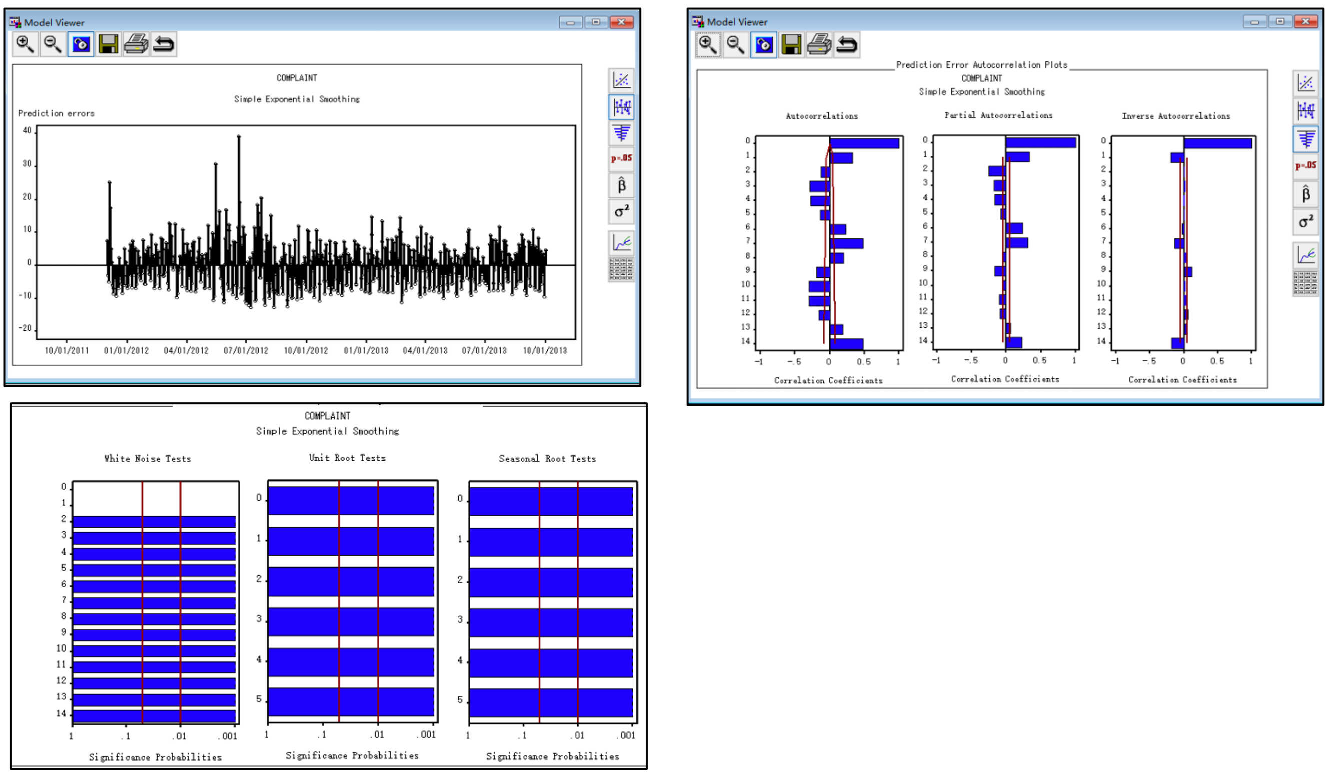
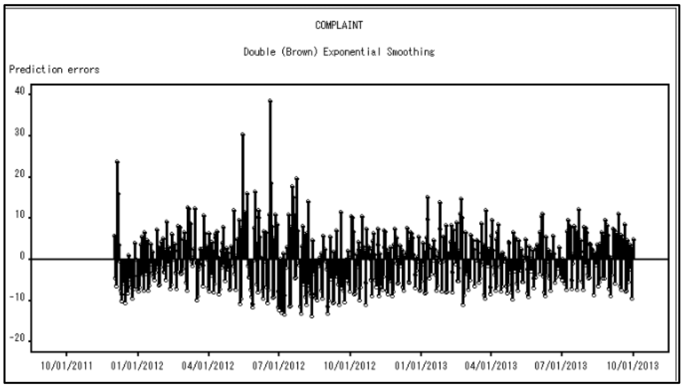
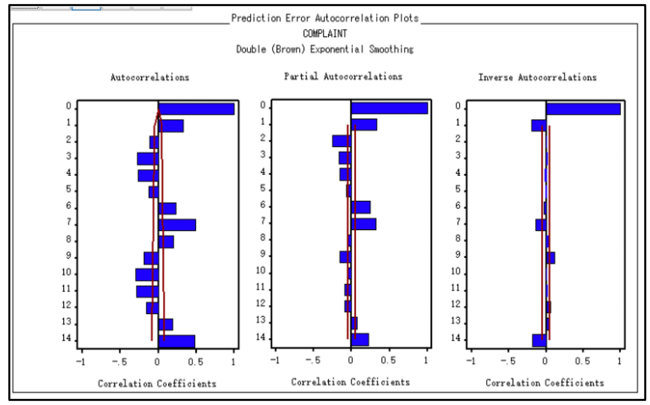


