Paper SAS282-2017

Development of Customer Intelligence Platform for the BFSI sector Using SAS Viya

Riya Adsul, Vedant Misra, Rishabh Tulshyan

ABSTRACT

The evolution of data science and advanced forms of analytics has given rise to a wide range of applications that are providing better insights and business value in the enterprise. Data science practices give organizations the capabilities they need to gain valuable information from ever-increasing amounts of highly variable data. The proposed project has been executed collaboratively with the Customer Success and Advisory team at SAS, utilizing the SAS Viya platform. SAS Viya is an advanced AI, analytic, and data management platform with modern and scalable architecture, leveraging the cloud analytics service (CAS). The project utilizes four of the platform's main features, namely SAS Visual Analytics, Visual Statistics, Built Models and Visual Text Analytics. The primary objective of the project is to develop a framework for extracting business-focused insights from various forms of data, that are relevant to business operations. To achieve this, it was imperative to have an in-depth understanding of how information and value propagate within a business and utilize this knowledge to identify opportunities for growth. The framework aims to identify opportunities for businesses by utilizing multiple use cases, such as identifying key data assets that can be transformed into data pipelines, enabling the delivery of sustainable tools and solutions, including Management Information Dashboards, Lead Scoring, Revenue Optimization, Risk Scoring, Customer Churn, Fraud Analytics, Customer Satisfaction etc.

In this paper, we apply a repeatable model of text analytics and machine learning techniques to the publicly available Prudential Life Insurance dataset, automobile insurance dataset and CFPB data. Specifically, we use SAS® VIYA: Visual analytics, Visual statistics, Build Models, and Visual Text Analysis to explore trend, patterns, sentiment, and machine learning techniques to generate predictive models and model the natural language available in each free-form complaint against a disposition code for the complaint. We also explore methods to structure and visualize the results, showcasing how areas of concern are made available to analysts using SAS® Visual Analytics and SAS® Visual Statistics. Finally, we discuss the applications of this methodology for overseeing financial institutions.

To generate insights, build model, test & validate the team has followed scientific methodology of developing the hypothesis and trying to disprove the same using various testing scenarios

To develop other components such as dashboards, reports, DDC objects, web app, mobile app & various integration teams has followed an agile approach by dividing the project into work streams and further achieving the milestone in a smaller sprint cycle.

INTRODUCTION

The technological landscape changes, and so the industries do. Nowadays a worldwide variety of insurance exists. Insurers can have a difficulty to understand customer behaviour, frauds, policy risk, and claim surety, which is mandatory before giving policy to someone. It took years for insurers to sell directly to their customers and issue policies online while competing on price comparison websites. Many companies still have not achieved it. With the prefiltration of data, the use of advanced math and financial theory to analyse and understand the customer behaviour

and costs of risks have been the stalwarts of the insurance industry. The analytics performed by actuaries are critically important to an insurer's continued profitability and stability. Traditionally companies are just looking for what happened in the past with Descriptive analytics. But now, the industry is demanding more such that what will happen in the future (predictive analytics) and how actions can change the outcome (Prescriptive analytics).Big data makes the insurance industry a perfect sphere for data analytics to construct basic patterns, get fundamental insights about the insurance business, and manage the complex relations between agents and clients. It might be a possibility that the client is fraud or life impaired that will create a huge problem for the insurer. Consistently evolving business environments are increasing competition and risk. Several other challenges, like theft and fraud, are also plaguing the insurance business.

The above challenges force insurers to generate insights from data to enhance pricing mechanisms, understand customers, safeguard fraud, and analyse risks. Data analytics collate more precise information about several transactions, product performance, customer satisfaction, etc.

- Data Analytics creates new capabilities that empower insurers to optimize every function in the insurance value chain with the help of data-driven decision-making.
- It can also analyse a customer's risk and determine which client is trustworthy or may give great loss.
- It can also detect fraud, like through which the greatest frauds happened.
- Customers can use data analytics to know which insurance company gives the minimum price with suitable offers.

Companies that underwrite insurance policies have to evaluate applications carefully. The pay-outs from insurance claims are very high relative to the insurance premiums that companies collect from an individual customer. For example, a person who purchases a $ 500 a year, $1 million sum-assured 40-year term plan would be paying $ 20,000 over 40 years but in case if a claim arose, the insurance company would have to pay $ 1 million to the dependents of this individual. So, the company has to be selective about people whom it chooses to insure to keep its business financially viable. Researchers have applied machine learning techniques to perform predictive analytics and automate the insurance application evaluation process. The underwriting process has been shown to be streamlined using predictive analytics, which also enhances decision-making [12]. There hasn't been a lot of study done in this area, however. Analytics help in the underwriting process to provide the right premiums for the right risk to avoid adverse selection. Predictive analytics approach in insurance Complex & Intelligent Systems mainly deals with modelling mortality rates of applicants to improve underwriting decisions and profitability of the business [10]. Risk profiles of individual applicants are thoroughly analysed by underwriters, especially in the life insurance business. Risk classification is a common term used among insurance companies, which refers grouping customers according to their estimated level of risks, determined from their historical data [11].

The first purpose of this research is to categorise the amount of risk in the life insurance sector using predictive modelling in SAS Viya. We then suggest the best model to evaluate risk and provide ideas to improve underwriting procedures. Here, an application is seen as a data point, with the applicant s data columns serving as its characteristics and the risk rating of the applicant serving as the output that we are attempting to forecast. By increased automation, we want to cut down on expenses for the business and greatly speed up the processing of life insurance applications while maintaining the accuracy of risk assessment.

Another major use case in an insurance organization is handling consumer complaints. Some challenges faced by an organization due to traditional Complaint Redressal Systems:

Delayed service delivery: Engaging with agencies is often hindered by long wait times as the staff struggles to coordinate a high volume of inquires across many points of contact.

Understand public sentiment: Understand public sentiment by analysing feedback to determine how dimensions of interest change and adjust your decisions based on the need of consumers

Identify areas of improvement: Uncover trends, customer demographics and spot opportunities for action as public comments are submitted, while also having the flexibility to explore how trends change over time.

Given the volume of the complaints, how can an overseeing organization assess the data for various trends, including the areas of greatest concern for consumers?

As more complaints are filed, is the solution to handling the increasing workload adding more readers to manually address the complaints and identify trends? There are 3 problems with this approach:

1. Unless very specific standards are adopted, the method that one reader uses to address and tag a complaint can be quite different from the method a second reader uses.
2. Reader fatigue ensures that the way a reader will address the first 10 complaints of the day will not necessarily be the same as the way they address the last 10 complaints. Vital information might be missed or skipped.
3. Suppose a trend is uncovered, and the directive arises to go back and retag all the data from the past year with this new trend. This is a case where manual analysis doesn't scale.

We aim to use Natural language processing using SAS VTA to respond quickly and accurately to citizen inquiries by offering real-time recommendations based on rapid classification of customer complaints and feedback, and to understand public sentiment by analysing feedback to assist in decision-making based on customer perspectives.

Consumer complaint narratives may be sent to an analysis engine to get categorised, useful information extracted and could be routed to other places (a workflow to escalate an issue), for a response. Similar citizen queries and complaints may have been raised in the past and addressed. A specific customer's text input could be used as the basis for search against an existing index, and recommendations/answers provided for past complaints of a similar nature could be used to handle this complaint. For this use case we have used the CFPB dataset. Consumer Finance protection bureau is one of a number of overseeing institutions that ensure that the consumer is treated fairly by corporations and financial institutions. The CFPB has handled more than one million complaints since its inception, and this number is increasing annually[1].

In the sections that follow in this paper, we will explore a short end-to-end implementation that showcases how an analyst can use SAS technology to quantitatively assess the complaint data for various trends. This includes the consumers' areas of greatest concern, as well as areas of complaint that need legislative correction. We show how to apply a sentiment model to the text as well as machine learning methods through SAS Visual Text Analytics to accomplish this. Finally, we will assess the results using visualization capabilities to highlight actionable information.

The purpose of the project is to develop complete integrated ecosystem for an organization (primary focus on Insurance and Banking sector) to leverage data analytics at various processes of different departments like:

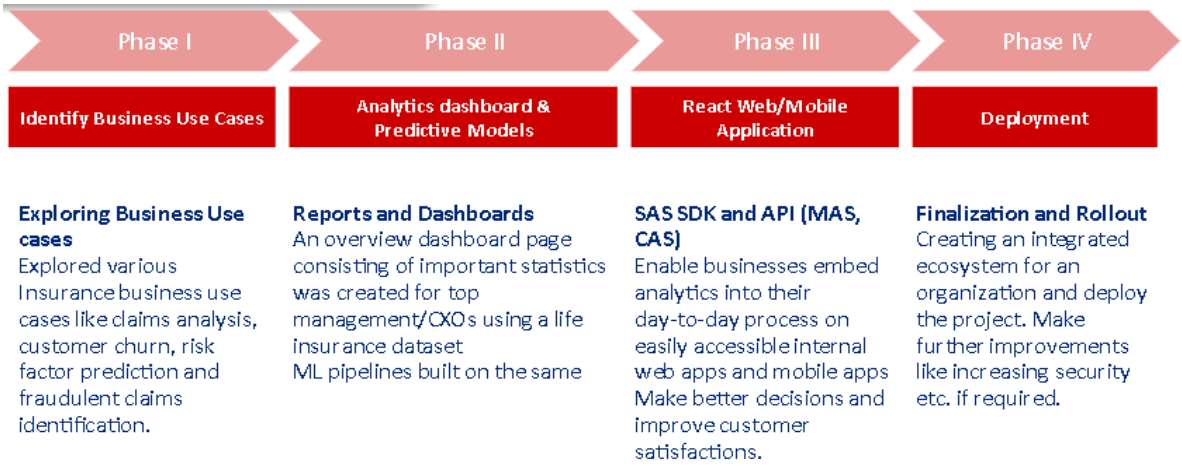
1. Finance – Provide a view of revenue and expenses distributed by various lines of business. Provide forecasts based on historical data
2. Sales – Provide insights on Sales achievement & factors affecting sales.
3. Marketing – Lead Scoring, Lead Generation, Customer Segmentation & Customer Interaction. Analyzing consumer complaints for enhancing customer satisfaction.
4. Management – Provide view of different business of the organization, highlight instances of missed opportunities and fraudulent activities, and offer insights into potential avenues for growth and development.

The project will be used as a reference for developing end-to-end analytics / data science use case journeys for organization thus improving opportunities of cost saving for end customers and well as organizations, improve processing / wait times, enhance customer experiences.

