The Vital Role of Insurance Claims in Protecting Your Present and Future

Introduction

Insurance is a crucial financial protection that provides coverage against unexpected losses for individuals and businesses. When policyholders experience incidents causing harm, damage, or loss, they file insurance claims seeking compensation from insurance providers per the policy terms. However, disputes can arise regarding claim approval, coverage limits, timeliness of payment, and more. These disputes manifest as formal complaints lodged by policyholders against insurance providers.

Analyzing insurance complaints and claims data can reveal meaningful insights into crucial pain points in the claims process from the consumer's perspective. It can also uncover differences in complaint types, confirmation rates, and resolution times across various insurance products. For Texas specifically, insurance complaints data aggregated at the state government level provides a unique window into regional consumer protection issues.

Studying insurance complaints and claims to identify frequent complaint triggers, products seeing more disputes, and factors impacting complaint resolution timelines. The goal was to unpack the consumer-side dynamics of the insurance claims process. Deriving data-driven intelligence on pain points can better inform policyholders and insurance providers on improving claims outcomes and preventing issues proactively where possible. These two aspects play a crucial role in the financial industry, offering valuable insights into customer satisfaction, regulatory compliance, and potential fraudulent activities. It is crucial to discern the differences between complaints and claims, as each serves distinct purposes in evaluating insurance processes, customer experiences, and the overall integrity of the insurance system. Overall, insurance complaints serve as an essential feedback channel that, if studied systematically, can enhance financial protection for citizens relying on risk transfer safeguards (Coalition Against Insurance Fraud).

Why?

Insurance (Nathan Weller)

- Insurance spending contributes to 13% of the GDP
- $5.2 trillion of insurance premiums each year
- 9.1% premium increase annually

Denial (Maria Clark)

- 1 in 7 of all claims get denied
- 30% of claims get denied on initial submission
- $262 billion in claims were denied

Fraud (Coalition Against Insurance Fraud)

- $308.6 billion stolen from consumers each year
- About 9,000 cars were intentionally set on fire to receive a payout
- $3.1 billion in false and fraudulent claims in 2020
- 48 states make insurance fraud a specific crime

Importance of the Problem

Financial Losses

Insurance fraud can lead to significant financial losses for insurance companies and the insured. These losses are often converted to higher premiums, affecting insurance affordability for everyone.

Economic Impact

It is possible to distort the market dynamics from claims frauds by increasing the cost of honest policyholders and businesses to compensate for the cost of frauds.

Legal and Regulatory Implication

Insurance fraud is considered a criminal offense in about 48 states because there are many factors that can affect the insurance industry, and significant legal and regulatory resources are required to investigate and prosecute.

Consumer Trust

Frequent occurrences of fraud from companies can erode consumer trust in the insurance company and the industry itself. This skepticism from customers can lead to decreased insurance purchases, which also increases risk for individuals.

Social Cost

If insurance fraud is not addressed, it can create an environment where unethical actions are practiced. This ignorance can lead to a domino effect where fraudulent activities continue.

Goals

Risk Prediction and Prevention

Develop predictive models to identify potential risks and patterns leading to insurance claims. Using historical data to predict and prevent future incidents can reduce claims' frequency and severity.

Fraud Detection and Prevention

Implement advanced analytics and machine learning algorithms to detect and prevent fraudulent insurance claims. By identifying and analyzing unusual patterns, insurers can enhance fraud detection mechanisms by saving resources and maintaining the integrity of the insurance system.

Utilize Predictive Modeling to Analyze Insurance Claim Data

The insurance industry can create predictive models and accurately forecast future outcomes by studying patterns.

1) Increasing insurance claim volume requires attention for the ability to prioritize. By implementing Artificial intelligence, insurance companies can rank claims by risk and severity so that they know what to focus on.
2) Assign the right level of experience to examine the claim appropriately. The low-cost claim requires minimal case management where artificial intelligence can be executed. Whereas human resources should manage complex and potentially high-cost claims.

Data

Objective

The objective of these datasets is to analyze insurance claims and insurance complaints in the United States. Datasets are gathered from The Texas Department of Insurance (TDI) and various insurance providers.

Data Source #1

The first dataset is sourced from the Texas Department of Insurance (TDI). Its primary focus is complaints against people, companies, agents, and adjusters. It contains information for each person and organization named in a complaint.

