Thi Ngoc Minh Nguyen

INsurance Detection project

Table of Contents

[1.1 Context 3](#_Toc207021656)

[2. Exploratory data analysis 3](#_Toc207021657)

[2.1 Data Transformation 3](#_Toc207021658)

[2.2 Genuine Claims Analysis: 3](#_Toc207021659)

[2.3 Incident Analytics: 4](#_Toc207021660)

[2.4 Discussion of the Results and Conclusions: 5](#_Toc207021661)

[3.1 Modelling 6](#_Toc207021662)

[3.2 Model Comparison 11](#_Toc207021663)

[3.3 Generate Business Insights and Strategize 12](#_Toc207021664)

[4. References 13](#_Toc207021665)

# **1.1 Context**

The report presents a key information on analyse patterns and indicators that may signify fraudulent activity based on historical motor insurance claims in the Australian context. The first part involves Exploratory Data Analysis using Power BI to visually identify patterns and key factors of fraudulent claims. The second part uses RapidMiner to build and compare supervised learning models to predict potential fraud.

**Findings**

From two dashboard, around 50% of claims are fraudulent and mostly occurring in urban areas with speeding as the leading cause. “Nose-to-tail” collisions under basic policies show high fraud rates. Fraudulent claims often involve SUVs valued $2,000–$2,900. Some claims have filing dates before accidents, with half confirmed fraudulent. These patterns are essential as a beginning step for company to understand about the fraud associated factors

Two model are Random Forest and Logistic regression both highlighted key a fraud indicator, of “Physical Therapist” as the top predictor. Logistic regression identified Saab vehicles, Queensland location while Random Forest identify Van and speeding as statistically significant variables. Although both models showed similar precision, logistic regression had better recall and identified more true fraud cases while missing fewer. Due to its lower false-negative rate, logistic regression is more suitable for early fraud detection, where missing fraud is costlier than false positives.

**Recommendation**

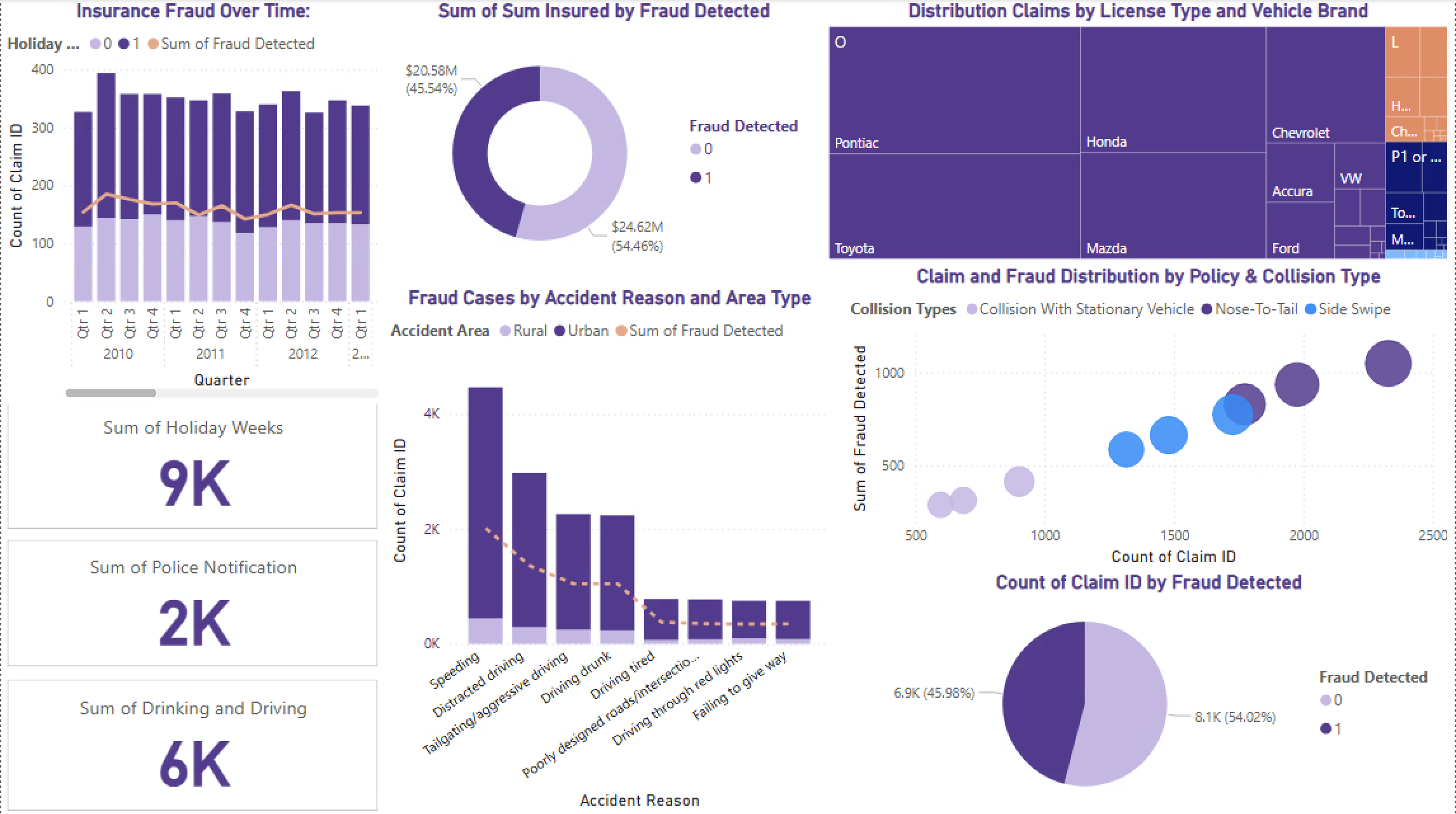
After successfully identify high risk factors, business can implement and improve the fraud control framework that align with these factors based on Three Lines model. This ensures the company identify the risk early, mitigate and early detect fraud claims.

