Project Title: Insurance Fraud & Customer Insight Analysis

Name: John Oluwatimilehin Hicks

Company Name: Scib Insurance Brokers

Table Of Contents

Table Of Figures

Table of Tables

Executive summary

This undertaking will help to enhance the integrity of insurance brokerage. Data analytics will be used to identify crash fraudulent claims. As we assist the clients in locating the best insurance covers, it is also necessary to protect against the bad faith actors. Fraud does not only lead to financial loss but also becomes a deteriorating element to the system. With the help of analytical insight, the project can increase decision-making, operational efficiency, and credibility of the company.

It is based on three critical objectives and each of these objectives aim to identify which claim channels (online, broker, agent, or call center) are most likely to experience fraud we aim to get an idea about customer demographics that are more likely to execute fraud, or identify behavioral patterns that render to a belief of a fraud. These observations will inform sound communication practices, enhance customer targeting, and will streamline fraud detection practices.

Introduction

Although getting people to procure the most appropriate insurance policies and plans that their money can secure is our major preoccupation, we should not snore too much to the fact that not everyone operates in good faith. Cases of fraud and people trying to defraud the system are not something favorable to the industry. This is why we, as brokers, need to find repetitive patterns and certain other distinctive characteristics that would either empower and encourage us or weaken and undermine our bonds with the potential clients. The decision to focus and approach the right audience through data-based decisions not only secures the business goals of the company but also reenforces the position of the company as a trustworthy, reliable broker in the long run.

Business Objectives

Channel Performance Analysis

Of which channels of claim (e.g., online, agent-based, call center, or Broker), are fraudulent claims more likely to be made?

Purpose: *To give priorities to monitoring resources and simplify secure channels of communication.*

Client Demographics Insight

What is the preferred demographic profiling (e.g., age, gender, education, employment status, etc) most likely to commit fraud and most profitable to serve?

Purpose: *To find the most honest and useful target audience, which helps in segmentation of customers and performing the strategic acquisition.*

Fraud Detection Objective

Which are the most salient features, as well as behavioral indicators, that are most correlated with any fraudulent insurance claims? Which customer segments are most and least trustworthy

Purpose: *To create fraud risk profiles and also improve the claim vetting procedures to minimize the risk of financial losses*.

Data Description & Preprocessing

Overview of dataset structure

The dataset was gotten from Kaggle and it was chosen to show what I as a data professional can do with Scib dataset in the future should I be given the chance and opportunity to be employed and work on the company's dataset.

The dataset has 17998 rows and 25 columns. It has various features that would enhance our research below is the primary state of the data before there was any steps towards preprocessing

Table 1 Data description Values

Column Name	Data Type	Missing Values	Unique Values
claim_number	int64	0	17998
age_of_driver	int64	0	87
gender	object	0	2
marital_status	float64	5	2
safty_rating	int64	0	100
annual_income	int64	0	2693
high_education_ind	int64	0	2
address_change_ind	int64	0	2

living_status	object	0	2
zip_code	int64	0	276
claim_date	object	0	731
claim_day_of_week	object	0	7
accident_site	object	0	3
past_num_of_claims	int64	0	7
witness_present_ind	float64	132	2
liab_prct	int64	0	101
channel	object	0	3
policy_report_filed_ind	int64	0	2
claim_est_payout	float64	17	17981
age_of_vehicle	float64	8	17
vehicle_category	object	0	3

vehicle_price	float64	0	17998
vehicle_color	object	0	7
vehicle_weight	float64	0	17998
fraud	int64	0	2

Table 1 shows us the nomenclature of the dataset, we can see here that there is a fraud column where records of fraudulent or non fraudulent claim have been recorded through the years there are many other features we believe can help fulfill our business objective.

Below were the changes made to the primary state of the data:

- ➔ Claim_number and Zip_code columns were removed because we simply do not need them for our analysis.
- ➔ After seeing that the unique values in the date column was as high as 731 we decided to convert the column to years only as days and months were not useful information for our dataset.
- ➔ Categorical columns were changed to the no and yes to make the data more meaningful to analyze and visualize it mostly in cases when presenting a plot or report.

Exploratory Data Analysis

Here, we'd be taking a close look at the dataset and see how we can answer the business objectives with visualizations of specific features