Features/Variables

Initially, this dataset had 242,000 rows and 17 columns such as:

- **Complaint number**: The number assigned to a specific complaint.
- **Complaint filed against**: The name of the person or organization the complaint was filed against
- **Complaint filed by**: Shows who filed the complaint.
- **Reason complaint filed**: Shows the reason the complaint was filed.
- **Confirmed complaint**: A "Yes" answer TDI confirmed the licensed person or organization was in error
- **How resolved**: Gives a brief description of how the complaint was resolved
- **Received date**: The date TDI received the complaint
- **Closed date**: The date TDI closed the complaint
- **Complaint type**: Shows if the complaint was about "property and Casualty" or "Life, Accident and Health" insurance

- **Coverage type**: Shows if the coverage was “Automobile.” “Homeowners.” “Accident and Health.” “Life & Annuity,” or “Miscellaneous.”
- **Coverage level**: Shows if the coverage was “Property or Casualty” or “Life, Accident and Health.”
- **Others involved:** Shows the other types of people and organizations involved in the complaint.
- **Respondent ID**: The number assigned to the person or organization the complaint was filed against
- **Respondent Role**: Shows the role of the person or organization that a complaint was filed against
- **Respondent type**: Shows if the complaint was filed against a person or organization
- **Complainant type**: Shows if the complaint came from a person or organization

https://data.texas.gov/dataset/Insurance-complaints-All-data/ubdr-4uff

Dataset Source #2

This dataset is a comprehensive collection of insurance claim records, with each row representing an individual claim and the column representing various features associated with that claim. This claims data is collected from various insurance providers, with information about the individual's background, claim specifics, associated documentation, and feedback from insurance companies. In addition, the dataset includes data about indicators and parameters that were examined during the claim's assessment, giving a broader look into the complexities of each claim. Each row is presented with a unique ID to ensure data privacy.

Features/Variable

This dataset contains 1001 rows and 40 columns

- **Months_as_customer**
- **Age**
- **Policy_number**
- **Polic_state**
- **Policy_deductable**
- **Incident type**
- **Coverage type**
- **Reason_complaint_filed**
- **Fraud_reported**
- **Confirmed fraud**

**The insurance claims in this dataset are subjected to rigorous examination, encompassing both manual assessments and automated checks. The result of this examination, specifically whether a claim was deemed fraudulent or not, is indicated for each record.*

https://data.mendeley.com/datasets/992mh7dk9y/2

Data Methodology:

Dataset:

A series of data reprocessing steps were taken to clean and prepare the data. By using essential libraries such as Pandas for data manipulation, scikit-learn for model selection and evaluation, Matplotlib, Seaborn, and RandomForestClassifier for classification, we were able to classify the target variable. The first step in reprocessing the dataset involved dropping unnecessary columns such as 'policy_number', 'policy_bind_date', 'incident_date', 'incident_location', and '_c39'. This step streamlined the dataset by removing irrelevant or redundant information. Next, we categorized variables within the dataset and used LabelEncoder to encode them. By doing this, we converted text-based categories into a numerical format that can be processed by machine learning algorithms. To prepare for the machine learning model, we defined the feature set(X) and the target variable (y), with 'fraud-reported' being the target variable. Then the data was split into a large testing set (80%) and a smaller training set (20%), using a random state for reproducibility.

Two classification models were created: RandomForestClassifier and Support Vector Machine. We initialized each model to train on the training set and then used it to make predictions on the test set. The RandomForestClassifier was configured with 100 estimators, and the SVM used default settings. The performances of both models were evaluated using metrics such as accuracy score, confusion matrix, and classification report. These results were printed, offering a comparative insight into the effectiveness of each model in predicting the target variable.

Additionally, a confusion matrix for the Random Forest model was visualized using Seaborn, displaying the matrix as a heatmap. By analyzing this visualization, we can understand the model's performance in terms of true positives, true negatives, false positives, and false negatives, with specific labels for "fraud' and 'No Fraud' both for predicted and actual values.