# **2. Exploratory data analysis**

* 1. **Data Transformation**

Before conducted visualising data, I have done several data transformation to ensure the best data quality. First, I converted all occupation entries were to lowercase for uniformity, and missing or ambiguous values were addressed by replacing "None" in the past number of claims with 0 and converting NA values to "Unknown". The holiday indicator was standardised by encoding it as 0 for "No" and 1 for "Yes". Also, monetary values were reformatted into decimal places to facilitate accurate analysis. I changed monetary values into decimal form and renamed the “make” column to “VehicleBrand” to improve clarity. Also, I have detected that, there are several claim dates occurred before the accident date, potentially flagging suspicious cases. I used 0 for “No” and 1 for “Yes” for this anomaly column. Lastly, all null values and redundant entries across columns were removed to maintain a clean dataset for analysis.

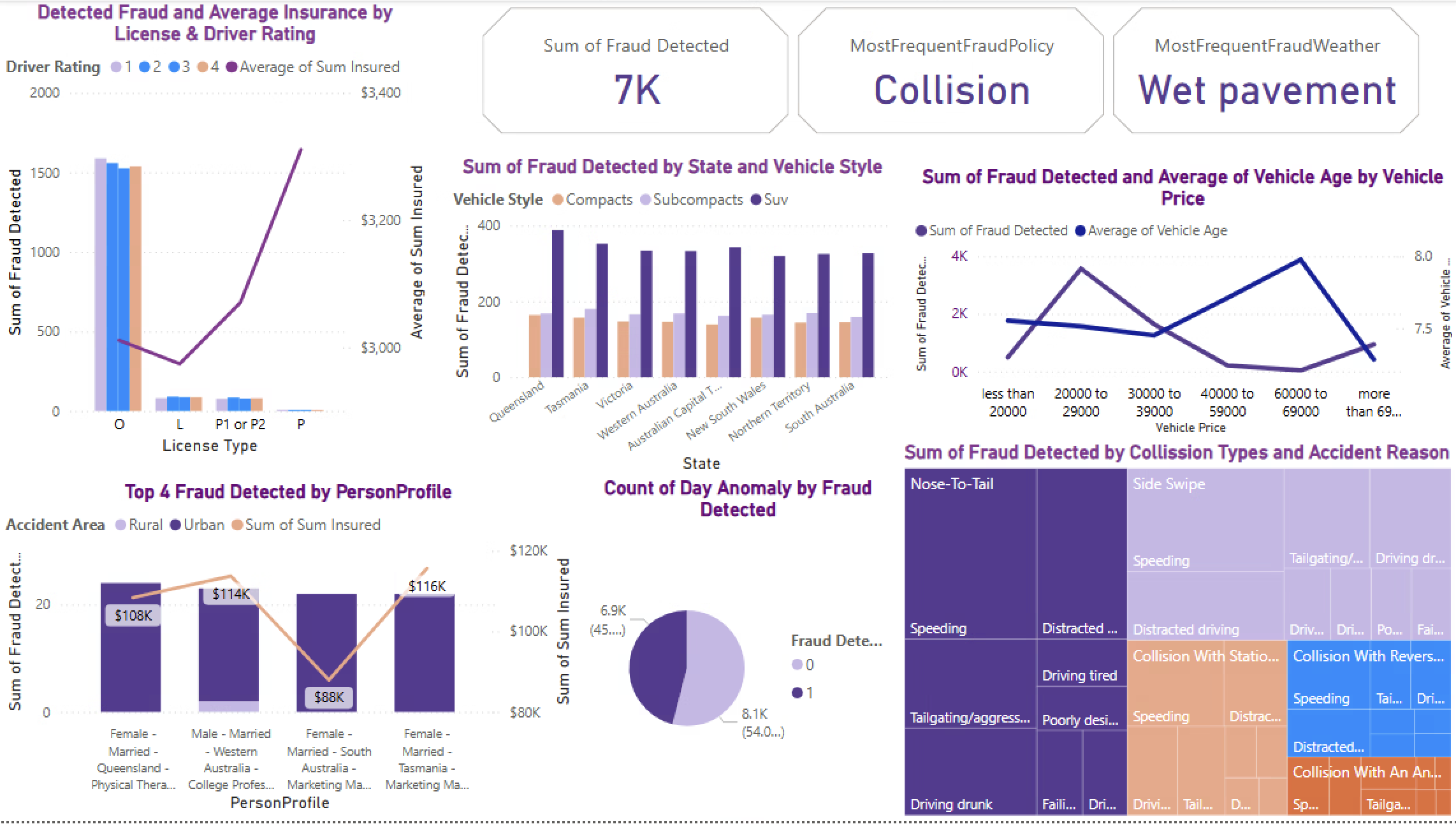
**2.2 Genuine Claims Analysis:**



As shown in the screenshot, fraud cases account for nearly half of the dataset, indicating that multiple factors may be contributing to this high proportion. Consequently, approximately 50% of the total insured claims are fraudulent. Most fraudulent cases occur in urban areas, with speeding being the most frequently reported cause of accidents.