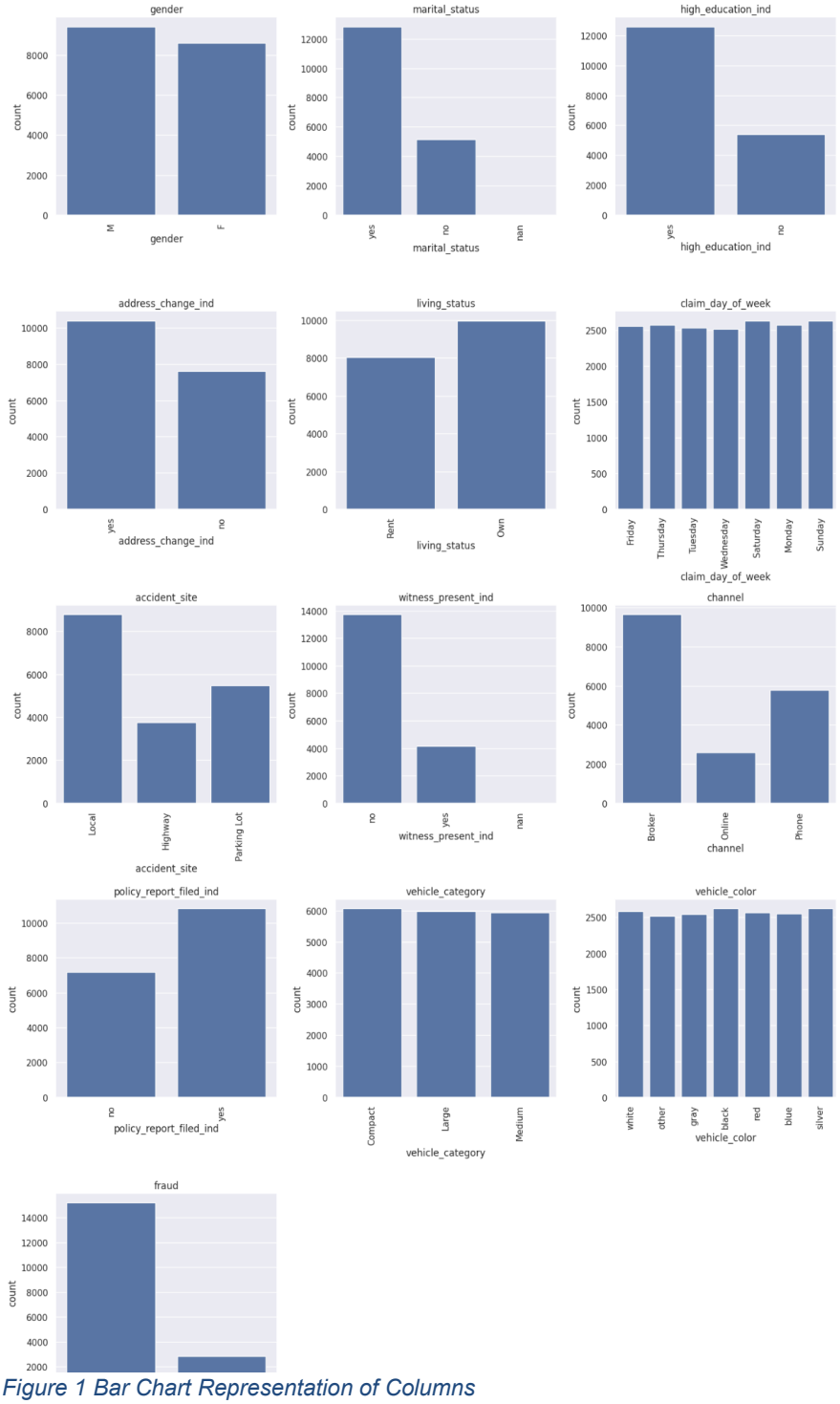

Figure 1 Bar Chart Representation of Columns

Key Insights From Our Dataset

- ➔ As seen in Figure 1 we can observe that there is a vast difference between non fraudulent and fraudulent claims with non fraudulent claims at 15,182 and fraudulent claims at 2,816 this goes to show that the threats are not so high but significant enough and deserves technical attention.
- ➔ With women being 9,414 in number and men at 8,584 the dataset is well balanced and has gender diversity.
- ➔ 12,824 people have a spouse and 5,169 people are single meanwhile we have 5 missing values in that column this shows that a lot of people that were interested in buying the car insurance are married people this makes sense if you consider the fact that individuals can now combine their finances and try to be more intentional after getting with a love interest.
- ➔ Also there's a huge difference between educated and uneducated individuals that bought the car insurance with educated at 12,584 and uneducated at 5,414 this goes to show that we have higher chances of success when pitching car insurance brokerage to an educated person.
- ➔ Most of the car insurance policy buyers fall in between 30,000 to about 45,000 annual income with the number of buyers (6,000 out of our dataset) at 40,000 annual income.
- ➔ 9,969 people own their apartments while 8,029 people have it rented.
- ➔ As for the claim channel distribution we observe that the broker channel is the highest with 9,633 then phone at 5,771, and online at 2,594.