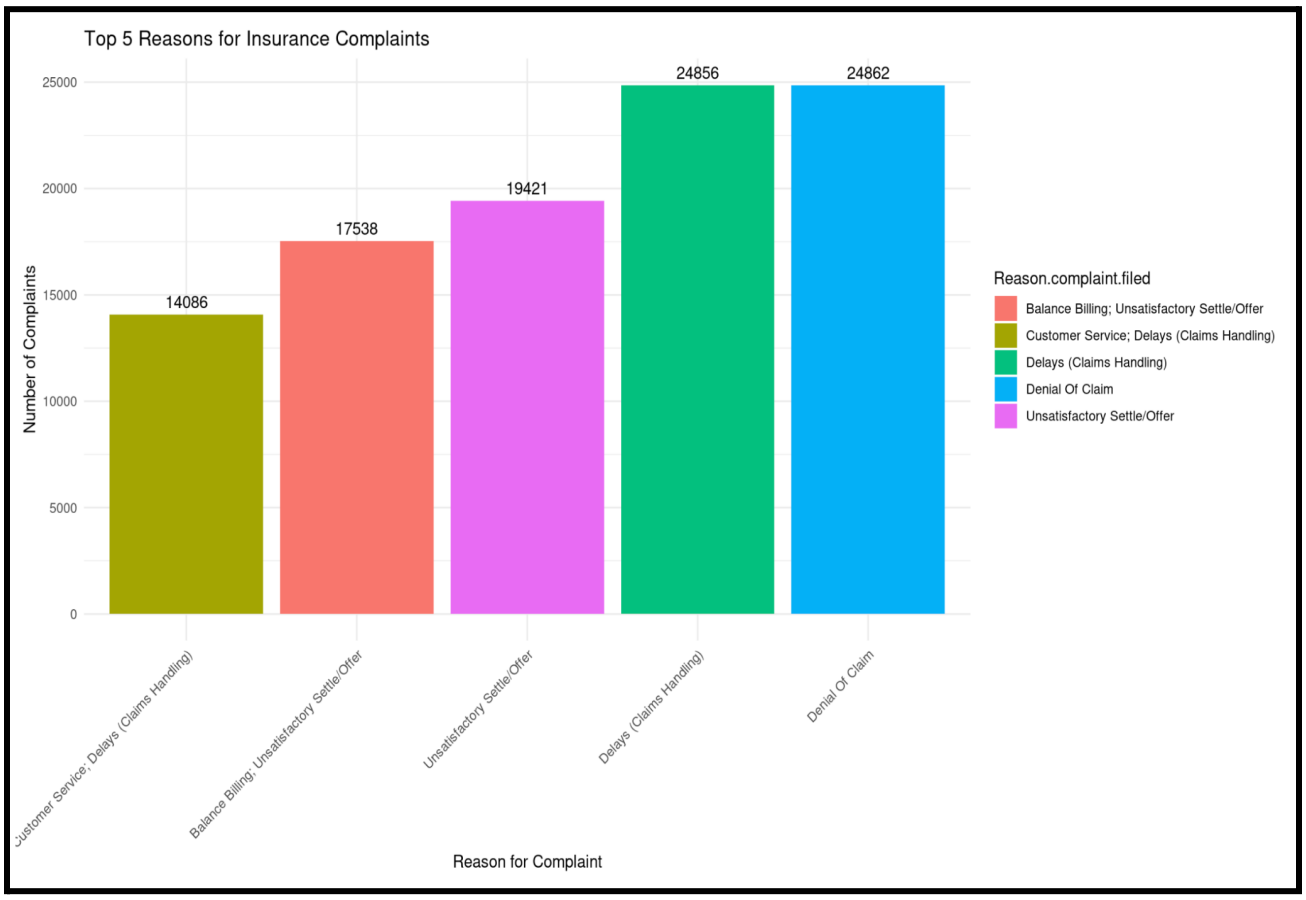
Specifically, we used the ‘feature_importances_’ attribute during the post-training to identify which features were most influential in the model’s decision-making process. To visualize the findings, the importance was then organized into a DataFrame sorted in descending order of importance alongside their corresponding feature names to highlight the most influential features. Finally, we used a horizontal bar plot from the Matplotlib library with feature names on the y-axis and their corresponding importance scores on the x-axis.

We also created a visualization using Seaborn and Matplotlib to explore the relationship between the incident severity and the frequency of fraud reporting in the dataset. By focusing on the ‘incident_severity’ column with the hue parameter set to ‘fraud_reported’ effectively segmenting the data based on the presence or absence of reported fraud, we were able to offer a clear and intuitive understanding of how severity correlates with the frequency of reported fraud in the dataset.

Exploratory Data Analysis

Hypothesis 1

Insurance claim denial is the top reason for insurance complaints. A common reason that claims are denied is lack of sufficient evidence or misrepresentation of evidence. This misrepresentation of evidence can be considered fraud, and insurance companies investigate these activities.