Interestingly, only 2,000 cases were reported to the police. Most accidents tend to occur during holiday weeks. In terms of collision type, "nose-to-tail" accidents is a common type under basic collision policies are associated with the highest number of fraud cases. Lastly, many claims are using Pontiac brand, with the majority drivers holding license type 'O'.

**2.3 Incident Analytics:**



The screenshot highlights key factors that influence and help flag fraudulent claims. Interestingly. Most fraud instances involve 'O' licenses with lower insured amounts, whereas 'P' license holders have larger average sums insured. This suggests that fraud can still occur even with lower-value claims.

Fraudsters typically use vehicles between $2,000 and $2,900, often SUVs with an average age of 7.8 years. Notably, claimants with certain demographic profiles consistently show a higher frequency of fraudulent activity and are responsible for significant monetary losses. Additionally, some claims were filed with dates that happen before accident date, and roughly half of these cases were confirmed as fraudulent.

# **2.4 Discussion of the Results and Conclusions:**

These claims involve SUVs valued between $2,000–$2,900 with an average vehicle age of 7.8 years. These mid-range vehicles appear to be favoured in staged claims. Speeding and nose-to-tail collisions dominate fraud cases, especially during holiday weeks. A significant number of claims are submitted before the accident date, with nearly half confirmed as fraudulent. Therefore, personal profiles, day anomaly, vehicle types and accident reasons are the key features to determine fraudulent claim.