In depth Feature Analysis Visualization with Box Plots

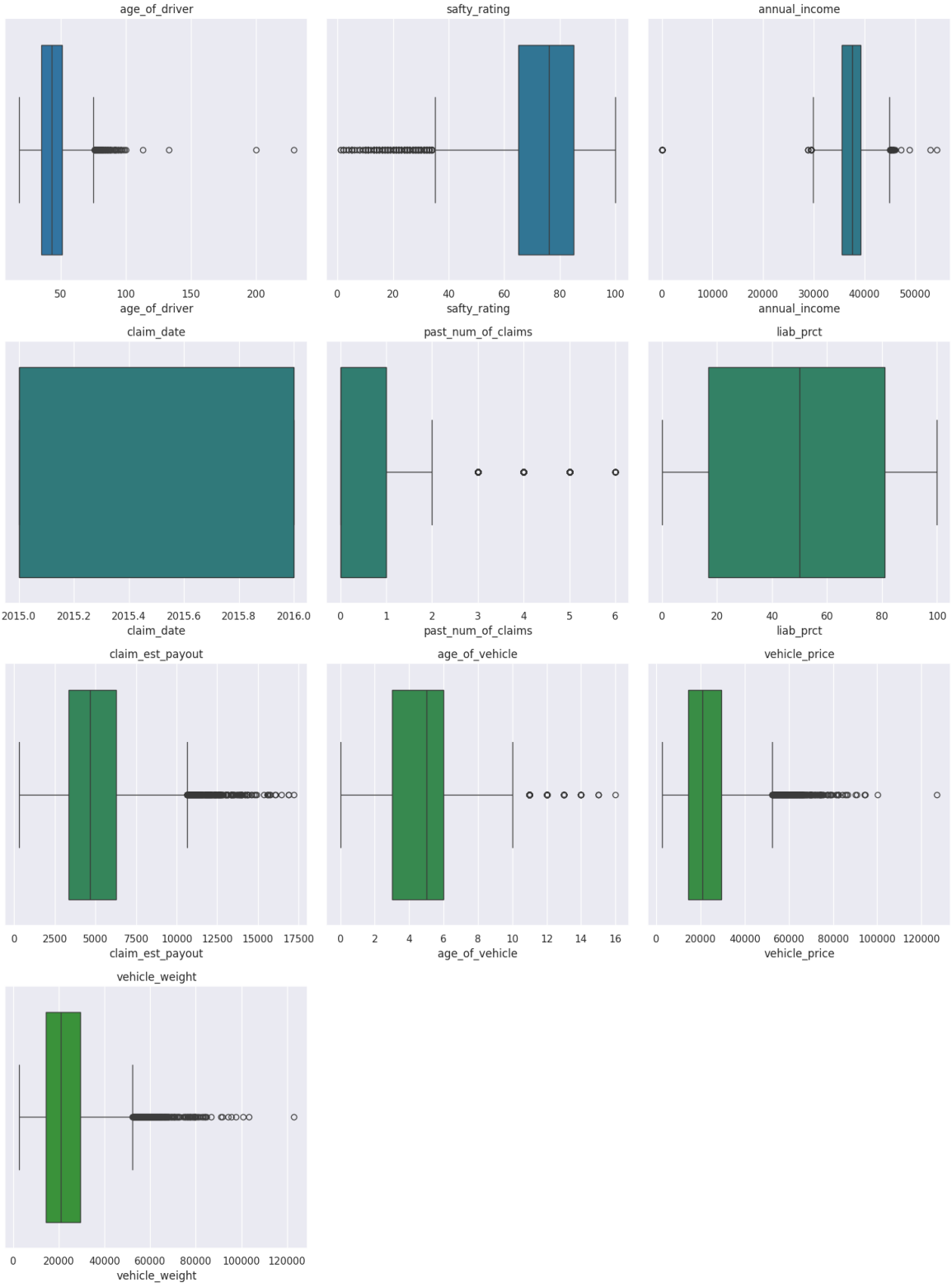

Figure 2 Visualizaation with Box Plot

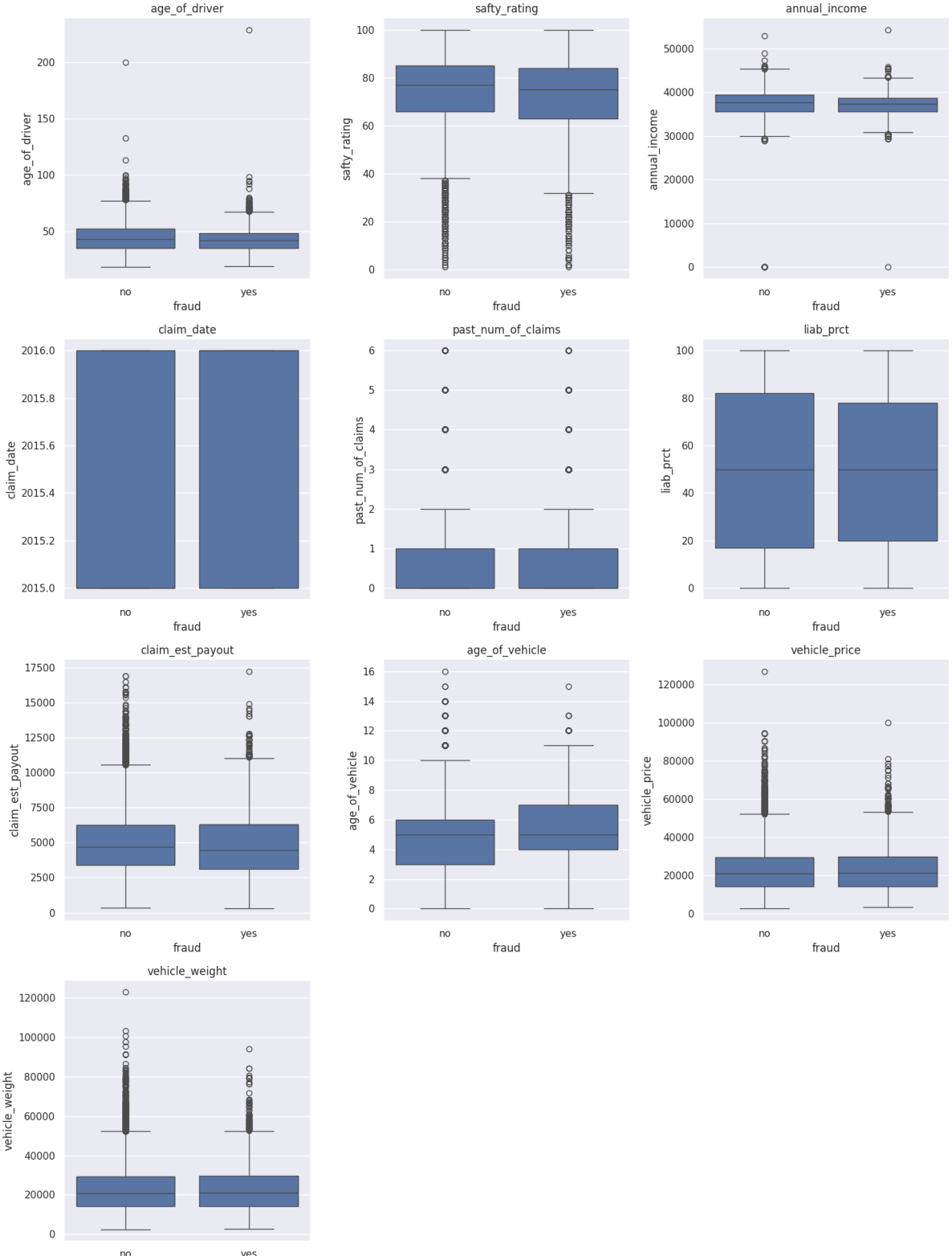

Figure 3 Visualization with Box Plot In Relation to Fraud

In depth Feature Analysis

In Relation to fraud Cases Our Box plots tell the following stories

- ➔ **Age of Driver**: The median and the interquartile range (IQR) of both fraud and non-fraud cases is very close in the distribution of drivers ages, with the non-fraud cases higher extreme found at higher ages. That implies that age might not be such a potent predictor of fraud in itself but extreme age might provide some scope of exploration.
- ➔ **Safety Rating:** Safety rating also has a slightly higher median value of fraud cases than that of non-fraud cases but the same IQR values overlap ranging between 0.200 and 0.500 and 0.100 to 0.600 respectively as shown in Figure 3. This doesn't provide any evidence that cars of lower safety may be slightly linked to fraud, but not on a scale that can become the single attribute.
- ➔ **Annual Income:** Fraud cases have lower median annual income when compared to non-fraud cases so with smaller IQR. It may mean that those drivers with the lower incomes tend to have higher chances in case of fraudulent claims, perhaps because of economic interests.
- ➔ **Claim Date:** The claim date distribution is broad and that of both groups are quite similar indicating that there are no changes in timelines in terms of fraud. This means that there is no seasonality of the fraud claims depending on the data presented.
- ➔ **Previous Claims:** Claims about the previous history of fraud demonstrate a higher median and bigger amount of outliers than non-fraud ones. This leads to the view that a history of many claims might be an indicator of possible fraud and as such, this is one of the main characteristics that should be tracked.
- ➔ **Liab_Pct:** The median of liability percentage (liab_pct) is similar in the two groups, although the IQR is slightly broader in the cases of fraud. This could mean that fraud claimants have a higher spectrum of liability distributions, which could be a sign of an undermined liability measure.
- ➔ **Claim Estimated Payout:** There is a high variability (wider IQR) of payout in fraud cases than non-fraud ones. This is an indication of the possibility of exaggerated payout claims as inflated claims that require close attention in confirmation of claims.
- ➔ **Age of vehicle:** There is similar median for the two groups of the age of vehicles, but the fraud cases have more outliers on older vehicles. It may happen that older cars are more vulnerable to fraudulent claims because repair services cost a lot or because of depreciation.
- ➔ **Fleet Price of vehicle:** The median vehicle price is higher and its range more variable in the fraud cases, suggesting that fraudulent claims may adopt the route of using expensive cars so that they can receive higher payments at the end of it all.
- ➔ **Vehicle Weight:** The distribution of the vehicle weight is virtually the same in both categories, with the fraction cases being higher in median and outliers. There is a possibility that the heavier vehicles might be linked to larger claims which may be used in fraud cases.