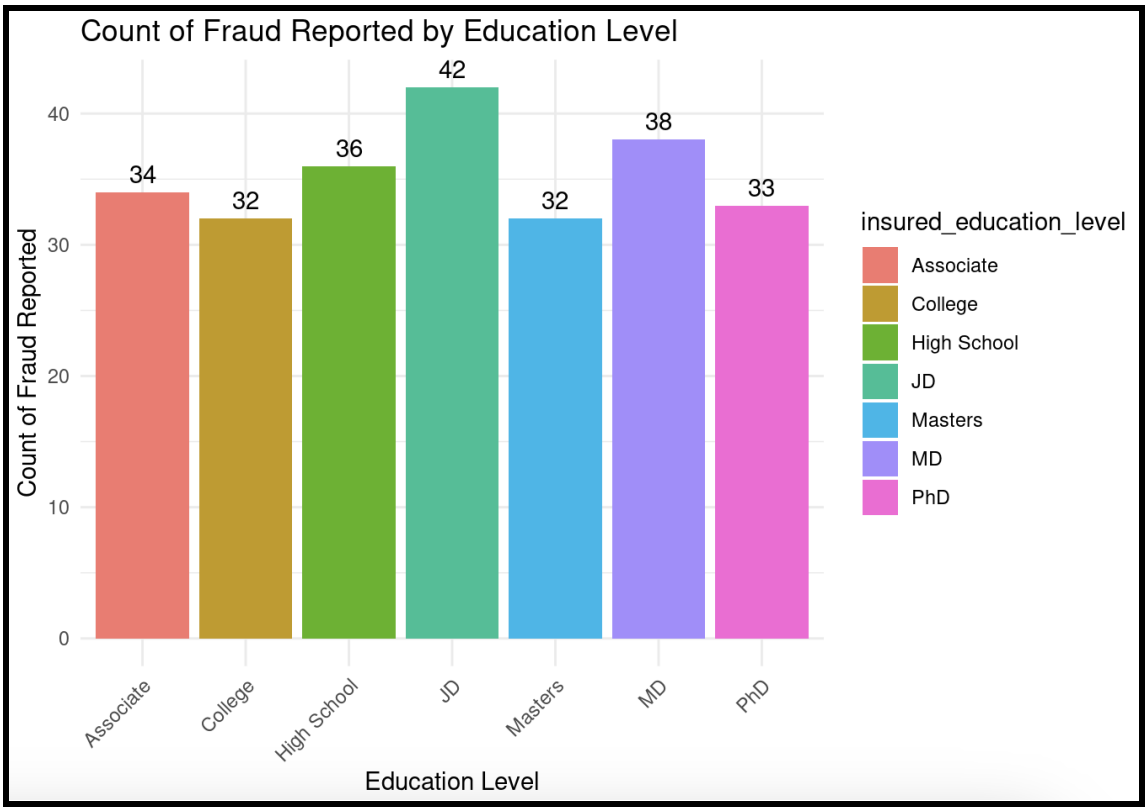


Visualization 1

As seen in the bar graph above, the most frequent reason for a claim complaint is for Denial of Claim. Claim handling delay is also a major reason why complaints are filed. Because of this, the hypothesis that claim denial is the top reason for insurance complaints is accepted. While some of these denials may be due to attempted fraud, another reason that claims are denied is the limitation of coverage. This occurs when individuals do not fully understand the extent of their coverage and may submit a claim that is not covered under their policy. Becuase of this, we cannot fully assume that all of the denials are due to fraud. However, we can extrapolate that denials of claims represent insurance companies attempting to regulate what claims are legitimate and are covered.

Hypothesis 2

Education level has a negative correlation with fraud, with a lower likelihood of fraud being reported by individuals with higher levels of education. Individuals with a higher level of education likely have a better understanding of the consequences of committing insurance fraud and, thus, will avoid activities that could elicit fraud. Because of this, insurance companies could use this as an indicator of who is more likely to commit fraud, which could improve their detection of it.