When executives across organisation have a clear view the dynamics of vehicle insurance. Executives can be able to navigate volatility and uncertainty more efficiently to build an identification capability for emerging risk (Michel-Kerjan & Serino, 2024). This could include a flexible tech infrastructure to collect, aggerate and monitor the risk in real time and link it to risk appetite dashboard. Other suggestions could be made are utilise AI and Machine Learning to investigate potential fraud claims. Deloitte predict that implementing AI technology around claims life cycle with real time analyses from server models could save around US$80 billions by 2023 (Kamalapurkar et al., 2025). Lastly, strength internal control and optimise business process though automation, internal control structure. An effective system will safeguard organisation assisted facilitate reporting and ensure legal compliance (Crime and Corruption Commission, 2018)

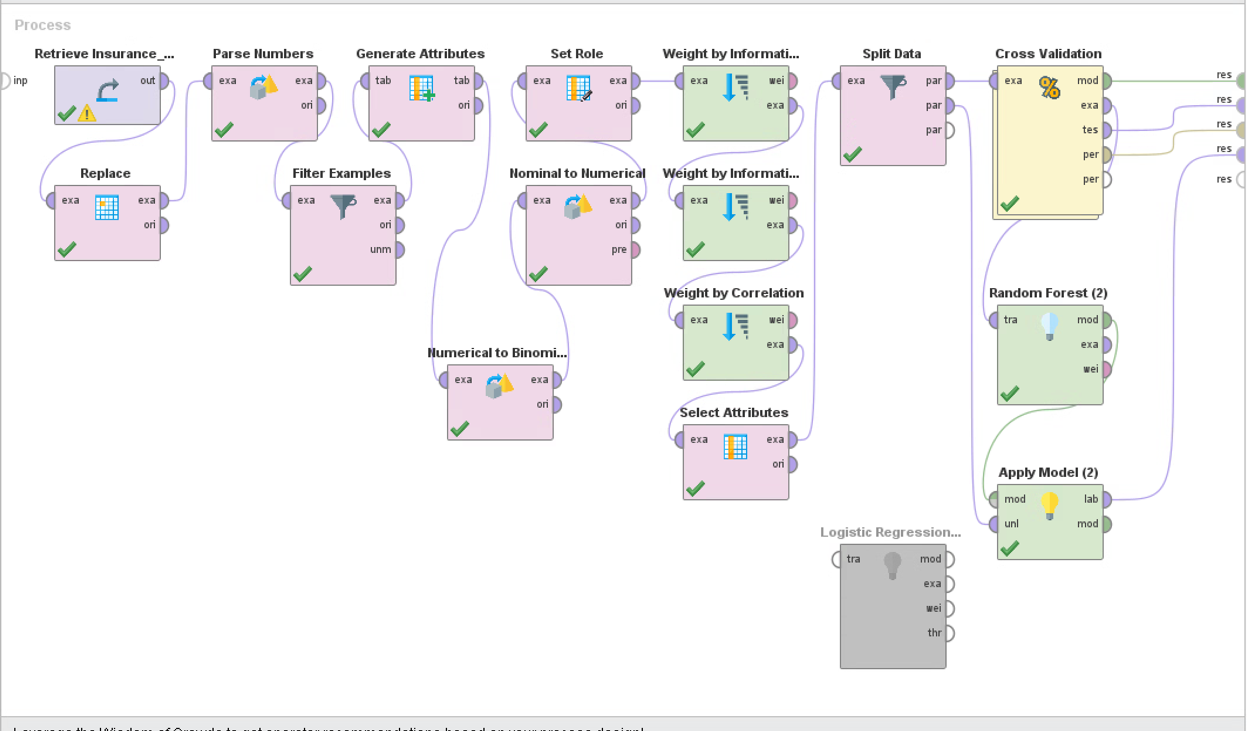
**3.1 Modelling**

Diagram

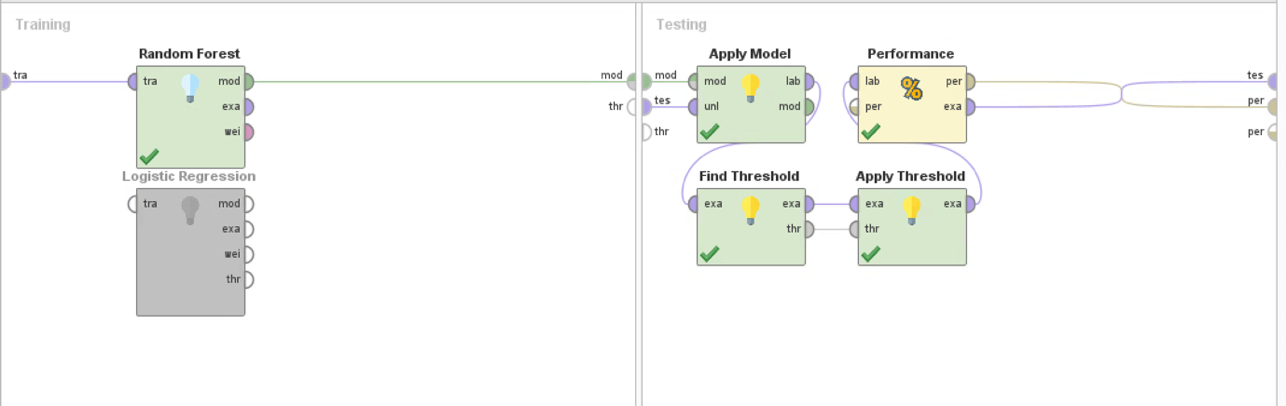
Description automatically generated

After accessing the process, check whether the process is properly set up and make corrections if necessary. Please document all changes and specifications below. Note: You can choose the models learned from this course or choose any other model outside the tutorial exercise.

1. **Deliverable: Screenshot of Analytical Process**



1. **Deliverable: Screenshot of Analytical Process:**



1. **Deliverable: Screenshot of Results of the first model**

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a white sheet

Description automatically generated

1. **Deliverable**