Figure 22: Single Exponential Smoothing Model Plots

Double Exponential Smoothing Model

For the Double Exponential Smoothing model, similar to single exponential model, we can observe white noise and few spikes at 7-day iteration at the lags.





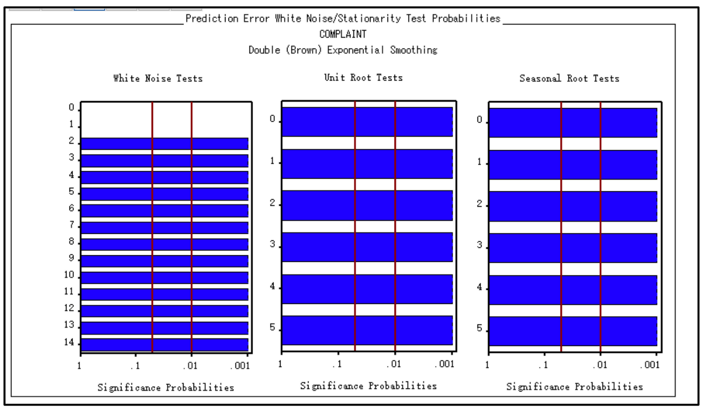


Figure 23: Double Exponential Smoothing Model Plots

Additive Seasonal Exponential Smoothing Model

For our third model which is the Additive Seasonal Exponential smoothing model, we can observe that there is not much white noise in our model and there are no spikes in the autocorrelation plots at the lags. Hence, we concluded that seasonal exponential smoothing is performing better over previous 2 models.