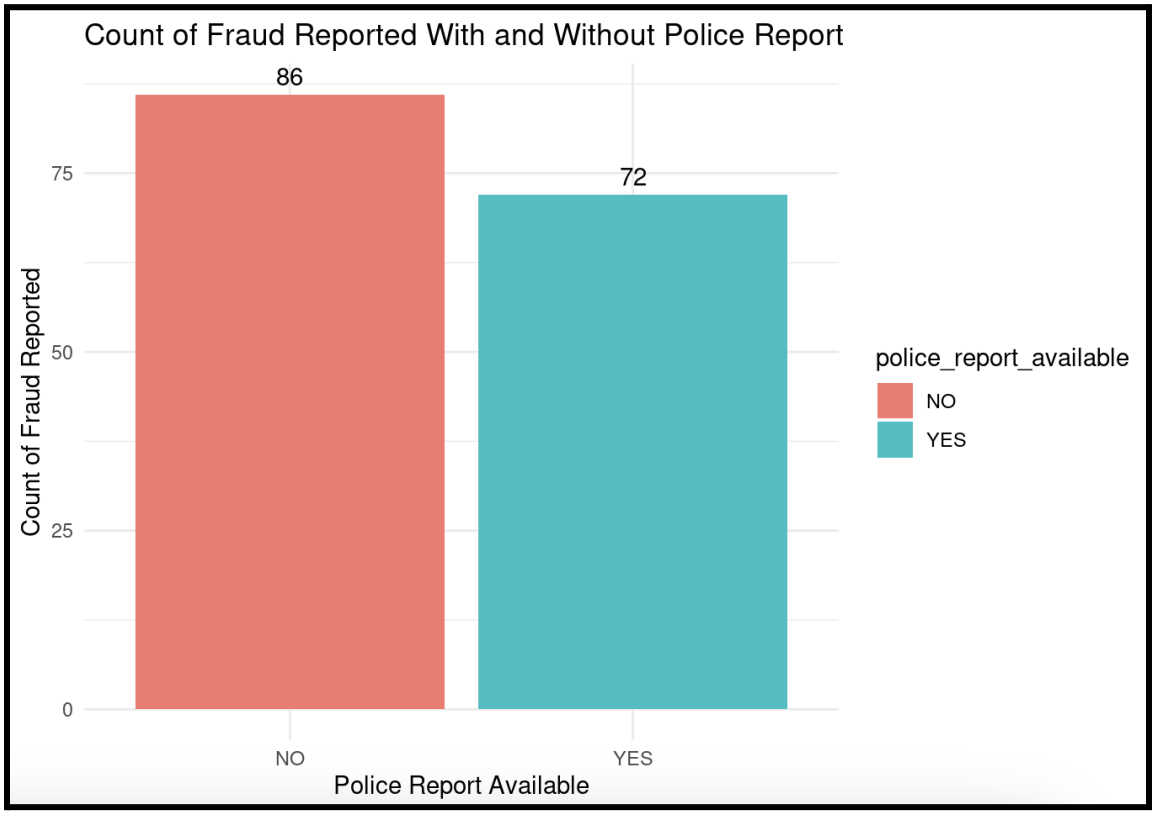


Visualization 2

The bar graph above visualizes all of the fraud cases reported and sorts them by the education level of the individual who filled out the fraudulent claim. As seen above, the frequency of fraud by education level is relatively equal across all education levels. This means that fraud is committed by individuals regardless of education level. In order to address fraudulent insurance claims, insurance claims should not place emphasis on the education level of an individual when looking for ways to identify potential fraudsters.

Hypothesis 3

The availability of a police report has a negative correlation with fraud reporting. In other words, the presence of a police report means that it is less likely that fraud will be reported. A police report provides official documentation of the incident and thus will provide less of an opportunity for the individual to commit fraud.



Visualization 3

The bar graph above visualizes all of the fraud cases reported and classified whether a police report was available with the original claim or not. As seen above, fraud cases that did not have a police report available were more frequent than those that did have a police report. From this, we can assume that fraud may be more difficult to commit when police are involved in the claim filing, as police reports are often submitted as documentation. Using this knowledge, insurance companies could use the availability of a police report as a potential indicator of whether fraud is likely or not in insurance claim cases.

Hypothesis 4

The likelihood of insurance claim fraud is correlated with cases that have higher financial stakes. This is because committing fraud is a very serious risk, so there must be a reasonable financial payout for fraudsters to make the risks worthwhile. A way to determine the financial stakes is by looking at the severity of damage, as more damage will result in a higher insurance payout.

**Visualizations in the Modeling Section*

Modeling

Initially, we decided to model the first dataset; however, the data was too large and was not good for modeling, so we decided to use dataset 2. First let's figure out which types of learning types we need to use for our hypotheses. For all of the hypotheses, the type of learning we will need to use is classification. Before starting any model, though, we had to start by encoding a great deal of variables so that the models ran smoothly.