METHODOLOGY

The four-step process for generating and using the framework is as follows:

1. Identify Business Use Cases: Explored various Insurance business use cases like claims analysis, customer churn, risk factor prediction, fraudulent claims identification and customer complaint analysis.
2. Analytics dashboard & Predictive Models: An overview dashboard page consisting of important statistics was created for top management/CXOs using a life insurance dataset ML pipeline built on the same
3. React Web/Mobile Application: Enable businesses embed analytics into their day-to-day process on easily accessible internal web apps and mobile apps Make better decisions and improve customer satisfactions.
4. Deployment: Creating an integrated ecosystem for an organization and deploy the project. Make further improvements like increasing security etc. if required.



DATA ACQUISITION AND PREPARATION FOR THE USE CASES

- **For Consumer Complaint Classification and Extraction of information using Visual Text Analytics**, the CFPB Complaints Dataset which is a publicly available dataset provided by the Consumer Financial Protection Bureau (CFPB) that contains consumer complaints related to financial products and services is used. The dataset includes over 2 million complaints received by the CFPB from 2011 to 2021. The complaints cover a wide range of financial products and services, including credit cards, mortgages, student loans, and bank accounts.
  Each complaint in the dataset includes information such as the type of product or service, the issue the consumer faced, the company the complaint was filed against, the zip code of the consumer, and a narrative description of the problem. The dataset also includes information on how the complaint was handled by the company, including whether the company provided a timely response, the company's response to the complaint, and whether the consumer disputed the company's response. It contains 18 features and 555957 observations of the 18 features, only the 'product' feature and the 'consumer_complaint_narrative' will be explored and further applied to modelling.

- **For Insurance Risk Factor Prediction and Premium Optimization:**
  The Prudential Life Insurance Dataset which is a collection consisting of 59,381 applications with 128 variables that define the characteristics of applicants for life insurance. The data set consists of anonymised nominal, continuous, and discrete variables. The following categories may be used to broadly group the features:

  1. Details on your health, including your height, weight, age, BMI, and any diseases you may have. While the definition of this phrase is not stated, it may comprise elements that show a person s marital status, the number of children they have had, and other factors.
  2. Insurance History: Again, it isn't mentioned clearly what this collection of information entails, but it may indicate whether the individual has ever missed paying a payment, whether they were previously covered, the degree of insurance coverage, etc.
  3. Personal Information: This category may contain details about a person s work, profession, pay grade, and seniority as well as details about the security of their PIN code for their home, if they own a vehicle, and other details that may be important when choosing a life insurance policy.
  4. Product information: This is information particular to the product. This may thus refer to a variety of factors, including the money insured, the duration of the plan, and whether payments are made in one lump sum or over time. It may not be immediately clear how a product s attributes might affect an applicant s risk assessment. It's possible that the firm may classify applicants who choose a 10-year plan which is considered short in the context of life insurance as high risk.

  There are 126 variables in all, divided into 60 categories, 48 dummy, 13 continuous, and 5 discrete features. The response or output variable is the risk rating, which has 8 values from 1 to 8, with 1 being the lowest risk rating and 8 the greatest.

- **For Fraud Prediction and Customer Churn**
  An auto insurance dataset consisting of 17000 rows was used for forecasting renewal rates based on information from clients whose insurance had already expired. The second dataset, which contained information on each policy officer's pending claims, was utilised to forecast the fraud rate.

  1. Incident Information:
  Authorities contacted: Details about whether any authorities were contacted following the incident, such as the police, fire department, or emergency medical services.
  Incident city: Details about the city where the incident occurred.
  Incident date: The date of the incident.
  Location: Specific location of the incident, such as an address or intersection.
  State: Details about the state where the incident occurred.
  Type: Type of incident that occurred, such as a car accident, theft, or vandalism.

  2. Vehicle Information:
  Auto make: The make of the vehicle involved in the incident, such as Ford, Toyota, or BMW.
  Auto model: Details about the model of the vehicle involved in the incident, such as Camry, Explorer, or X5.
  Collision type: The type of collision that occurred, such as head-on collision, rear-end collision, or side impact collision.

  3. Policyholder Information:
  Education level: Information about the education level of the policyholder or the driver involved in the incident.
  Occupation: The occupation of the policyholder or the driver involved in the incident.
  Relationship: Details about the relationship between the policyholder and the driver involved in the incident, such as spouse, child, or friend.
  Sex: The sex of the policyholder or the driver involved in the incident.

  4. Claim Information:
  Fraud reported: Whether fraud was suspected or reported in relation to the incident.
  Severity: The severity of the incident, such as minor, moderate, or severe.
  By grouping the columns, it is possible to identify patterns and relationships among the variables, which can be useful for data analysis and modeling. For example, combining vehicle information with policyholder information could be useful for predicting the likelihood of an incident occurring based on certain driver profiles. Similarly, incident information combined with claim information can be useful for predicting the severity of a claim or the likelihood of fraud based on the specific circumstances of the incident.

  Overall, this database is a valuable resource for insurance companies and analysts to understand the factors that influence auto insurance claims and develop models to predict the likelihood of an incident occurring or the severity of the incident. It can also help identify patterns and trends related to fraud or other suspicious activities, which can help prevent losses and improve overall risk management strategies.

INSURANCE RISK FACTOR PREDICTION AND PREMIUM OPTIMIZATION

- **Data preparation**

There are 128 characteristics and 59,381 occurrences in the data set. The SAS Viya Prepare Data function was used in the data pre-processing stage to find the missing data.

1. The Missing Data Mechanism: The qualities with more than 30% of the data missing would not be included in the analysis [54]. The only characteristics that are kept for further analysis are the attributes Employment Info 1, Employment Info 4, Employment Info 6, and Medical History 1. To impute the missing values for these four properties, a treatment must be applied.
2. Imputation of missing data: Multiple imputation is a suitable strategy to replace the missing values in the features if the data are taken to be MAR. Multiple imputation is a statistical method that projects missing values using the information that is currently available. According to [56], multiple imputation entails three steps: imputation, analysis, and pooling. As multiple imputation considers the uncertainty in missing data, it is more trustworthy than single imputation methods like mean or median imputation [57, 58].
3. Multiple imputation involves the following steps:
    - Imputation: Depending on the number of imputations specified, this phase includes imputed the missing data many times. Several full data sets are produced as a consequence of this phase. A predictive model, like linear regression, often does the imputation to replace missing values with anticipated ones based on the other variables in the data set.
    - Analysis: The different fully formed data sets are examined. Standard errors and parameter estimations are assessed.
    - Pooling: To get the result, the analysis s findings are then combined.

- **Risk Factor Prediction Model**

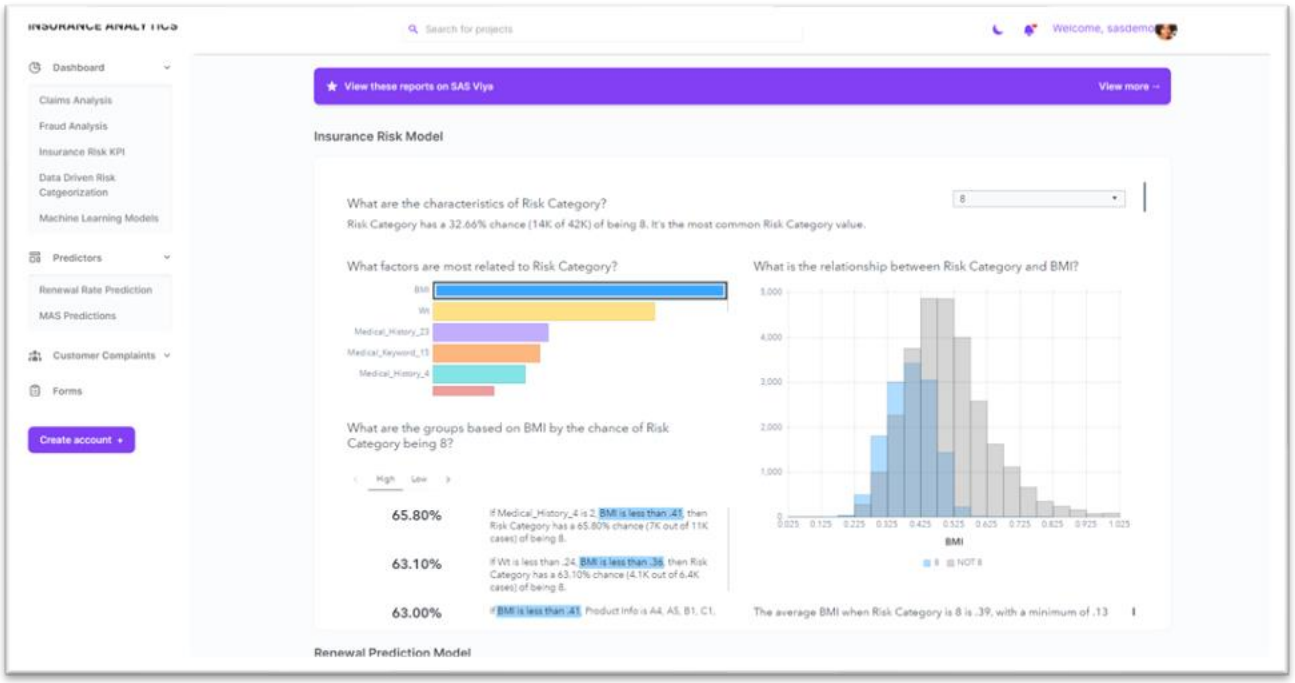
6 pre-existing models were applied to the dataset and model comparison was performed after applying ensemble learning on all the models. The models chosen were:

1. Neural Network: Neural networks are a series of algorithms that identify underlying relationships in a set of data. These algorithms are heavily based on the way a human brain operates. These networks can adapt to changing input and generate the best result without the requirement to redesign the output criteria. In a way, these neural networks are like the systems of biological neurons.
2. Gradient Boosting: Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. When a decision tree is the weak learner, the resulting algorithm is called gradient-boosted trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalises44 the other methods by allowing optimization of an arbitrary differentiable loss function.
3. Logistic regression: Logistic Regression is used to solve the classification problems in machine learning. They are like linear regression but used to predict the categorical variables. It can predict the output in either Yes or No, 0 or 1, True or False, etc. However, rather than giving the exact values, it provides the probabilistic values between 0 & 1.