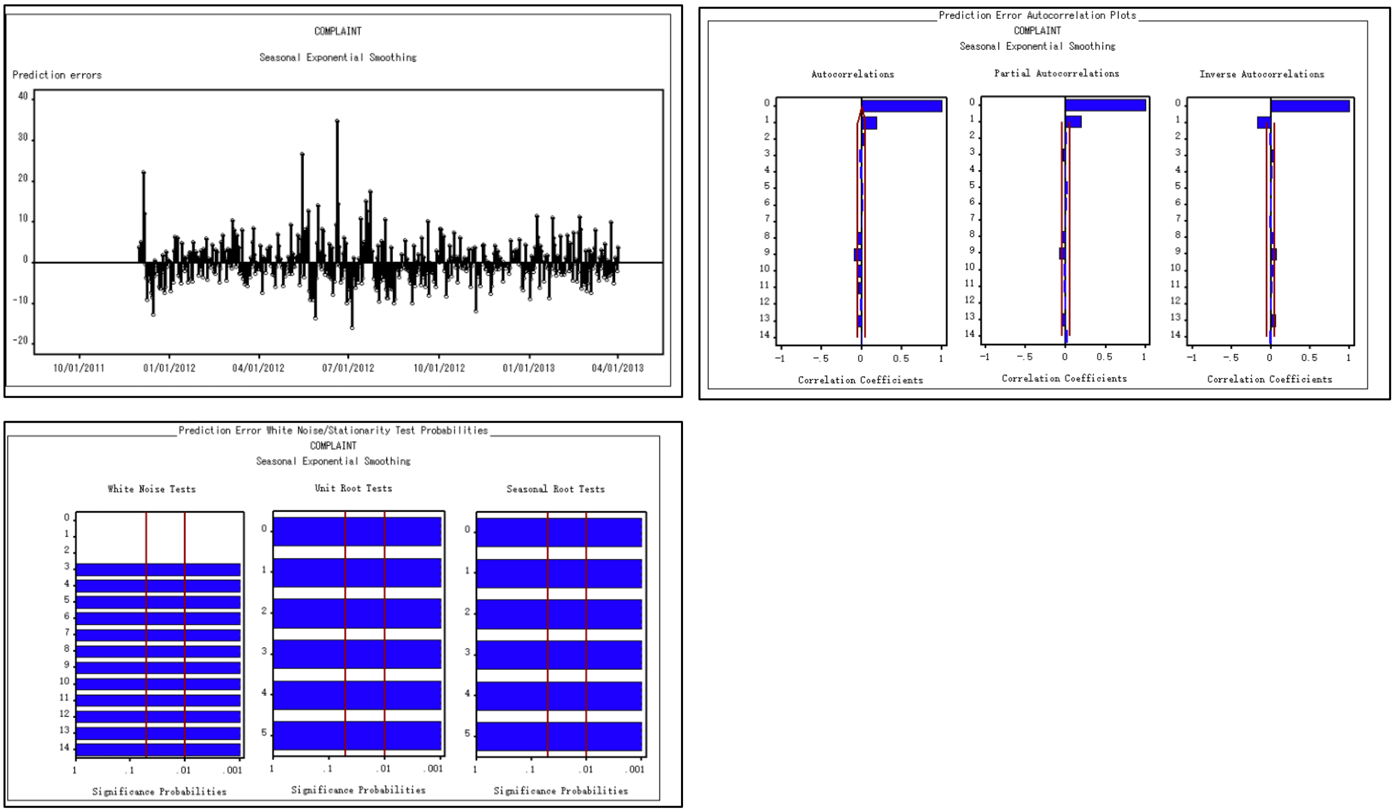


Figure 24: Additive Seasonal Exponential Smoothing Model Plots

Linear Trend + Seasonal Dummies Model

For our fourth model which is the Linear Trend + Seasonal Dummies model, we can observe that there is not much white noise in our model, but there are no spikes in the autocorrelation plots at the lags. Hence, we concluded that Linear + seasonal dummies is performing better over first 2 models.