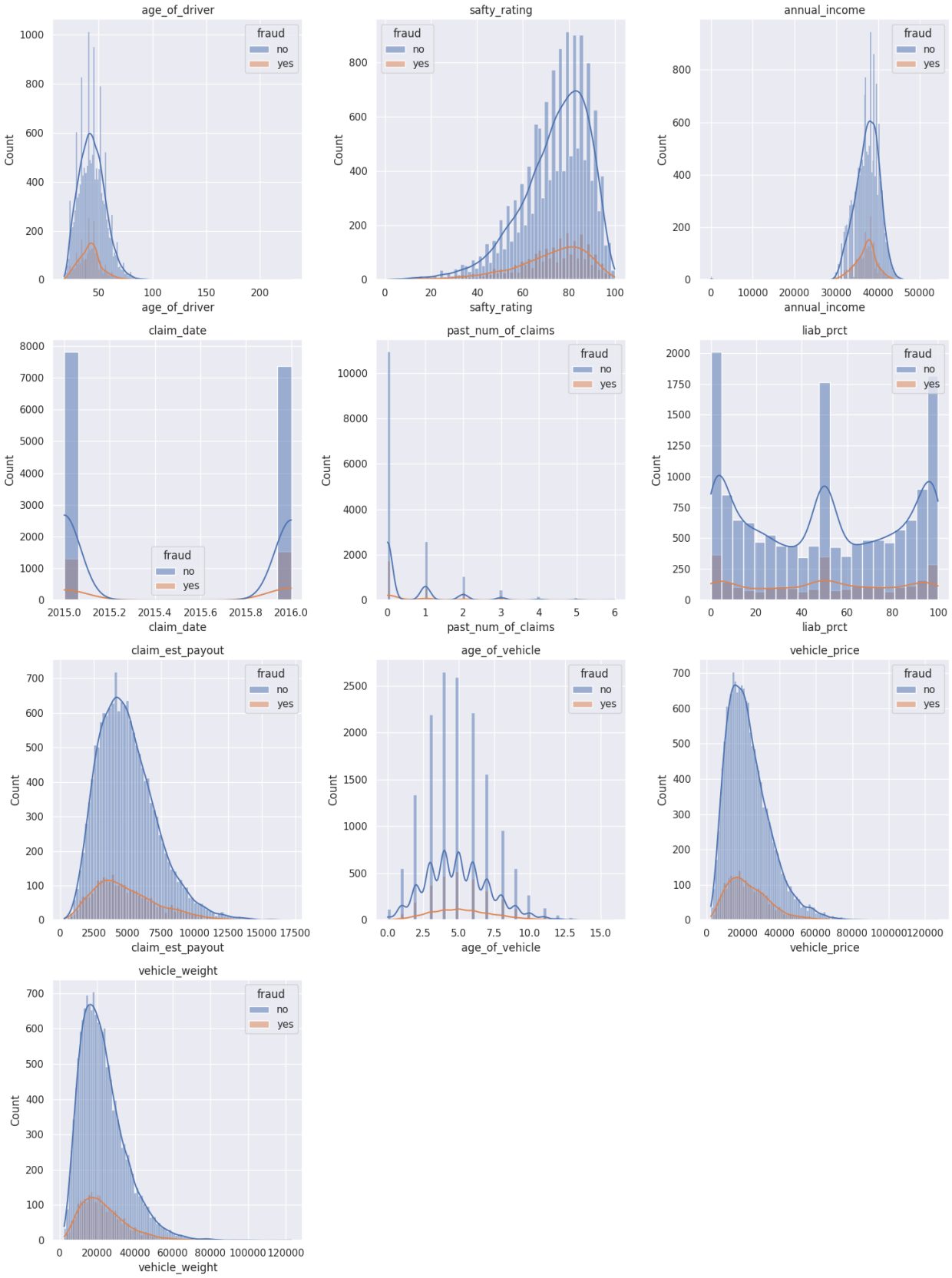

Figure 4 Histogram Representation of Features and Fraud Distribution

Narration Of Histograms Visualization

These histograms in Figure 4 describes exploratory characteristics that would be used in revealing possible fraud patterns in car insurance claims.

The drivers' age is rather similar in fraud and non-fraud groups, indicating a lack of significant signal in terms of the age and fraud; therefore, there is no strong age based fraud indicator.
The ratings of safety, however, indicate that cars with low-safety ratings standard (0-20) are more prevalent in fraud occurrences, and this offers a marginal risk indicator. In the same way, people who have a lower income per year seem to be marginally more addressed to fraud claims.

Volumes of days on which the claims are made are distributed across the period, which indicates no particular trend of fraud related to time. Few claims with a history of 12 have 12 past claims, and the majority of claimants' histories are zero, implying that there is a slight connection between the history of the claim and fraud. Greater variation in the percentages of liability can be seen among the cases of fraudulent cases which may indicate manipulation. In terms of claim payments, most of the values fall within 500-1000 range, but the fraud instances at times entail higher claims of payment.

As to transportation, there is also a minor increase in fraud association with older and heavier cars. Trends in the price of vehicles are also evident in that more fraud may be committed in cars of higher prices. All these observations indicate marginal but significant variables which the insurance brokers can keep a closer watch on, in making claim evaluations.

Extracted Insights

- ➔ **Use of Features:** Features such as safety rating, previous claim number, payout estimate on claim, vehicle price, and vehicle weight should be in the top priority because they will indicate minuscule variations between fraud and non-fraud cases to do predictive modeling.

- ➔ **Error Test Missing values:** Testing: Testing of the tails in the safety ratings, estimated payout, vehicle price and weight to fraud cases, as this may identify fraud.

- ➔ **Risk Profiling:** Profiling of drivers who have poor safety ratings, older vehicles, or vehicles with higher prices, because these are all elements that indicate some relationship to fraud.

- ➔ **Claim Risk:** Claims with large estimated payouts, as well as expensive/heavy vehicles, should be given increased scrutiny since they are slightly overrepresented in such cases.