4. Random Forest: Random Forest is the ensemble learning method, which consists of many decision trees. Each decision tree in a random forest predicts an outcome, and the prediction with most votes is considered as the outcome. A random forest model can be used for both regression and classification problems. For the classification task, the outcome of the random forest is taken from most votes. Whereas in the regression task, the outcome is taken from the mean or average of the predictions generated by each tree.
5. Decision Tree: Decision trees are the popular machine learning models that can be used for both regression and classification problems. A decision tree uses a tree-like structure of decisions along with their possible consequences and outcomes. In this, each internal node is used to represent a test on an attribute; each branch is used to represent the outcome of the test. The more nodes a decision tree has, the more accurate the result will be.

The advantage of decision trees is that they are intuitive and easy to implement, but they lack accuracy.

Chosen Model: The Gradient Boosting Model gave the highest accuracy and least misclassification rate for the, so it was chosen as the final model.

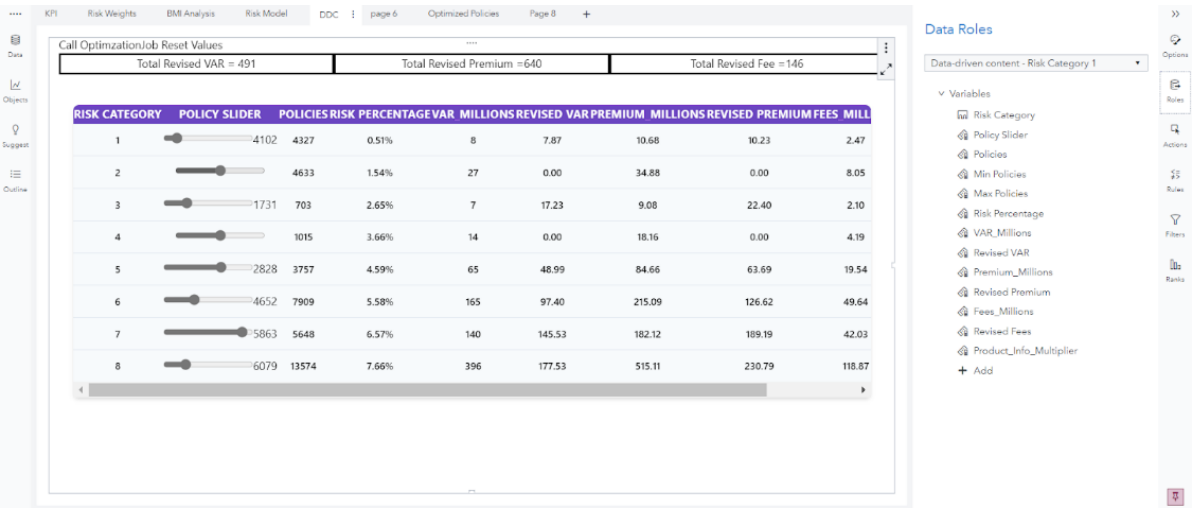


- **SAS Jobs and Optimization**

We also had to recommend the optimal number of policies the company should have to maximize its gains while reducing its value at risk. We used the SAS PROCOPT model which used a linear solver to get the answer. We provided different constraints to our model like minimum number and maximum number of policies required in each risk category, or the sum of total value at risk should be below a certain number.

```
/*    constraints */

/*    num max_Fees, min_RiskPer, max_Prem, min_VARisk;   */
/*    con Fee_con: sum{i in RiskCat}Fee[i]*OPTSOL[i]*Policy[i] >= 250; */
/*    con Prem_con: sum{i in RiskCat} Prem[i]*Policy[i]*OPTSOL[i] >= 1000; */
   con Policylb_con: sum{i in RiskCat}Policy[i]*OPTSOL[i] >= 41566;
   con Policyub_con: sum{i in RiskCat}Policy[i]*OPTSOL[i] <= 50000;
/*    con GainUb_con: sum{i in RiskCat}Policy[i]*RiskPer[i]*0.5*(1-OPTSOL[i]) >=0; */
```

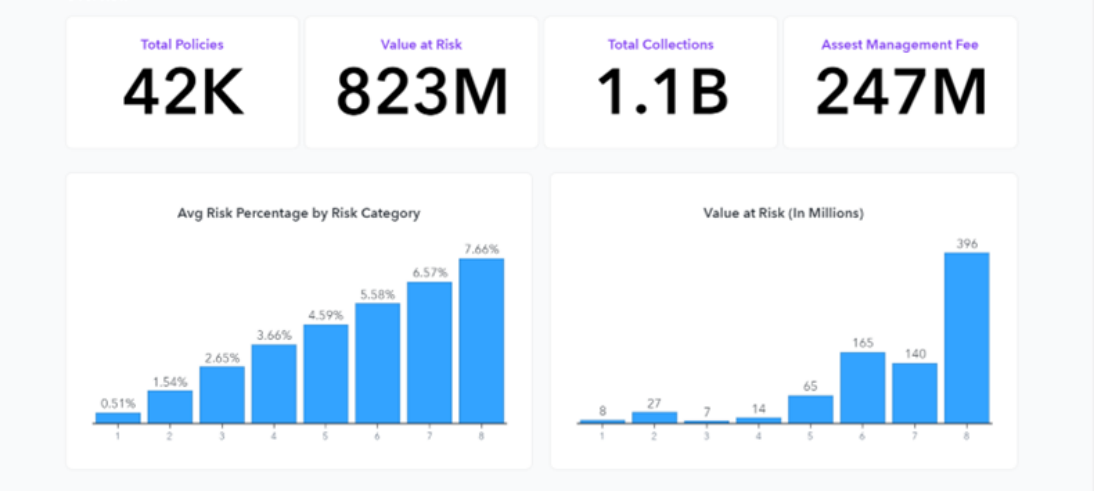


RiskCat	Risk Percentage	Actual Policies	Policies Post Optimization	Actual Premium Collection	Premium Collection (Post Optimization)	Actual Value at Risk	Value at Risk (Post Optimization)	Assest Mgmt Fees	Fees Post Optimization	Gains post optimization (In Mil $)
1	0.51%	4327	4000	$14.21	$13.26	$11.00	$10.20	$3.28	$3.06	$-0.22
2	1.54%	4633	4000	$46.38	$40.04	$36.00	$30.80	$10.70	$9.24	$-1.46
3	2.65%	703	1500	$12.09	$25.84	$9.00	$19.87	$2.79	$5.96	$3.17
4	3.66%	1015	1000	$24.16	$23.79	$19.00	$18.30	$5.58	$5.49	$-0.09
5	4.59%	3757	4000	$112.17	$119.34	$86.00	$91.80	$25.88	$27.54	$1.66
6	5.58%	7909	6564	$286.65	$238.07	$221.00	$183.13	$66.15	$54.94	$-11.21
7	6.57%	5648	6000	$241.03	$256.23	$185.00	$197.10	$55.62	$59.13	$3.51
8	7.66%	13574	14000	$675.55	$697.06	$520.00	$536.20	$155.90	$160.86	$4.96
	4.10%	**41566**	**41064**	**$1,412.24**	**$1,413.63**	**$1,087.00**	**$1,087.41**	**$325.90**	**$326.22**	**$0.32**

- **Reports and Dashboards**

To acquire a deeper understanding of the data, dynamic visualisations were created in SAS Visual Analytics using the cleaned data set. SAS Viya is a well-known analytical tool with a user-friendly interface that makes it simple to build interactive visualisations for easy interpretation and effective reporting. The final cleaned data set had 59, 381 occurrences and 118 variables. The dashboard, which was developed using the Prudential insurance data collection, is shown in Figure. Many interactive graphs are shown on the dashboard. In the data set containing the answer variable, this dashboard primarily shows how KPIS, and demographic factors are distributed. For example, the

relationship between BMI, age, weight, and family history and the various risk categories. Such a dashboard offers perceptions into the client data. As a result, the life insurance firm has greater interaction and knowledge of its candidates. In the visualization phase, a life insurance dataset with 42000 rows was used. An overview dashboard page consisting of important statistics and KPI like total policies, value at risk, total collection, asset management fees, profit, turnover etc. An analysis on each measure is performed. A prediction and forecasting model pipeline risk factor prediction was created in Model Studio.

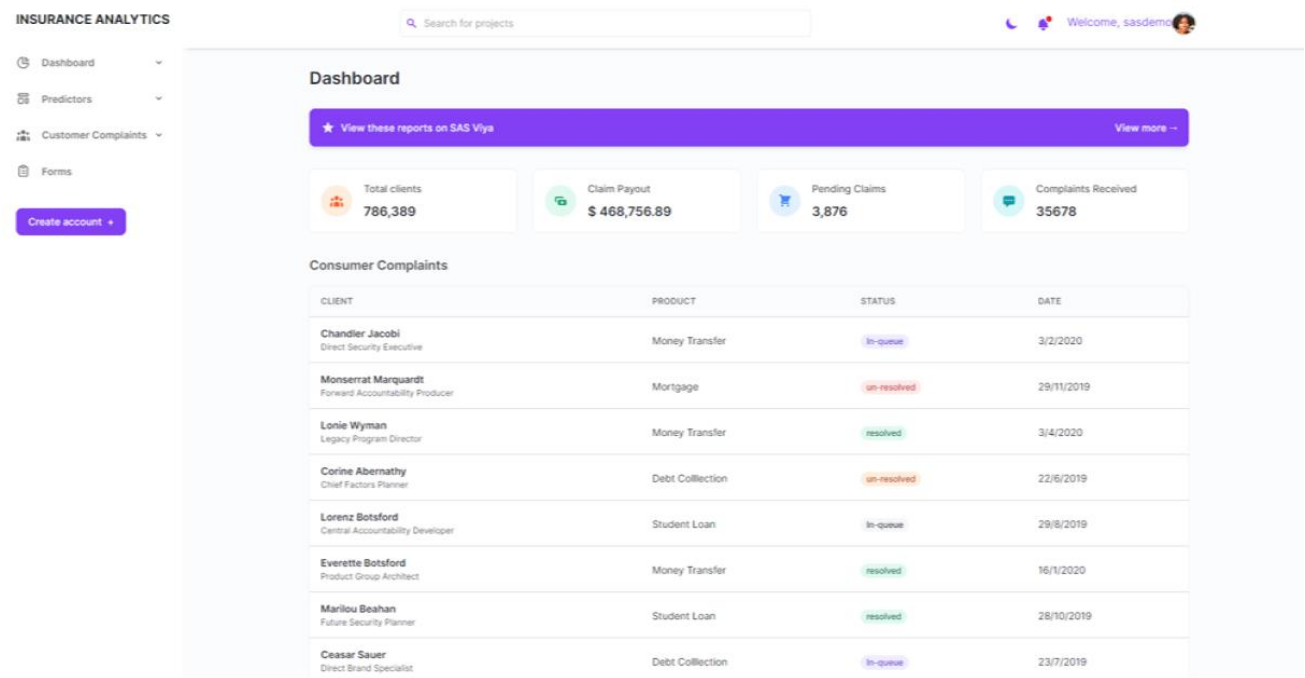


- **Integration in Web Application (ReactJS) & Data Driven Content**

To integrate the different charts and visualisations created, a web application was created using React JavaScript library. We chose React because it is well documented, it is used by many large companies and more importantly, the code is easy to read and to understand which is I think the most important point as the objective is to learn about developing web applications that use SAS REST APIs as their data source.