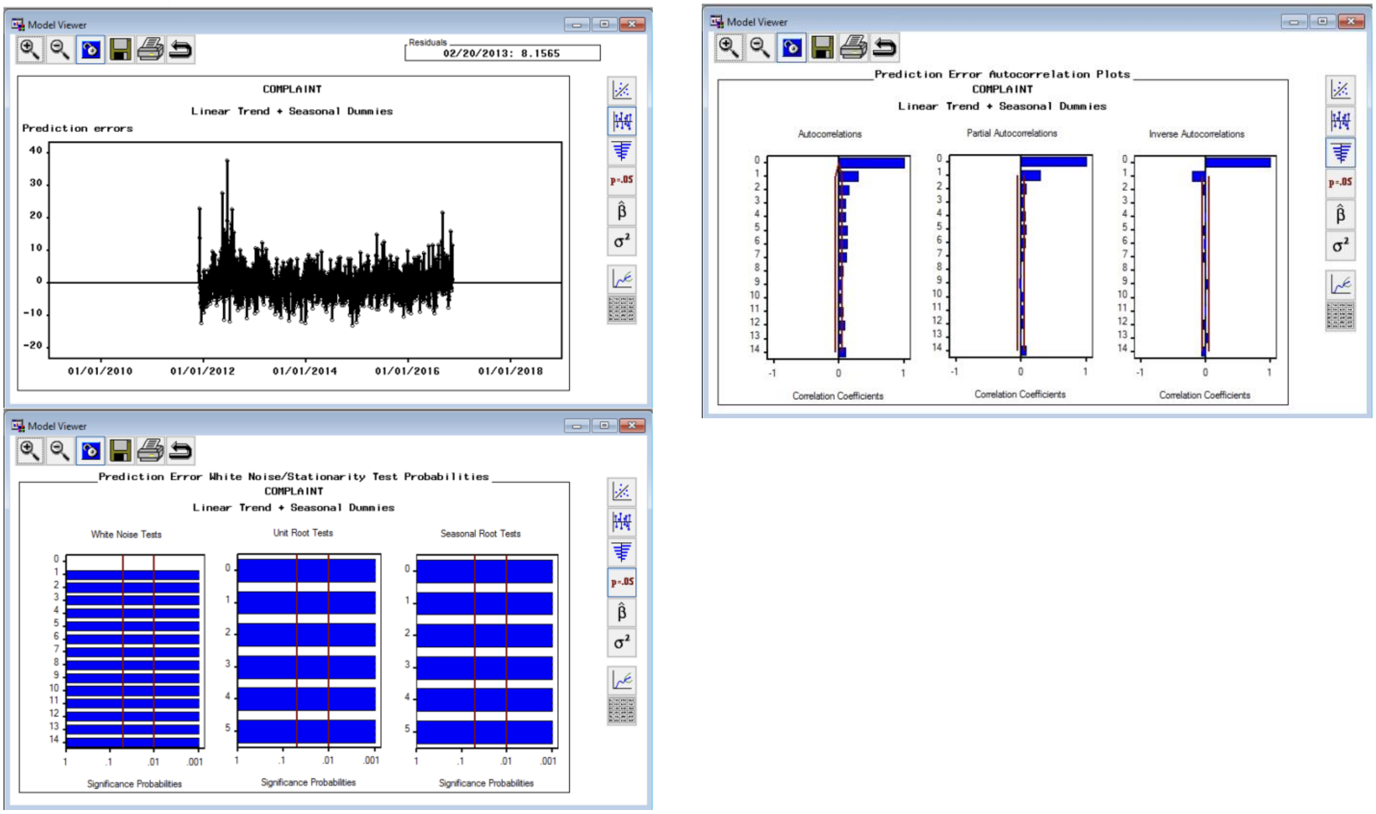


Figure 25: Linear Trend + Seasonal Dummies Model Plots

Forecasting

Then we use the full data set to do the forecast.