Occupation = physical therapists > 0.500

| Vehicle style = Van > 0.500

| | Accident Reason = Speeding > 0.500: true {false=0, true=1}

| | Accident Reason = Speeding ≤ 0.500

| | | License Type = L: true {false=0, true=1}

| | | License Type = O

| | | | date\_diff = false

| | | | | BasePolicy = Collision > 0.500: true {false=0, true=2}

| | | | | BasePolicy = Collision ≤ 0.500

| | | | | | Marital Status = Married > 0.500: false {false=3, true=0}

| | | | | | Marital Status = Married ≤ 0.500

| | | | | | | PastNumberOfClaims = 1: true {false=0, true=1}

| | | | | | | PastNumberOfClaims = 2 to 4: false {false=1, true=0}

| | | | date\_diff = true: false {false=6, true=0}

| Vehicle style = Van ≤ 0.500

| | Vehicle style = Executive > 0.500

| | | Claim Date > 19/04/2011

| | | | Drinking-and-Driving = false: true {false=0, true=3}

| | | | Drinking-and-Driving = true

| | | | | Accident Reason = Speeding > 0.500: true {false=0, true=1}

| | | | | Accident Reason = Speeding ≤ 0.500: false {false=1, true=0}

| | | Claim Date ≤ 19/04/2011: false {false=1, true=0}

| | Vehicle style = Executive ≤ 0.500

| | | Vehicle style = SUV > 0.500

| | | | Claim Date > 06/07/2020: true {false=0, true=7}

| | | | Claim Date ≤ 06/07/2020

| | | | | VehiclePrice = 30000 to 39000 > 0.500

| | | | | | AddressChange\_Claim = 2 to 3 years: true {false=0, true=3}

| | | | | | AddressChange\_Claim = no change

| | | | | | | Repair amount > 489

| | | | | | | | Collission Types = collision with an animal: true {false=1, true=3}

| | | | | | | | Collission Types = collision with reversing: false {false=5, true=0}

| | | | | | | | Collission Types = collision with stationary vehicle: false {false=11, true=2}