With that being said, react is designed to create single page applications (SPA). This means that when the end-user connects to the application, the architecture of the page is loaded and then when the user navigates to the different pages, there is no need to query the back-end server except to get data. It makes a clear separation of concerns between the client application and the back-end server. Which is nice because we are using REST APIs to get data from the CAS server, the Compute server, the Viya Job and from MAS.

The application, in the snippets, has different pages. Each page represents a specific option and reuse the same components: A login page for OAuth authentication to the SAS server, a dashboard page which give an overview of all the visualizations made, a KPI (Key Performance Indicators) page and DDC (Data Driven Object Page).



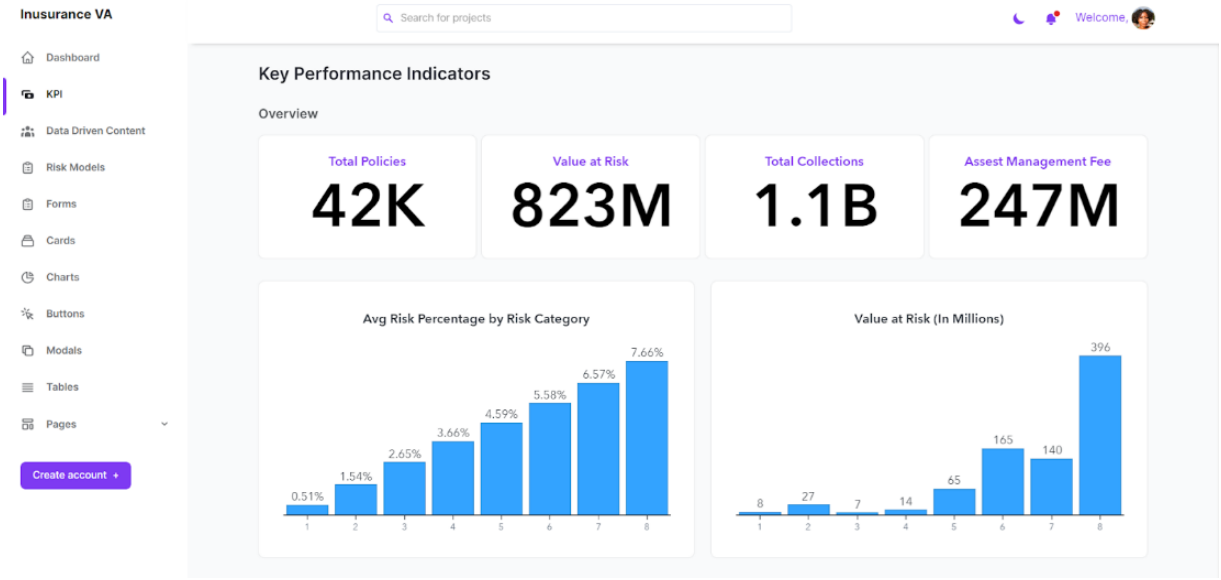
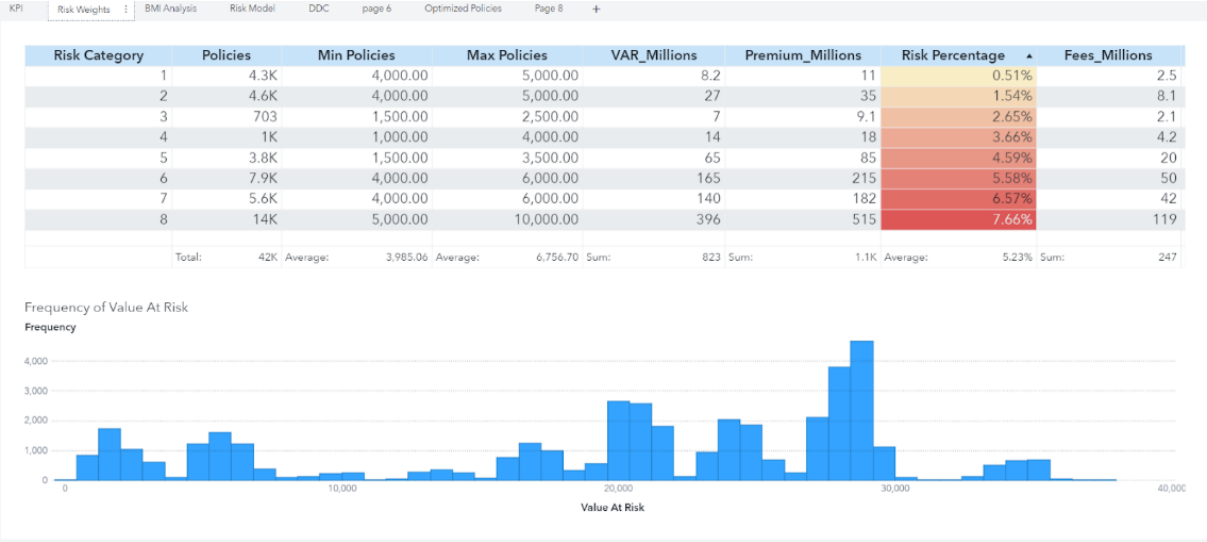


Fig 4.3 KPI Dashboard



Risk Category	Policies	Min Policies	Max Policies	VAR_Millions	Premium_Millions	Risk Percentage	Fees_Millions
1	4.3K	4,000.00	5,000.00	8.2	11	0.51%	2.5
2	4.6K	4,000.00	5,000.00	27	35	1.54%	8.1
3	703	1,500.00	2,500.00	7	9.1	2.65%	2.1
4	1K	1,000.00	4,000.00	14	18	3.66%	4.2
5	3.8K	1,500.00	3,500.00	65	85	4.59%	20
6	7.9K	4,000.00	6,000.00	165	215	5.58%	50
7	5.6K	4,000.00	6,000.00	140	182	6.57%	42
8	14K	5,000.00	10,000.00	396	515	7.66%	119
	Total: 42K	Average: 3,985.06	Average: 6,756.70	Sum: 823	Sum: 1.1K	Average: 5.23%	Sum: 247

Fig 4.5 Risk percentage calculation DDC Object

FRAUD AND CLAIMS ANALYSIS: FRAUDULENT CLAIMS CLASSIFICATION AND CUSTOMER CHURN MODEL

Fraud Prediction and Analysis:

Fraud is a significant challenge faced by the insurance industry. Fraudulent claims lead to significant losses for insurance companies and can also impact premiums for honest policyholders. Data analytics has emerged as a valuable tool to identify and prevent fraud in the insurance industry. By leveraging advanced analytics techniques, insurers can identify patterns and anomalies in data that may indicate fraudulent activity. Predictive modeling can be used to identify claims that are likely to be fraudulent and flag them for further investigation. Machine learning algorithms can also be trained on historical data to identify patterns that may indicate fraudulent behavior. By detecting fraud early, insurers can reduce losses and prevent future fraudulent claims. Fraud analytics also helps insurers to build trust with their customers by ensuring that only valid claims are paid out, and policyholders are protected against losses caused by fraudulent activities. Overall, fraud prediction and analysis is a crucial aspect of risk management for insurance companies, and data analytics is a powerful tool in the fight against fraud.

Fraud prediction in the automobile insurance industry involves using advanced analytics techniques to identify fraudulent claims submitted by policyholders. The use of data analytics helps insurance companies to detect fraudulent claims more quickly and accurately, thereby reducing financial losses caused by fraudulent activities.

The automobile database typically contains various columns such as age, auto year, bodily injuries, capital gains, capital loss, and incident hour. These columns provide important information about the policyholder, the insured vehicle, and the circumstances surrounding the claim.

By analyzing this data, we developed predictive models that can identify patterns and anomalies in the data that may be indicative of fraudulent activity. For example, policyholders who submit claims with high capital gains or capital losses may be more likely to commit fraud as they may be seeking to recoup financial losses.

Similarly, policyholders who report accidents during unusual hours or who have a history of making multiple claims for bodily injuries may also be more likely to engage in fraudulent activity. By identifying these patterns and anomalies, insurance companies can prioritize and investigate claims that have a higher likelihood of being fraudulent, thereby reducing the overall risk of fraudulent activity

Customer Churn Prediction and Claims Analysis:

Customer churn refers to the number of customers who stop using a company's products or services over a given period. In order to predict customer churn in an insurance company, various factors such as the number of premiums paid, age, underwriting score, count of 3-6 months late, count of 6-12 months late, income can be considered. These factors can provide insights into the likelihood of a customer churning.

For instance, customers who have a history of making late payments or have a lower underwriting score may be at a higher risk of churning. Similarly, customers who have been with the company for a longer time and have a higher number of premiums paid are more likely to stay with the company.

By analyzing these factors, companies can identify the customers who are at a higher risk of churning and take proactive measures to retain them. This can include offering incentives, providing better customer service, or introducing new products or services that cater to their needs. By reducing the number of customers who churn, companies can improve their customer retention rates, reduce costs associated with acquiring new customers, and increase their profitability in the long run.

Claims analysis, on the other hand, involves analyzing data related to insurance claims to identify patterns and insights that can help improve the claims management process. This analysis can involve looking at factors such as claim frequency, severity, and duration, as well as identifying common types of claims and their causes. By understanding these patterns, insurance companies can develop strategies to reduce the frequency and severity of claims, improve claims processing times, and identify potential fraud or abuse

Integration with React JS Application:

One for forecasting renewal rates based on information from clients whose insurance had already expired. The second dataset, which contained information on each policy officer's pending claims, was utilised to forecast the fraud rate.

The first dashboard is an overview dashboard giving us an insight on the pending claims, frauds reported previously, the state wise distribution of claims amount and frequency of the make of auto by the incident severity reported during the filling of claims.

The Claims dashboard shows us the distribution of claims amount (vehicular, property and injury) with respect to auto make, policy state the attributes of the customer like age, sex occupations etc.

The Fraud dashboard gives us an overview of the frauds that have taken place in the insurance company and showcases their distribution with respect to accident type, authority contacted, months as a customer etc.