- ➔ **Temporal Analysis:** Claim dates do not follow any particular trend and so, analyzing more detailed temporal information (such as month or day) may reveal certain trend in frauds.

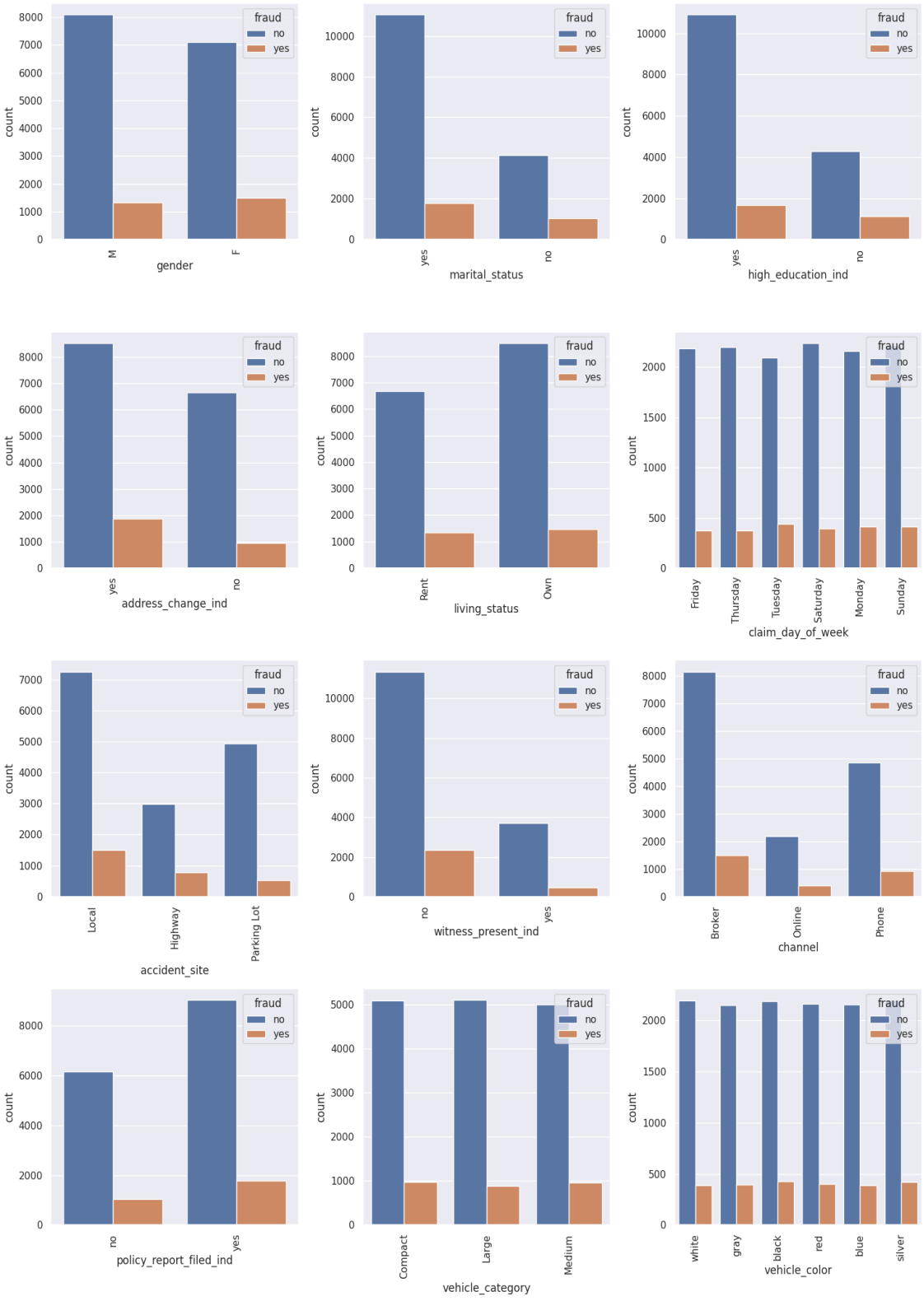

Figure 5 Box Plot Representation of Relational Differences between Fraud and other Features