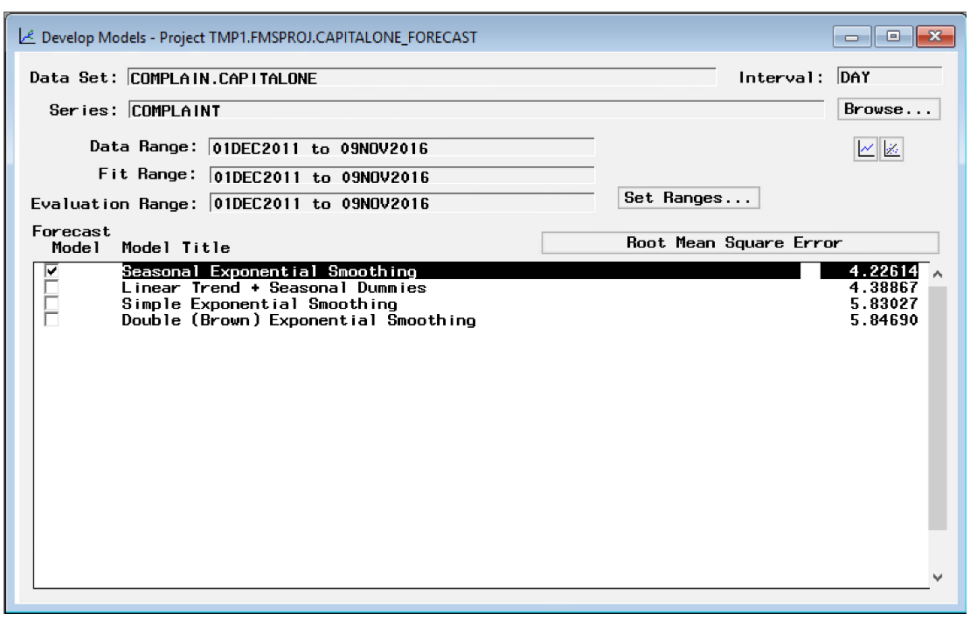


Figure 26: Model Forecasting

Model Comparison

Model - Fit Statistics	MAPE	RMSE	MAE	MSE
Simple Exponential Smoothing Model	**96.82687**	**5.89027**	**4.68159**	**33.99207**
Double Exponential Smoothing Model	**98.21463**	**5.84690**	**4.70079**	**34.18620**
Additive Seasonal Smoothing Model	**47.71251**	**4.22614**	**3.16234**	**17.86028**
Linear Trend + Seasonal Dummies Model	**47.86584**	**4.38867**	**3.20302**	**19.26040**

Table 4: Fit Statistics Model Comparison

Model - Forecast Statistics	MAPE	RMSE	MAE	MSE
Simple Exponential Smoothing Model	**94.85576**	**5.60292**	**4.56406**	**31.39272**
Double Exponential Smoothing Model	**93.41458**	**5.61047**	**4.56673**	**31.47773**
Additive Seasonal Smoothing Model	**46.76294**	**3.97340**	**3.03336**	**15.78793**
Linear Trend + Seasonal Dummies Model	**46.47611**	**4.07551**	**3.06624**	**16.60978**

Table 5: Forecast Statistics Model Comparison

Conclusion & Recommendation (Forecasting)

- Based on the forecasting results, Capital One can implement a rotational program in which staff would be rotating around different departments where they foresee a surge in complaints in near future.
- Predicting whether a complaint will cost the bank to pay a monetary relief will help in prioritizing those sets of complaints and get them resolved in a speedy way so that if possible, we can avoid monetary settlement with the customers.

References

- https://www.consumerfinance.gov/about-us/newsroom/cfpb-releases-updated-covid-19-consumer-complaint-data/ (CFPB July 2020)
- https://www.fa-mag.com/news/financial-complaints-soared-during-pandemic--reports-say-57161.html (Sergeant 2020)
- https://iopscience.iop.org/article/10.1088/1742-6596/1187/5/052036/pdf (Xin Xu et Al 2019)

Appendix

Data Sample Node

.. Property	Value
General	
Node ID	Smpl
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Sample Method	Default
Random Seed	12345
Size	
Type	Percentage
Observations	.
Percentage	10.0
Alpha	0.01
PValue	0.01
Cluster Method	Random
Stratified	
Criterion	Level Based
Ignore Small Strata	No
Minimum Strata Size	5
Level Based Options	
Level Selection	Event
Level Proportion	70.0
Sample Proportion	20.0
Oversampling	

Properties & Results

Data=DATA

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Closed_with_Monetary_relief	.	No	10772	84.9795	Closed_with_Monetary_relief
Closed_with_Monetary_relief	.	Yes	1904	15.0205	Closed_with_Monetary_relief