MAS is designed to execute micro requests. It can be used in any web application which needs real-time or close to real-time processing. As soon as a model or decision is published to MAS, you can send data using REST APIs and retrieve the scoring data. If you compare the usage of MAS with the SAS Viya Jobs. Using this service we created a form predicts the renewal rate using our model for a particular customer.

To give you a better view on the model that we are using, it is logistic regression model which evaluates the probability of renewal rate using the following input:

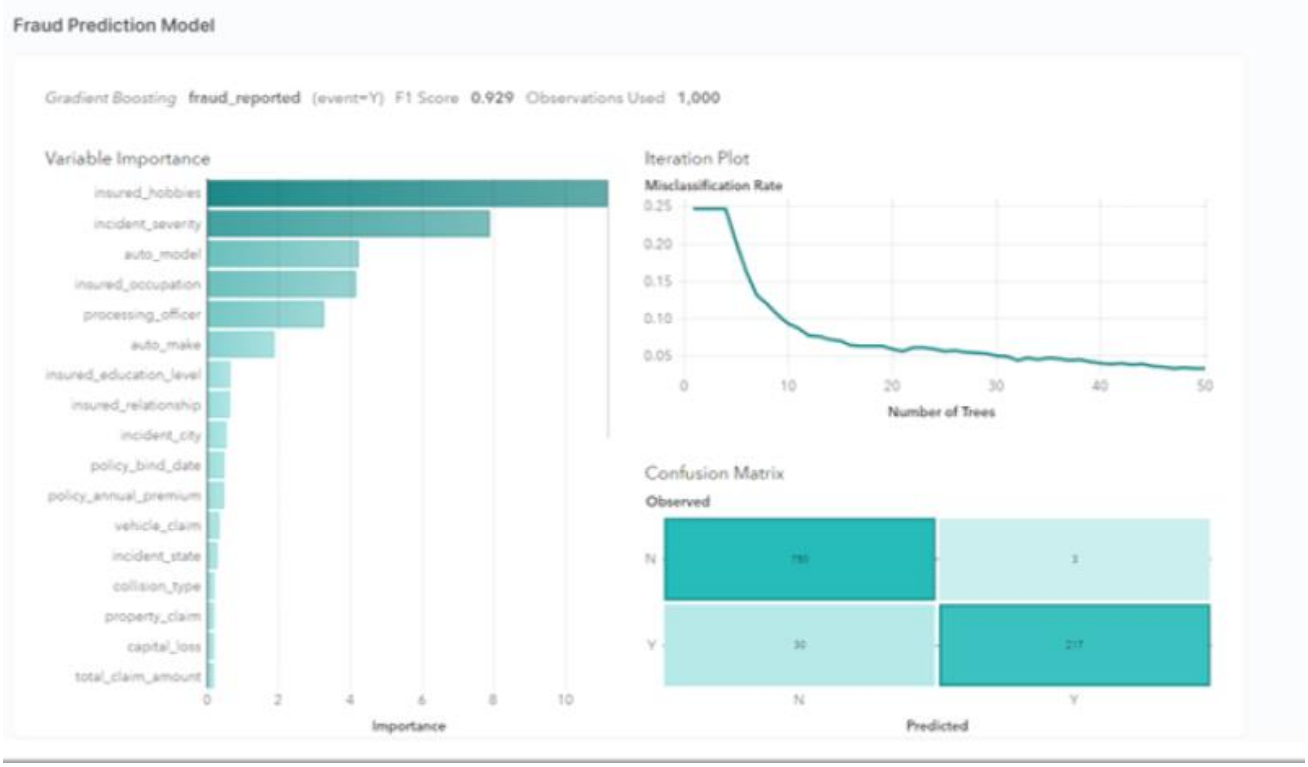
- Cholesterol level
- Diastolic level
- Systolic level
- Cigarettes consumption per day
- Age at start
- Weight
- Sex

Two models for prediction of possible fraud and insurance renewal prediction were also created using these datasets.

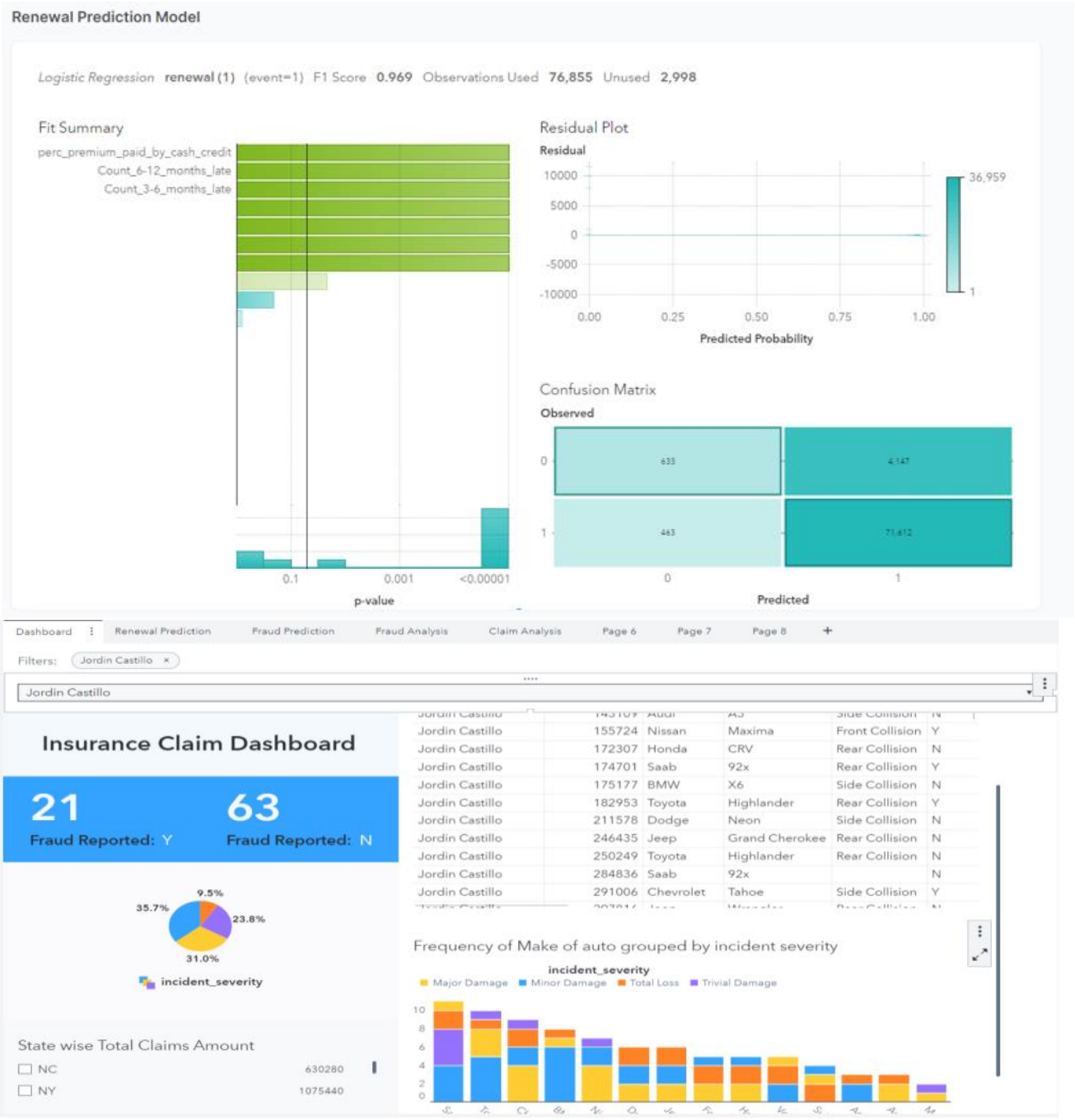
Data was loaded in the model followed by imputation and variable selection.

6 pre-existing models were applied to the dataset and model comparison was performed after applying ensemble learning on all the models.

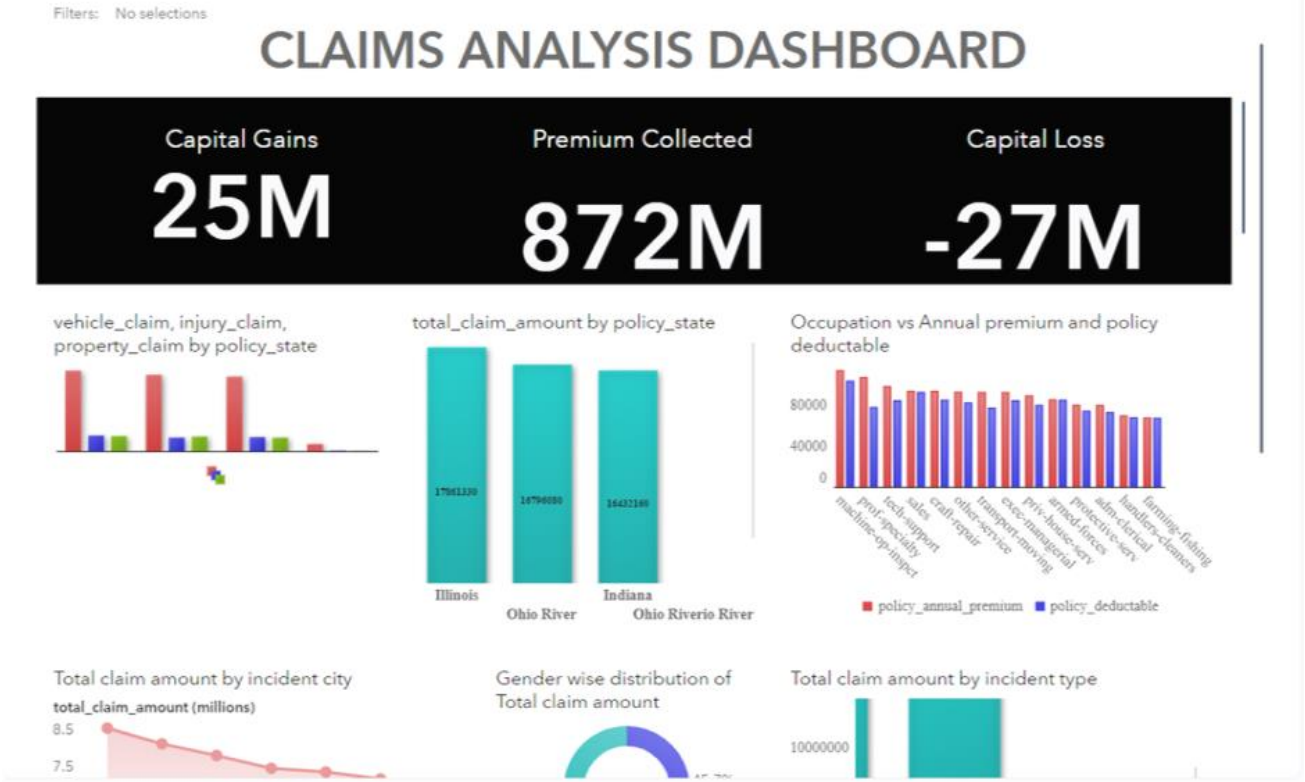
Chosen Model: The Gradient Boosting Model (92%) gave the highest accuracy and least misclassification rate for the prediction of fraud rate, so it was chosen as the final model.