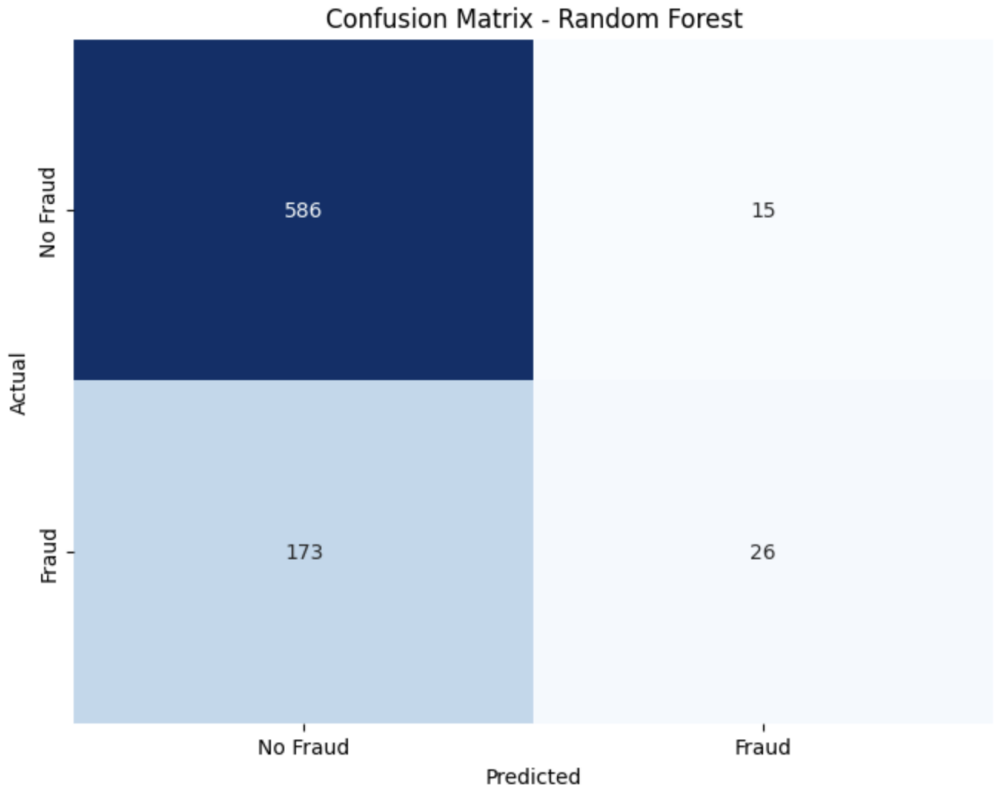
We encoded binary categorical variables by assigning a 0 one value and a 1 to another. We then set up our feature and target variables. The target variable was fraud_reported. For the feature variables, we had to drop a great deal of variables that were not important for our hypotheses, such as 'policy_number' and 'policy_bind_date.'. We then ran a random forest and SVM model. The accuracy scores were similar, but we decided to go with the random forest.

```
Random Forest Results:
Accuracy: 0.7650

SVM Results:
Accuracy: 0.7512
```

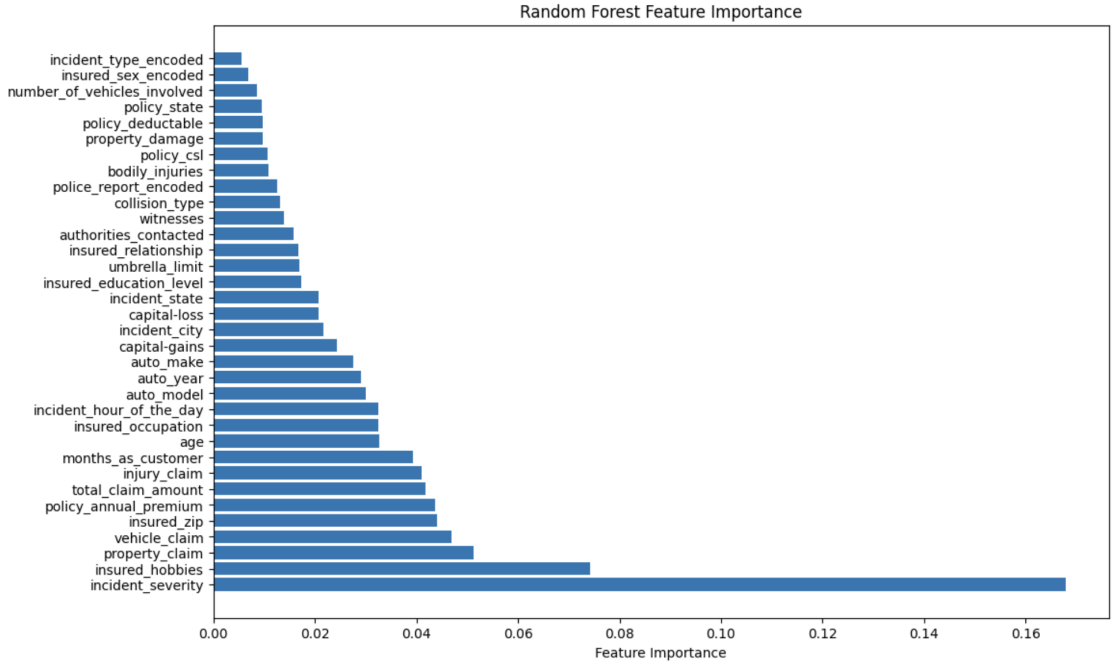
The accuracies for both are pretty low, so that is something important for stakeholders to understand before using this model.