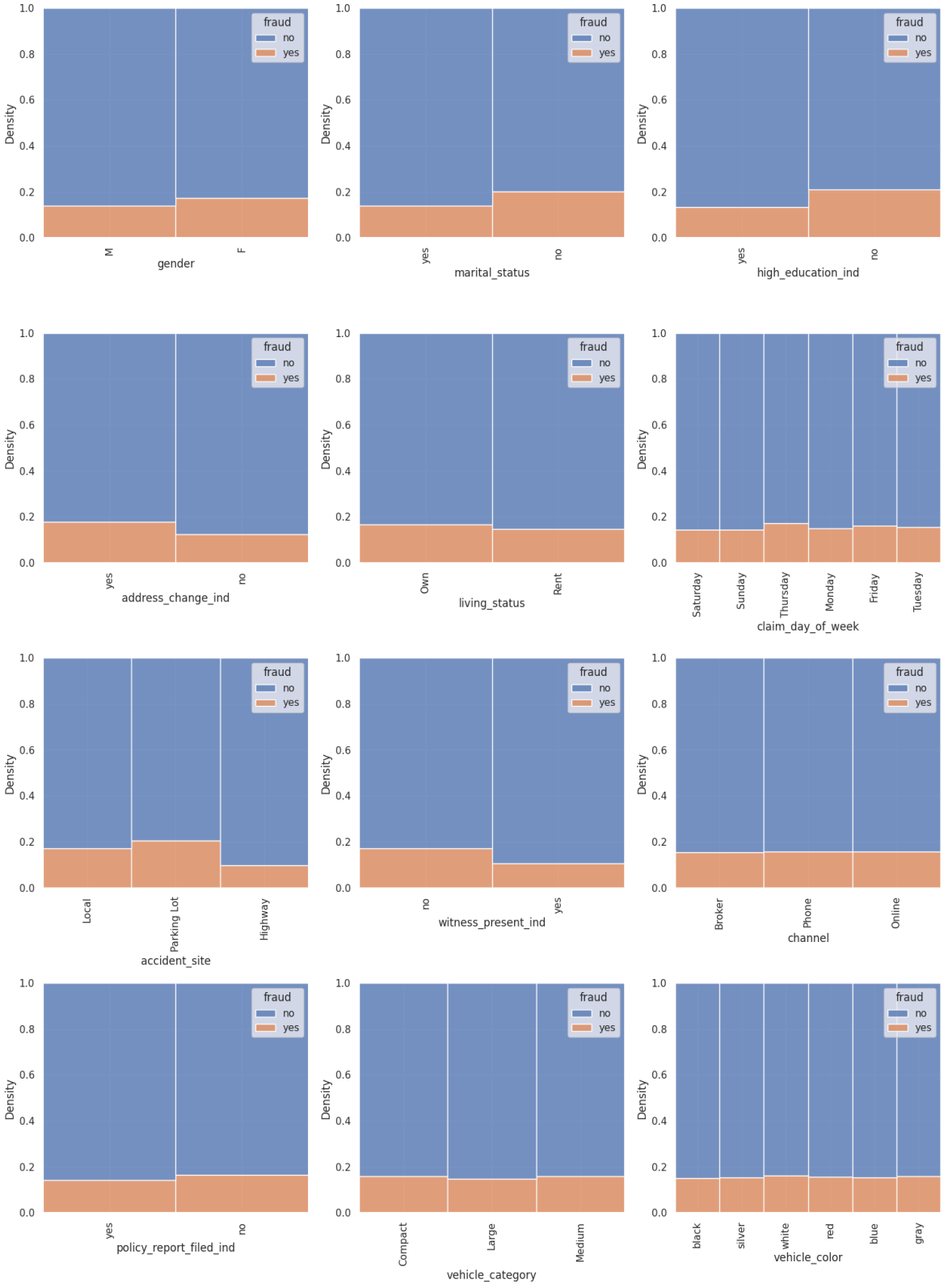

Figure 6 Desity Plots distribution of categorical variables for Fraud and Non Fraud cases

- ➔ **Address Change Indicator:** Fraud cases have a lower count and density for address changes ("yes") compared to non-fraud cases, suggesting stable addresses may be more linked to fraud.

- ➔ **Accident Site:** The fraud cases have lower numbers in cases of local and highway, and a little bit higher in parking lot, which suggests there are more parking lot accidents connected with the fraud cases.

- ➔ **Witness Present Indicator:** Fraud cases are far less in number and density when the witness is present ("yes"), and the fraud could thus be correlated with the absence of witnesses.

- ➔ **Policy Report Filed Indicator:** On fraud cases, numbers and density of reported policies are lesser compared to non-fraud cases, which suggests that the non-report of policies might be high in fraud cases.

Data Preprocessing:

Data Preprocessing is an iterative process, you really cant be done with preprocessing data once and for all so I decided to go back and drop the missing values in our data set as it would be appropriate to do so being that we still have enough authentic data to train our model with moving forward.

Also the columns that were converted to Yes / No for easy visualization were converted back to 1s and 0s for label encoding.

Correlation Analysis

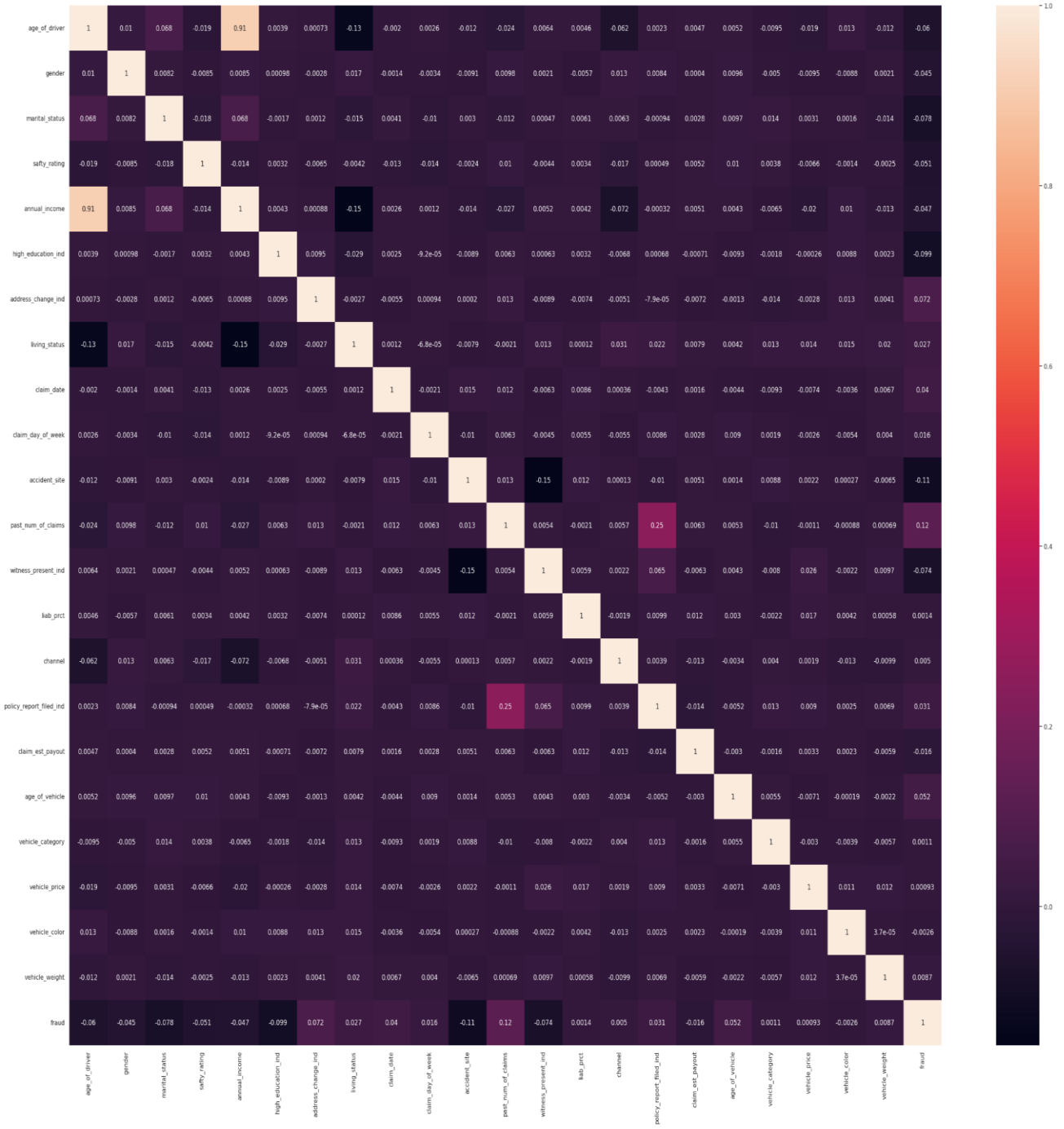

Figure 7 Correlation Analysis

Heatmap Correlations of Car Insurance Fraud Claims

The Correlation heatmap in Figure 7 depicts the interrelationships among the different features on the car insurance fraud claims data set by the ranges of color gradation where the dark purple color is characterized to have low/negatively low correlation with each other and the light yellow color with a high/positively high correlation with each other. Notably, one can single out such strong positive correlations as claim_date and claim_dow with around 0.6 and vehicle_price and vehicle_weight with about 0.74, which signify that there are predictable relationships in the timing of claims and vehicle pricing. The correlations between many pairs of variables, in particular categorical variables (e.g. the variables gender and vehicle_color) are very low (approaching 0) and the corresponding correlations reflect little linear dependence.