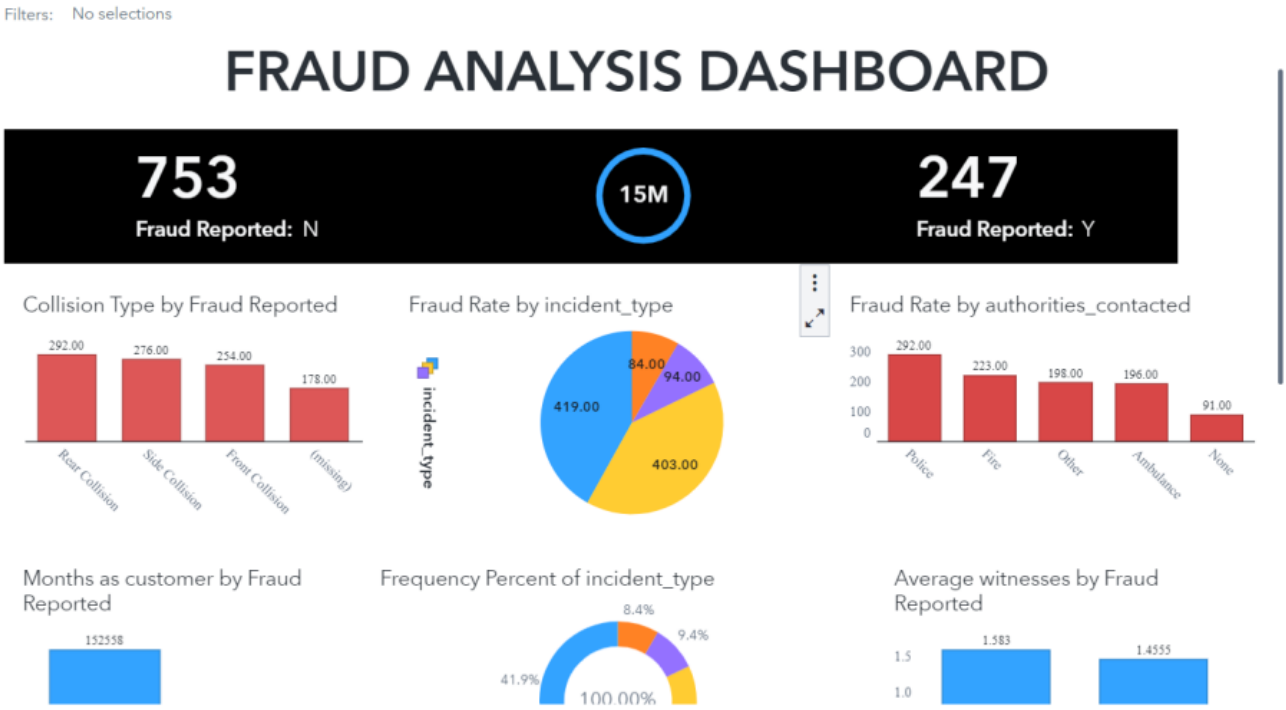


Chosen Model: The Logistic Regression Model (96%) gave the highest accuracy and least misclassification rate for the prediction of renewal rate, so it was chosen as the final model.


Renewal Prediction Model
Logistic Regression renewal (1) (event=1) F1 Score 0.969 Observations Used 76,855 Unused 2,998
Fit Summary
perc_premium_paid_by_cash_credit
Count_6-12_months_late
Count_3-6_months_late
0.1
0.001
<0.00001
p-value
Residual Plot
Residual
10000
5000
0
-5000
-10000
0.00
0.25
0.50
0.75
1.00
Predicted Probability
36,959
1
Confusion Matrix
Observed
633
4,147
463
71,612
0
1
Predicted
Dashboard
Renewal Prediction
Fraud Prediction
Fraud Analysis
Claim Analysis
Page 6
Page 7
Page 8
Filters:
Jordin Castillo
Jordin Castillo
Insurance Claim Dashboard
21
Fraud Reported: Y
63
Fraud Reported: N
9.5%
35.7%
23.8%
31.0%
incident_severity
State wise Total Claims Amount
NC 630280
NY 1075440
Jordin Castillo 155724 Nissan Maxima Front Collision Y
Jordin Castillo 172307 Honda CRV Rear Collision N
Jordin Castillo 174701 Saab 92x Rear Collision Y
Jordin Castillo 175177 BMW X6 Side Collision N
Jordin Castillo 182953 Toyota Highlander Rear Collision Y
Jordin Castillo 211578 Dodge Neon Side Collision N
Jordin Castillo 246435 Jeep Grand Cherokee Rear Collision N
Jordin Castillo 250249 Toyota Highlander Rear Collision N
Jordin Castillo 284836 Saab 92x N
Jordin Castillo 291006 Chevrolet Tahoe Side Collision Y
Frequency of Make of auto grouped by incident severity
incident_severity
Major Damage
Minor Damage
Total Loss
Trivial Damage
10
8
6
4
2
0

Analysis


Filters: No selections
CLAIMS ANALYSIS DASHBOARD
Capital Gains
25M
Premium Collected
872M
Capital Loss
-27M
vehicle_claim, injury_claim, property_claim by policy_state
total_claim_amount by policy_state
17041330
16796050
16432160
Illinois
Indiana
Ohio River
Ohio Riverio River
Occupation vs Annual premium and policy deductable
80000
40000
0
machine-op-inspct
prof-specialty
tech-support
sales
craft-repair
other-service
transport-moving
exec-managerial
priv-house-serv
armed-forces
protective-serv
adm-clerical
handlers-cleaners
farming-fishing
policy_annual_premium
policy_deductable
Total claim amount by incident city
total_claim_amount (millions)
8.5
7.5
Gender wise distribution of Total claim amount
Total claim amount by incident type
10000000


Filters: No selections
FRAUD ANALYSIS DASHBOARD
753
Fraud Reported: N
15M
247
Fraud Reported: Y
Collision Type by Fraud Reported
292.00
276.00
254.00
178.00
Rear Collision
Side Collision
Front Collision
(missing)
Fraud Rate by incident_type
incident_type
84.00
94.00
419.00
403.00
Fraud Rate by authorities_contacted
300
200
100
0
292.00
223.00
198.00
196.00
91.00
Police
Fire
Other
Ambulance
None
Months as customer by Fraud Reported
152558
Frequency Percent of incident_type
8.4%
9.4%
41.9%
100.00%
Average witnesses by Fraud Reported
1.5
1.0
1.583
1.4555

CONSUMER COMPLAINT CLASSIFICATION AND ANALYSIS USING VISUAL TEXT ANALYTICS

DATA PREPARATION FOR VISUAL ANALYSIS

In this section we have analysed the Consumer Complaint dataset (between the years 2018–2021) obtained from the Consumer Financial Protection Bureau website. We brought that data into SAS Data Studio in which we created an ID column with unique values and in SAS Model studio we began to run some visual text analytics.
For variables such as “State”, “Issue”, and “Product” we assigned them the role of a categorical variable. For “Consumer Narrative “we assigned it to be a text variable as that would be the variable that I would be creating concepts and categories for. Through Model Studio, we was able to create a VTA pipeline, which is a process flow diagram whose nodes represent tasks in the text analytics process. The nodes we focused on during this project were concepts, sentiment, and categories

Concepts and Categories

A concept is a data element or pattern — such as named entities or fact relationships — that you wish to extract from a larger text field because they match a specific context. In VTA, we wrote rules recognizing concepts that were important to the context of the common frauds committed by organizations. The four main concepts we created were: Cannot Access, Closed Account, Fraudulent transactions, Sections mentioned, and Unauthorized Transfers. These concepts were created because as we used visual text analytics to interactively explore the narratives, we found that people tended to have problems with not being able to access their account, banks would close their account without an explanation, customers couldn’t access their account funds, or they saw unauthorized money transfers show up on their bank statements. These concept rules use language interpretation for textual information (LITI) syntax, which includes boolean and distance operators. For example, in the Unauthorized Transfers concept, our rule was Concept_Rule: (Sent, “_c{transferred@}”, (OR, “money@”, “knowledge@”)). This rule looks for any synonyms of transferred in the same sentence with the synonym of money or knowledge as a way to fetch all the consumer narratives in which the consumer complaints of having their money transferred from their account without their knowledge. In the concept node, we also included VTA’s predefined concepts that include the following: Person Names, Location Names, Organization Names, Dates, Times, Currency Amounts, and Percentage Amounts. Overall, concepts are important because they influence the way in which text is parsed as we are pulling out specific pieces of information.