Diving deeper, we decided to make a heatmap to visualize the confusion matrix. Analyzing this confusion matrix, I can see that there were a great deal of false negatives, which is understandable given the accuracy.



There is a very small amount of false positives because, looking at the feature importance, we noticed that incident_severity was the most important feature, which is important in hypothesis 4. It seems like there is a higher deal of fraud when the severity is reported, and hobbies are reported.

Our classification modeling pipeline involved splitting the insurance complaints dataset into an 80/20 train-test split to validate model performance. We tuned parameters through 5-fold cross-validation on the training data to prevent overfitting. Additionally, we incorporated upsampling methods in our algorithm to account for imbalanced categories of claims in the raw data. These approaches helped safeguard model generalizability by reducing variance and bias.


Random Forest Feature Importance
incident_type_encoded
insured_sex_encoded
number_of_vehicles_involved
policy_state
policy_deductable
property_damage
policy_csl
bodily_injuries
police_report_encoded
collision_type
witnesses
authorities_contacted
insured_relationship
umbrella_limit
insured_education_level
incident_state
capital-loss
incident_city
capital-gains
auto_make
auto_year
auto_model
incident_hour_of_the_day
insured_occupation
age
months_as_customer
injury_claim
total_claim_amount
policy_annual_premium
insured_zip
vehicle_claim
property_claim
insured_hobbies
incident_severity
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Feature Importance

Discussion

Discussing the Random Forest model which we chose in our modeling portion of the report, Random Forest excels in determining feature importance, making it a valuable tool in data analysis and machine learning applications. The algorithm assesses the significance of each feature by measuring the decrease in predictive accuracy when that particular feature is excluded from the model.

In terms of performance metrics, the model had an accuracy score of 0.76, showcasing a pretty low prediction performance. The precision and recall scores for Class 0 (No Fraud) are 0.77 and 0.98, respectively, indicating an above-average score toward predicting instances within this class. However, for Class 1 (Fraud), the model had a precision and recall score of 0.63 and 0.13. In summary, the model performs relatively well for Class 0 with good precision and recall, indicating a balanced prediction. However, for Class 1, the model's performance is less satisfactory, particularly in terms of recall, suggesting that it struggles to effectively identify instances of Class 1.

The limitations placed on our model by factors such as incomplete data, catastrophic events, underreporting, and the complexity of causation are profound and multifaceted. Incomplete data hampers the model's ability to provide accurate predictions or insights, as it relies on comprehensive information for robust analysis. Also, due to the incomplete data, our group needed to make assumptions as to what to omit, which may also hamper the model. Catastrophic events introduce unpredictability, disrupting the patterns on which models are trained and challenging their adaptability. Underreporting exacerbates these limitations by introducing biases and inaccuracies, distorting the model's perception of reality. Moreover, the intricate web of factors contributing to events, encapsulated in the complexity of causation, poses a significant challenge. Models struggle to capture the nuanced interactions between variables, leading to oversimplified representations that may not accurately reflect the true dynamics at play. As a result, these limitations call for continuous refinement and

adaptation of our modeling to better account for uncertainties, improve data integrity, and enhance their capacity to navigate the intricacies of real-world scenarios.

While our insurance fraud prediction model performed reasonably well in the initial testing phases, going forward, acquiring larger volumes of validated fraudulent claim data could help enhance model robustness through training on more representative samples. Furthermore, exploring additional engineered features derived from timestamps, claim change histories, and policyholder data could unveil new model improvement opportunities not currently captured strictly from the isolated claim snapshot data available.

The main obstacle we came across was data misinterpretation. At first, we faced the challenge of understanding the difference between insurance complaints and insurance claims. After learning the differences, we decided to include a specific dataset to analyze. Generally, the industry consists of a large amount of data, where studying all types of insurance was difficult. In the future, we would like to explore different types of insurance, such as Life, Health, etc.. Additionally, we aim to perform an analysis of the behavior of the customers and see if we can form new findings to help the insured party. As for our dataset, we aim to search for more accurate and specific data to fit our study.

Ethics

Social Need of This Research (Anticipate Purpose)

The implications of this project are socially significant, as the observed correlations between independent and dependent variables bear relevance to both the prosperity of insurance companies and the well-being of their customers. For insurance companies, the findings contribute crucial insights into factors influencing the frequency of insurance claims and the effectiveness of fraud detection mechanisms. For consumers, the project serves as an educational resource, enhancing understanding of the insurance system, fostering an environment open to improved integrity, and providing insights on minimizing both claims and fraudulent activities. This comprehensive approach aims to benefit both stakeholder parties by fostering a more informed and mutually beneficial insurance landscape.