Some of the central and most critical issues on correlation with fraud are as given below.

Negative Correlation with Fraud:

Witness Present Indicator (witness_ present_ind): With a normal correlation of -0.41 with the variable what you should know is that claims with witnesses are least likely to be incurred as fraud and as such, absence of witnesses is a key risk factor.

Policy Report Filed Indicator (policy_report_filed_ind): To a greater extent it holds a negative correlation with the variable of interest, i.e. fraud (approximately -0.47), indicating that failure to report policy is more likely to have a fraudulent claim which is a crucial point of fraudulent claims to be avoided.

Positive Correlation with Fraud: There are not many strong positive associations found with fraud most of the positive correlations are weak (e.g. "accident_site" or "claim_day_of_week"), which indicates that there are minor contextual effects.

Columns Relevant to fraud
Column policy_report_filed_ind (approximately -0.47) has the highest (negative) correlation with the fraud column and thus it is the most important indicator. This implies that one of the possible measures to minimize cases of fraud could be by making sure that policies are well reported.

Business Implication
The concentration on the claims with unreported policies and absent witnesses can increase the possibility of detecting fraud, and the results lack of strong positive correlations prove that the analysis of fraud patterns should be performed utilizing multivariate models.

Predictive Analysis

After the data set was split in two for train and test where train is 80% and test is 20% Decision Tree classifier was employed to build a model.
The accuracy score was at 63.87
F-1 Score : 0.6387331838565022
Precision Score : 0.6387331838565022
Recall Score : 0.6387331838565022
Jaccard Score : 0.46921968293185096
Log Loss : 13.021375902066145