In the VTA pipeline, we also created custom categories through the process of tagging documents as belonging to a specific category using linguistic rules. we created categories such as Managing an Account, Unauthorized Transactions, Fraud or Scam, and Closing Your Account and SAS VTA generated other linguistic rules using supervised learning on how the document issues were manually tagged. Categories are rules generated based on Boolean logic for the presence of specific terms and phrases; the code uses OR, AND, and NOT commands to dictate what keywords will be looked for in the consumer narratives. In the categories tab, we can see all the consumer narratives that are matched and their sentiment

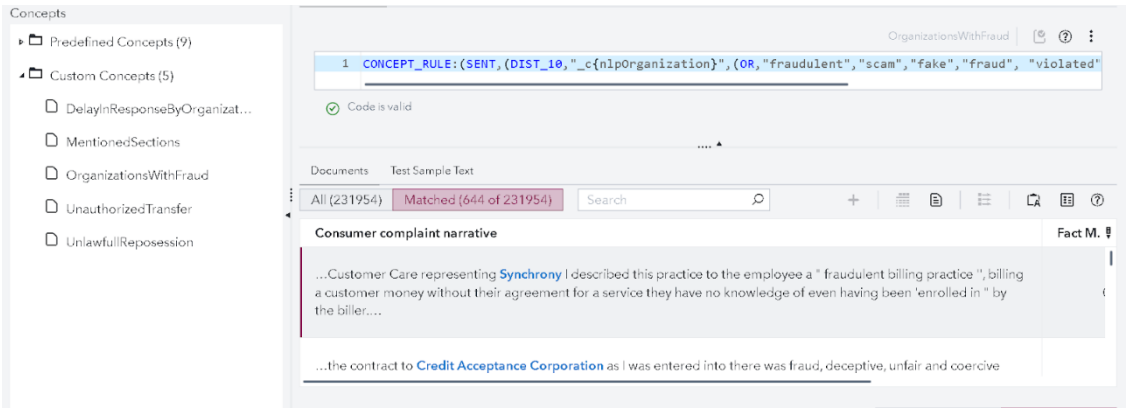


Fig 4.17 Custom Concepts

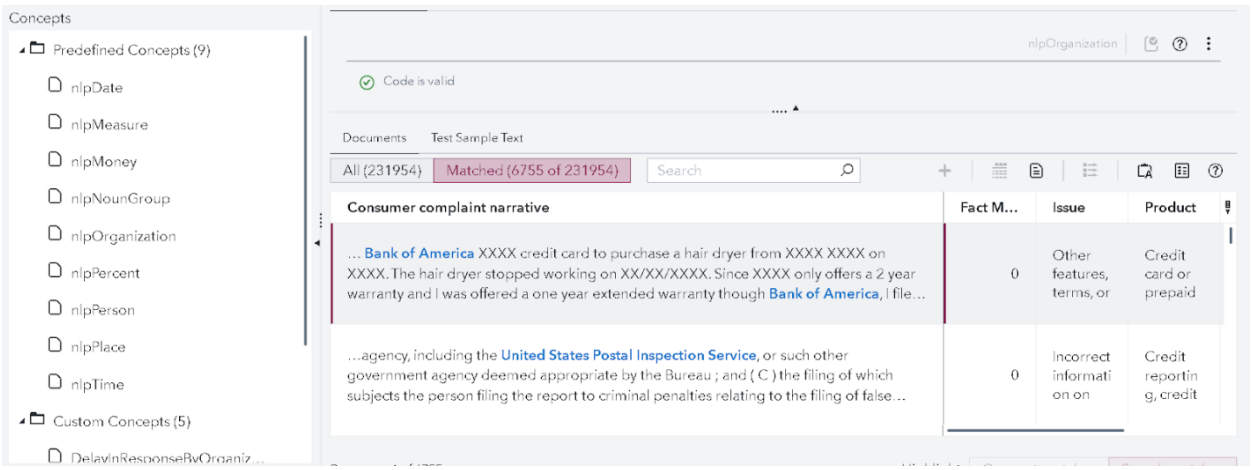


Fig 4.18 Pre-Defined Concepts

INTERACTIVE REPORT GENERATION AND USE

After creating all our concepts and categories, we ran the VTA pipeline to be able to download the score code for categories, concepts, and sentiment from VTA results. The score code was then entered into a code window in SAS Studio and then brought into Visual Analytics to allow exploration and visualization of the data.

The next step was to create visualizations to visualize the distribution patterns in the complaints and the information extracted from the complaint narratives which is shown in the images below