Who Matters? Representativeness (Engage People)

The research was prioritized to address the interests of the two key stakeholders engaged in the study: insurance companies and their customers. The primary objective is to empower insurance companies to augment their customer base by finding a delicate balance between approving and denying claims, which would minimize complaints and enhance fraud detection mechanisms. Simultaneously, it provides a valuable avenue for customers to deepen their understanding of their insurance policies, which fosters an opportunity for increased integrity by minimizing fraudulent activities. Additionally, customers can identify the factors that contribute to the highest claim approval and denial rates and make informed decisions in order to avoid potential pitfalls.

Is the Research Controversial? (Reflect on Purpose)

In line with the previously outlined research goals and methodologies, our study was strictly objective-driven, aiming to enhance both customer satisfaction as well as the overall success of insurance companies. Thus, the nature of our research is inherently non-controversial. However, to safeguard against potential misinterpretations or the unintentional creation of controversy, seeking guidance from experts within the insurance field and gathering input from stakeholders will serve to minimize any likelihood of misunderstanding or misrepresentation of our findings. This collaborative approach ensures an overarching understanding of the research as well as transparent interactions and relationships with stakeholders in the industry.

Conclusion

This analysis of insurance complaints and claims data reveals meaningful insights into pain points in the claims process from the consumer perspective. We confirmed that claim denial is the top trigger for policyholder complaints, indicating that enhancing approval rates could alleviate this key issue. We also found mixed correlations between fraud likelihood and factors like education level and police report availability. However, we discovered a strong link between fraud risk and claim severity, suggesting financial incentives primarily drive this behavior.

Overall, this project successfully met its goal of deriving data-driven intelligence on dynamics influencing claims outcomes and customer satisfaction. Additional analysis should focus on the factors driving the high denial rates for certain claims. More examination of the language, documentation, and other evidence provided could reveal patterns linked to unsuccessful claims. This knowledge can better equip policyholders to avoid common pitfalls when filing. For insurers, an automated system that checks for these red flags could flag claims needing more information or intervention to bolster approval odds. This would serve policyholders by reducing frustration while saving insurers resources spent processing doomed claims.

Regarding fraud, further modeling work should center on quantifying the relationship between claim sizes, payouts, and fraud rates. This would empower more precision in fraud likelihood assessments. External demographic, geographic, and psychographic data could also boost fraud prediction accuracy. The next phases could also construct streamlined interfaces to quickly translate these risk scores into fraud investigation workflows. This would accelerate integrity oversight across expanding volumes of claims. Additional data gathering across more states and insurance providers could also bolster the insights and power of more granular segmentation analyses. By continuing to build on these initial findings, insurers can work towards improved customer experiences and integrity across their claims processes. Enhanced data utilization and collaboration will be key to unlocking further efficiency and oversight improvements in the insurance sector.

After analyzing our findings and researching the benefits of claim analysis, we believe that insurance claim data is typically quantifiable and structured, allowing us to do statistical analysis and modeling. By identifying patterns, trends, and correlations. we can use this data to help insurance companies assess and mitigate risks. Ultimately, these data can be used to develop a more accurate and efficient underwriting process where fewer resources are required. Furthermore, We can also utilize machine learning to understand how the claim is processed so that we can free up more resources and use those resources for complex claims.

Acknowledgment

Members	Contributions	Grade
Simaak Siddiqi	Presentation Delivery, Report Writing, Presentation Preparation, Regression and Classification	16% - 100
Ugochukwu Izuegbunam	Report Writing, Website, Presentation Preparation	14% - 87.5
Christopher Jones	Presentation Preparation, Report writing	11% - 68.75
Karim Ladak	Presentation Delivery, Report Writing, Presentation Preparation	14% - 87.5
Mohammad Morsy	Report Writing, Presentation Preparation	14% - 87.5
Hongzhi Luo	Presentation Delivery, report writing, Presentation Preparation	16% - 100
Stephen	Hypothesis development, presentation preparation, Graphs, report writing.	15% - 93.75

References

Clark, Maria. “30+ Mind-Boggling Health Insurance Claim Denial Statistics.” Etactics, Etactics | Revenue Cycle Software, 7 Mar. 2022, etactics.com/blog/health-insurance-claim-denial-statistics.

“Fraud Stats.” InsuranceFraud.Org, 20 June 2023, insurancefraud.org/fraud-stats/#:~:text=Insurance%20fraud%20steals%20at%20least,of%20property%2Dcasualty%20insurance%20losses.

“Insurance Fraud.” FBI, FBI, 17 Mar. 2010, www.fbi.gov/stats-services/publications/insurance-fraud.

"OpenAI. (2022). GPT-3.5, a language generation model developed by OpenAI."