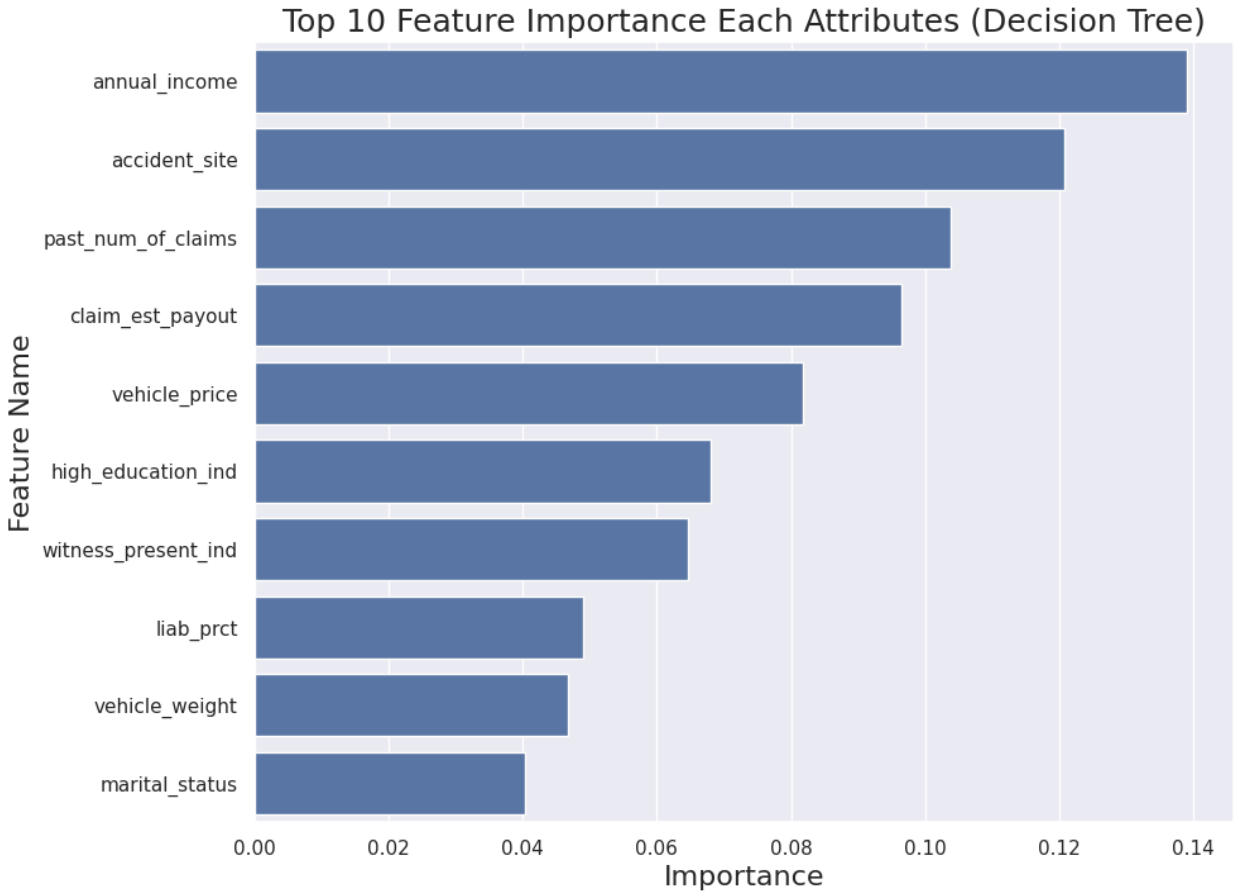

Figure 8 Feature Importance Ranking

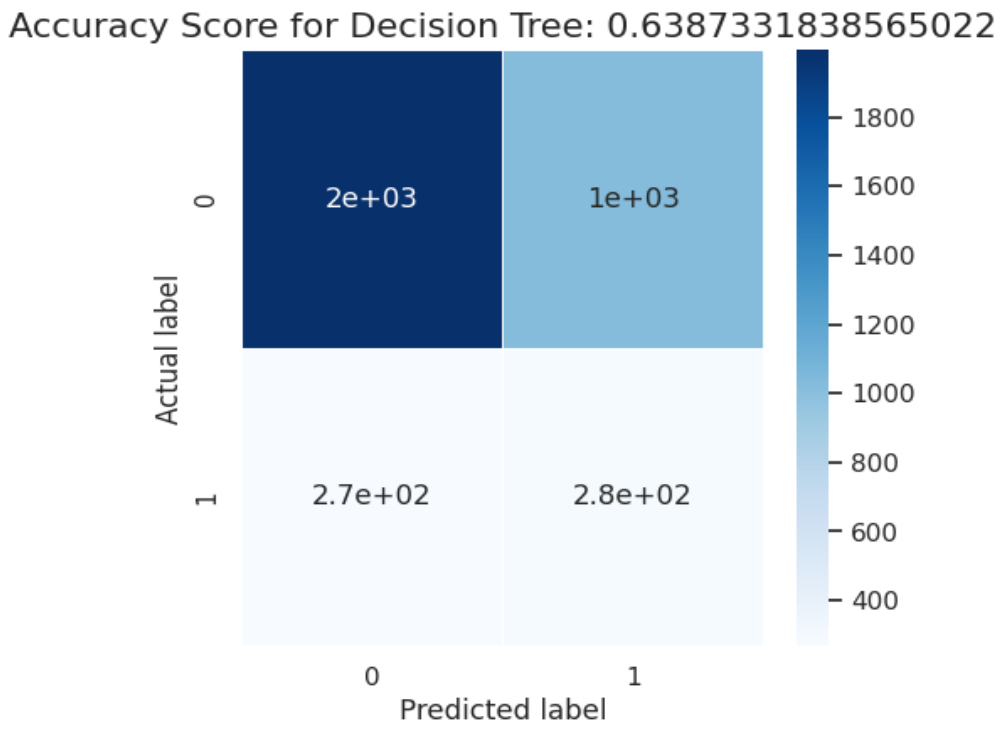

Figure 9 Confusion Matrix

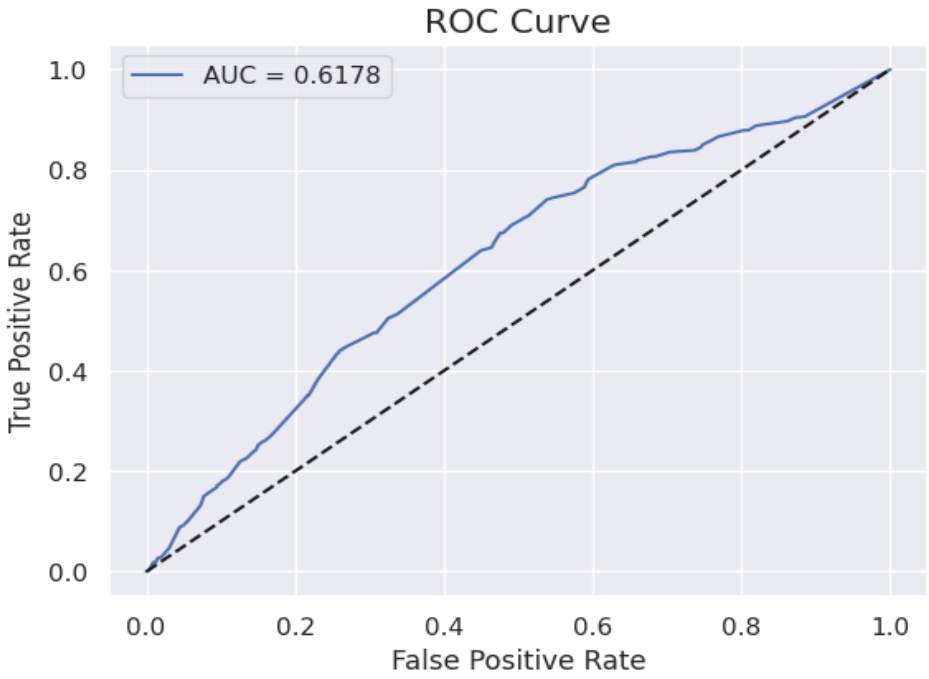

Figure 10 Reciever Operating Characteristic Curve

Dissatisfied with this evaluation metrics, another classifier was employed which was Random Forest classifier

The random forest classifier performed better with These Metrics:

Accuracy Score : 84.53 %
F-1 Score : 0.8452914798206278
Precision Score : 0.8452914798206278
Recall Score : 0.8452914798206278
Jaccard Score : 0.7320388349514563
Log Loss : 5.576260277688529

Insights From Predictive Analysis

Evaluation measures which are obtained using the Random Forest Classifier on insurance fraud detection dataset represent that the model performed very well in various aspects. The Accuracy Score of 84.53 would indicate that the model was able to classify correctly the cases of fraud/non-fraud in more than 84 out of all instances as an indication of good overall predictive ability.

The F1 Score with values at 0.8453 that balances between precision and recall further confirms the fact that the model is working well in addressing the issue of class imbalance which is critical in fraud detection problems where both false negatives as well as false positives have business impacts. Interestingly, they also match with the Precision and Recall scores having a value of 0.8453, meaning that the model is as good at avoiding false positive (incorrectly identifying genuine claims as a fraud) as it is at avoiding false negative (failing to identify a fraud case).

The Jaccard Score of 0.732 indicates that there is a high value overlap of the predicted labels with the actual ones which speaks of the robustness of the model. Nonetheless, the Log Loss value of 5.576 implies that the predictions are correct but the confidence levels of the probabilistic outputs are somewhat scattered indicating that there might have been some miscalibrated predictions.

On the whole, the combination of these metrics suggests a Random Forest model that performs rather well, and hence would be useful in initiatives of detecting fraudulent activities within the context of insurance brokerage business, and help us pick the right clients.

Conclusion

This discussion reveals how data analytics can be used in the reinforcement of an insurance brokerage company activities such Scib. We have been able to detect actionable insights by analyzing the demographics of car insurance policy buyers by age group and how fraudulent the clients can be through claim channel, type of vehicle being amended, and financial metrics. Our predictive models in particular the Random Forest classifier is promising in the level of accuracy and precision and this proves that data-driven fraud detection is not only possible but in fact necessary In Insurance. The application of the mentioned insights can facilitate the streamlining of operations, reduce potential financial losses, and cement the superiority of the firm as a sensible intelligent broker in the insurance industry.