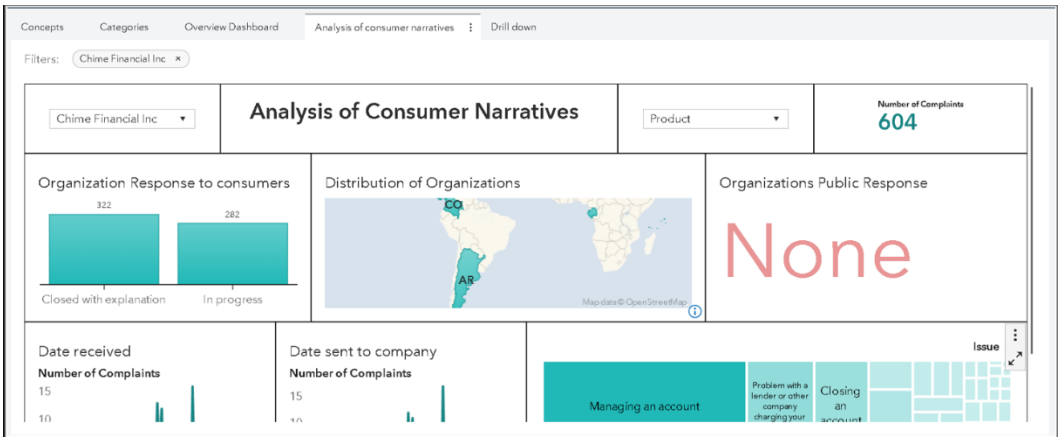


Fig 4.10 Consumer narrative Dashboard

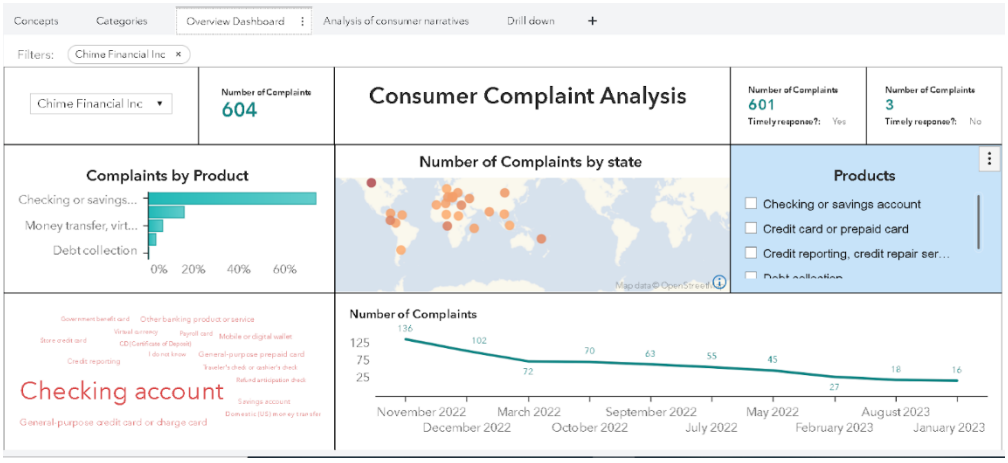


Fig 4.11 Consumer Complaint Dashboard

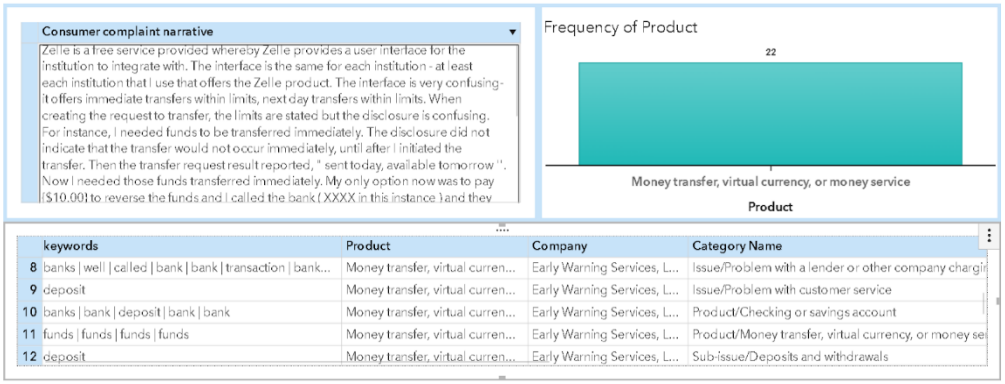


Fig 4.12 Information Extracted By using nlp

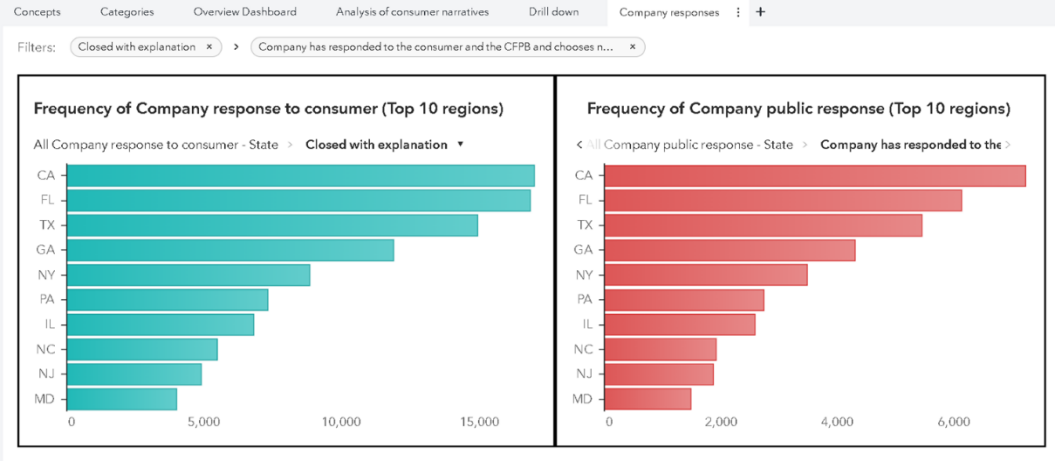


Fig 4.13 Company Response Dashboard

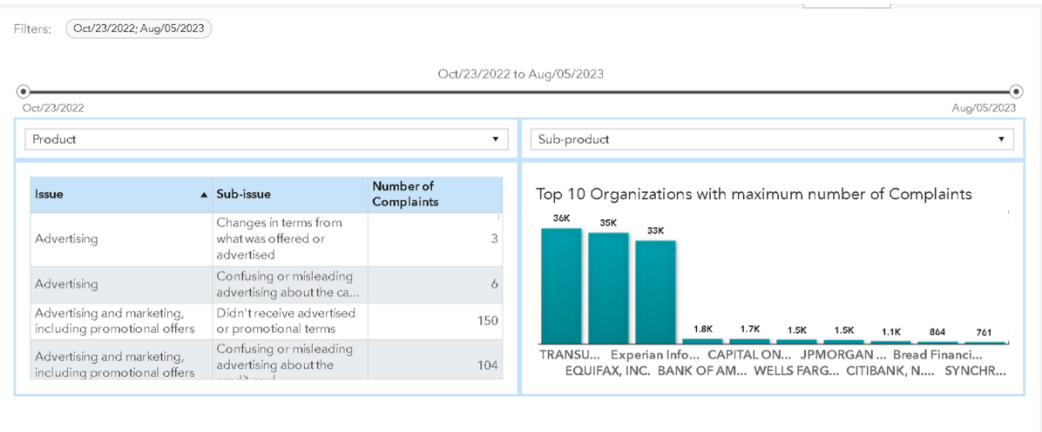


Fig 4.14 Complaint Drill down

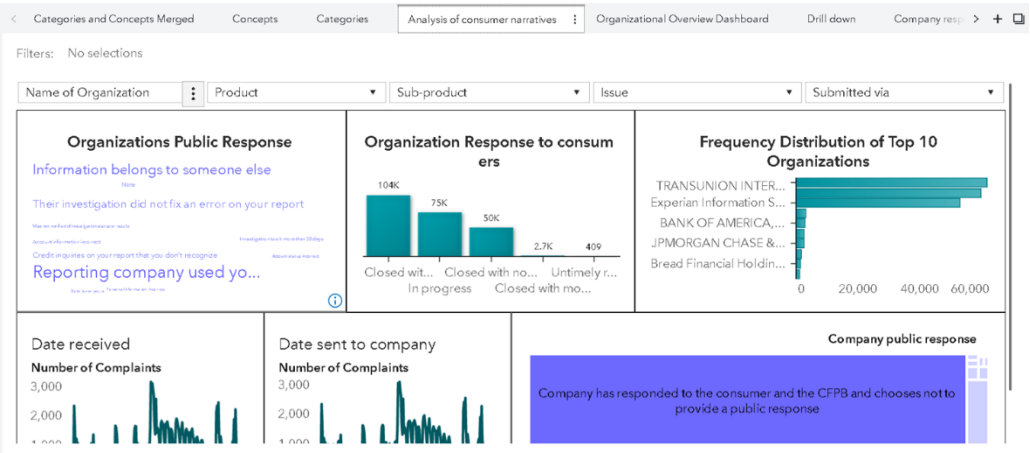


Fig 4.15 Analysis of Consumer Narrative

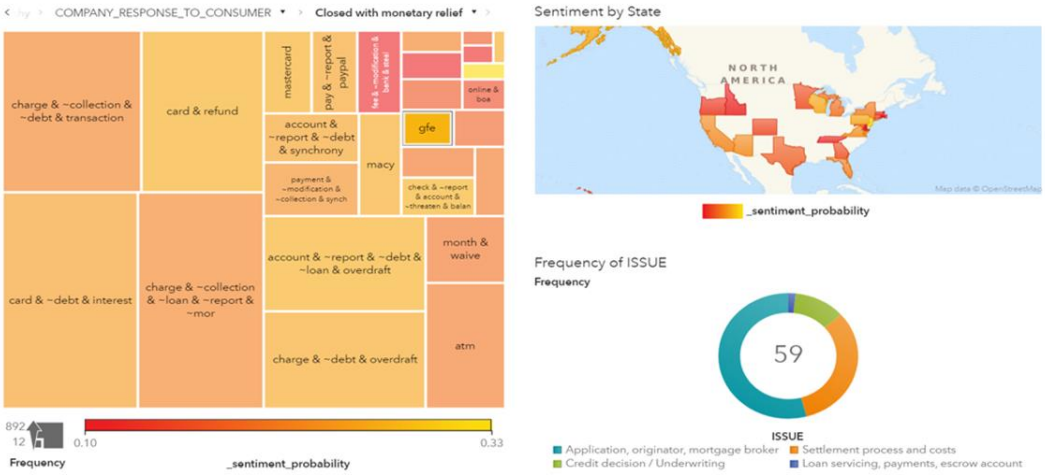


Fig 4.16 Distribution of Consumer Sentiment

CONCLUSION

As technology advances, the ability to analyze massive amounts of data is driving up the requirement within the insurance industry for well governed models that can be quickly deployed. Insurers recognise the opportunity that large quantities of structured and unstructured data from new sources bring to develop AI and Machine Learning models. The benefits of this scale of modeling can include improved pricing, offering more tailored and relevant products, better fraud detection and improvements in customer service.

Machine learning and Natural Language Processing (NLP) have become increasingly important tools for the insurance industry in predicting fraudulent claims, identifying potential customer churn, and extracting valuable insights from customer data. With the explosion of digital data, these advanced technologies enable insurers to analyze vast amounts of unstructured data, such as social media posts, emails, and customer service conversations, to detect patterns and anomalies that might indicate fraudulent activity or dissatisfaction among customers. By analyzing customer behavior and sentiment, insurers can proactively intervene to prevent customer churn and improve the overall customer experience. Machine learning and NLP are also useful in extracting relevant information from customer data, allowing insurers to personalize their offerings and improve their risk assessment models. As the insurance industry continues to evolve and become more data-driven, these technologies will play an increasingly critical role in improving operational efficiency and profitability while enhancing customer satisfaction. The ability to manage model risk is an existential challenge. With the advent of new regulations and new technology driving complexity within models, as well as organizational challenges such as risk and finance integration, the old approach – using multiple shared files and spreadsheets together with generic governance solutions – simply doesn't scale.

Enterprise model risk management requires firms to go beyond a checklist approach toward understanding model risk dynamically and in context. Any model risk solution also needs to manage a wide range of model types, including open source, multivendor and AI and Machine Learning models.The use of big data analytics in the insurance industry is rising. Insurance companies invested $3.6 billion in 2021. Companies who invested in big data analytics have seen 30%more efficiency, 40% to 70% cost savings, and a 60%increase in fraud detection rates. Both the customers and companies benefit from these solutions, allowing insurance companies to target their customers more precisely.

REFERENCES

1. Website of the Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/. Accessed on February 1, 2017.
2. Yeo, A.C., Smith, K.A., Willis, R.J. and Brooks, M., 2001. Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. Intelligent Systems in Accounting, Finance & Management, 10(1), pp.39- 50.
3. Boodhun, N. and Jayabalan, M., 2018. Risk prediction in life insurance industry using supervised learning algorithms. Complex & Intelligent Systems, 4(2), pp.145- 154.
4. Wang, Y., Li, B., Li, G., Zhu, X. and Li, J., 2019. Risk factors identification and evolution analysis from textual risk disclosures for insurance industry. Procedia Computer Science, 162, pp.25-32.
5. Poornima, A. and Priya, K.S., 2020, March. A comparative sentiment analysis of sentence embedding using machine learning techniques. In 2020 6th international conference on advanced computing and communication systems (ICACCS) (pp. 493- 496). IEEE.
6. Verma, A., Taneja, A. and Arora, A., 2017, August. Fraud detection and frequent pattern matching in insurance claims using data mining techniques. In 2017 tenth international conference on contemporary computing (IC3) (pp. 1-7). IEEE.
7. Tao, H., Zhixin, L. and Xiaodong, S., 2012, October. Insurance fraud identification research based on fuzzy support vector machine with dual membership. In 2012 international conference on information management, innovation management and industrial engineering (Vol. 3, pp. 457-460). IEEE.
8. Burri, R.D., Burri, R., Bojja, R.R. and Buruga, S., 2019. Insurance claim analysis using machine learning algorithms. International Journal of Innovative Technology and Exploring Engineering, 8(6S4), pp.577-82.
9. Spiteri, M. and Azzopardi, G., 2018, September. Customer churn prediction for 29 a motor insurance company. In 2018 Thirteenth International Conference on Digital Information Management (ICDIM) (pp. 173-178). IEEE.
10. Jayadi, R., Kelvin, A., Rifyansyah, P., Mufarih, M. and Firmantyo, H.M., 2020, September. Predicting Customer Churn of Fire Insurance Policy: A Case Study in an Indonesian Insurance Company. In 2020 6th International Conference on Science and Technology (ICST) (Vol. 1, pp. 1-4). IEEE.
11. He, Y., Xiong, Y. and Tsai, Y., 2020, April. Machine learning based approaches to predict customer churn for an insurance company. In 2020 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-6). IEEE.
12. Soares, J.H.C., Barbosa, J.L.N., Lopes, L.A., Júnior, G.V.M., Rabêlo, R.D.A.L., Passos, E.B. and dos Santos Neto, P.D.A., 2019, June. How to Avoid Customer Churn in Health Insurance/Plans? A Machine Learn Approach. In 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS) (pp. 559- 562). IEEE

13. Sabo, Tom. 2014. “Uncovering Trends in Research Using Text Analytics with Examples from
Nanotechnology and Aerospace Engineering.” *Proceedings of the SAS Global Forum 2014 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS061-2014.pdf.

14. Sabo, Tom. 2015. “Show Me the Money! Text Analytics for Decision-Making in Government Spending.” *Proceedings of the SAS Global Forum 2015 Conference*. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings15/SAS1661-2015.pdf.

15. Sabo, Tom. 2016. “Extending the Armed Conflict Location and Event Data Project with SAS® Text Analytics.” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf.

16. “Consumer Complaint Database" Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data. Accessed on February 1, 2017.

17. “How we improved the disclosures” Consumer Financial Protection Bureau. Available http://www.consumerfinance.gov/know-before-you-owe/compare/. Accessed on February 1, 2017.

18. Sabo, Tom. 2014. SAS Institute white paper. “Text Analytics in Government: Using Automated Analysis to Unlock the Hidden Secrets of Unstructured Data.” Available http://www.sas.com/en_us/whitepapers/text-analytics-in-government-106931.html.

19. Albright, Russ. Cox, James. Jin, Ning. 2016. “Getting More from the Singular Value Decomposition (SVD): Enhance Your Models with Document, Sentence, and Term Representations” *Proceedings of the SAS Global Forum 2016 Conference*. Cary NC: SAS Institute Inc. Available https://support.sas.com/resources/papers/proceedings16/SAS62412016.pdf.

20. Osborne, Mary. Maness, Adam. 2014. “Star Wars and the Art of Data Science: An Analytical

21. Approach to Understanding Large Amounts of Unstructured Data.” Proceedings of the SAS Global Forum 2014 Conference. Cary NC: SAS Institute Inc. Available http://support.sas.com/resources/papers/proceedings14/SAS286-2014.pdf.

22. Chakraborty, G., M. Pagolu, S. Garla. 2013. Text Mining and Analysis; Practical Methods, Examples, and Case Studies Using SAS®. SAS Institute Inc.

23. Reamy, Tom. 2016. Deep Text; Using Text Analytics to Conquer Information Overload, Get Real Value from Social Media, and Add Big(ger) Text to Big Data. Medford NJ: Information Today, Inc.

ACKNOWLEDGMENTS

Thanks to Prof. VIJAYETHA THODAY and Mr. AJAY PANJWANI for providing insights into the visualizations that would best convey the analytics results of this project to a wider audience.