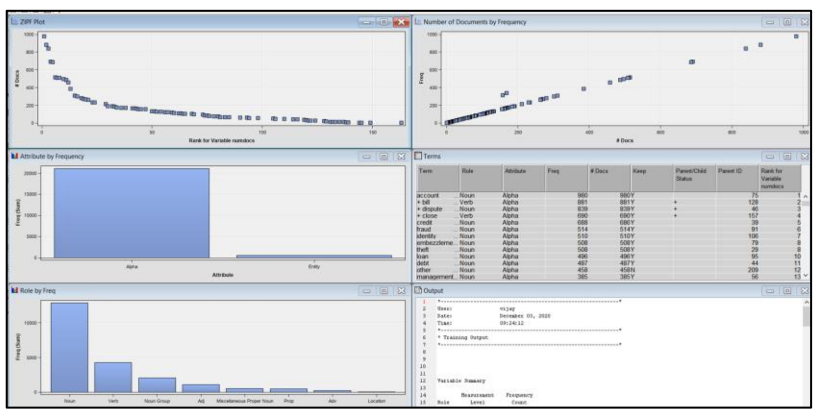
Data=SAMPLE

Variable	Numeric Value	Formatted Value	Frequency Count	Percent	Label
Closed_with_Monetary_relief	.	No	5328	80	Closed_with_Monetary_relief
Closed_with_Monetary_relief	.	Yes	1332	20	Closed_with_Monetary_relief

Text Parsing Node

.. Property	Value
General	
Node ID	TextParsing
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Parse	
Parse Variable	Issue
Language	English
Detect	
Different Parts of Speech	Yes
Noun Groups	Yes
Multi-word Terms	SASHELP.ENG_MULTI
Find Entities	Standard
Custom Entities	
Ignore	
Ignore Parts of Speech	'Aux' 'Conj' 'Det' 'Interj'
Ignore Types of Entities	...
Ignore Types of Attributes	'Num' 'Punct'
Synonyms	
Stem Terms	Yes
Synonyms	SASHELP.ENGSYNMS
Filter	
Start List	...
Stop List	SASHELP.ENGSTOP

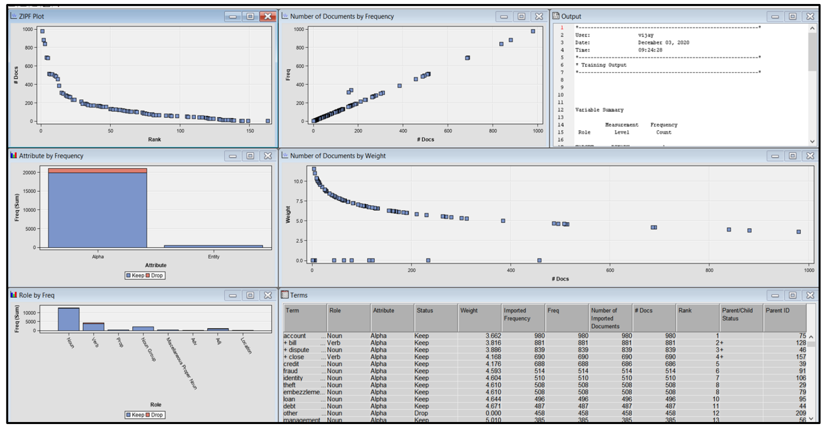
Properties & Results



Text Filter Node

.. Property	Value
General	
Node ID	TextFilter
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Spelling	
Check Spelling	No
Dictionary	...
⊟Weightings	
Frequency Weighting	Log
Term Weight	Inverse Document Freque
⊟Term Filters	
Minimum Number of Docu	4
Maximum Number of Term	.
Import Synonyms	...
⊟Document Filters	
Search Expression	
Subset Documents	...
⊟Results	
Filter Viewer	...
Spell-Checking Results	...
Exported Synonyms	...

Properties & Results



Text Cluster Node

.. Property	Value
General	
Node ID	TextCluster
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
⊟Transform	
SVD Resolution	High
Max SVD Dimensions	20
⊟Cluster	
Exact or Maximum Number	Maximum
Number of Clusters	40
Cluster Algorithm	Expectation-Maximization
Descriptive Terms	15
Status	
Create Time	12/3/20 6:54 AM
Run ID	3161f972-4eee-4b02-aa93-
Last Error	
Last Status	Complete
Last Run Time	12/3/20 9:25 AM
Run Duration	0 Hr. 0 Min. 21.57 Sec.
Grid Host	
User-Added Node	No

Properties & Results