| | | | | | | | Collission Types = nose-to-tail: false {false=14, true=11}

| | | | | | | | Collission Types = side swipe: false {false=12, true=2}

| | | | | | | Repair amount ≤ 489: true {false=0, true=2}

| | | | | VehiclePrice = 30000 to 39000 ≤ 0.500

| | | | | | Accident Date > 12/04/2010

| | | | | | | Accident Reason = Driving tired > 0.500: true {false=0, true=2}

| | | | | | | Accident Reason = Driving tired ≤ 0.500

| | | | | | | | Days\_Policy\_Claim = 15 to 30: true {false=0, true=1}

| | | | | | | | Days\_Policy\_Claim = more than 30: false {false=79, true=75}

| | | | | | Accident Date ≤ 12/04/2010: true {false=0, true=7}

| | | Vehicle style = SUV ≤ 0.500

| | | | Gender = false

| | | | | Sum Insured > 6129: true {false=0, true=1}

| | | | | Sum Insured ≤ 6129

| | | | | | Sum Insured > 6099.300: false {false=2, true=0}

| | | | | | Sum Insured ≤ 6099.300

| | | | | | | Repair amount > 356

| | | | | | | | Accident Reason = Driving tired > 0.500: true {false=2, true=7}

| | | | | | | | Accident Reason = Driving tired ≤ 0.500: true {false=66, true=67}

| | | | | | | Repair amount ≤ 356: true {false=0, true=1}

| | | | Gender = true

| | | | | Sum Insured > 5841.450: true {false=0, true=4}

| | | | | Sum Insured ≤ 5841.450

| | | | | | AddressChange\_Claim = 1 year: true {false=0, true=2}

| | | | | | AddressChange\_Claim = 2 to 3 years: true {false=0, true=1}

| | | | | | AddressChange\_Claim = 4 to 8 years: false {false=2, true=0}

| | | | | | AddressChange\_Claim = no change

| | | | | | | Sum Insured > 5700.150: false {false=1, true=0}

| | | | | | | Sum Insured ≤ 5700.150

| | | | | | | | Vehicle style = Midsize > 0.500: false {false=12, true=8}

| | | | | | | | Vehicle style = Midsize ≤ 0.500: true {false=37, true=62}

Occupation = physical therapists ≤ 0.500

| Accident Date > 29/12/2020: true {false=0, true=4}

| Accident Date ≤ 29/12/2020

| | Accident Date > 05/01/2010

| | | Repair amount > 200.500

| | | | Occupation = college professor > 0.500

| | | | | Repair amount > 245.500

| | | | | | AddressChange\_Claim = 1 year

| | | | | | | Police Notification = false

| | | | | | | | DriverRating > 2: true {false=0, true=7}

| | | | | | | | DriverRating ≤ 2: false {false=2, true=0}

| | | | | | | Police Notification = true: true {false=0, true=2}

| | | | | | AddressChange\_Claim = 2 to 3 years

| | | | | | | License Type = L: false {false=1, true=0}

| | | | | | | License Type = O

| | | | | | | | DriverRating > 1.500: true {false=1, true=7}

| | | | | | | | DriverRating ≤ 1.500: false {false=2, true=0}

| | | | | | AddressChange\_Claim = 4 to 8 years

| | | | | | | VehiclePrice = more than 69000 > 0.500: false {false=2, true=0}

| | | | | | | VehiclePrice = more than 69000 ≤ 0.500

| | | | | | | | VehiclePrice = 30000 to 39000 > 0.500: false {false=5, true=0}

| | | | | | | | VehiclePrice = 30000 to 39000 ≤ 0.500: true {false=7, true=14}

| | | | | | AddressChange\_Claim = no change

| | | | | | | Days\_Policy\_Claim = 15 to 30: false {false=1, true=0}

| | | | | | | Days\_Policy\_Claim = more than 30

| | | | | | | | Accident Reason = Driving drunk > 0.500: true {false=34, true=35}

| | | | | | | | Accident Reason = Driving drunk ≤ 0.500: false {false=296, true=211}

| | | | | Repair amount ≤ 245.500: true {false=0, true=3}

| | | | Occupation = college professor ≤ 0.500

| | | | | Accident Date > 13/01/2010

| | | | | | BasePolicy = Collision > 0.500

| | | | | | | Occupation = dentists > 0.500

| | | | | | | | Sum Insured > 6189.300: false {false=2, true=0}

| | | | | | | | Sum Insured ≤ 6189.300: true {false=106, true=120}

| | | | | | | Occupation = dentists ≤ 0.500

| | | | | | | | Occupation = marketing manager > 0.500: false {false=115, true=109}

| | | | | | | | Occupation = marketing manager ≤ 0.500: false {false=1949, true=1547}

| | | | | | BasePolicy = Collision ≤ 0.500

| | | | | | | Sum Insured > 6648.200: false {false=1, true=0}

| | | | | | | Sum Insured ≤ 6648.200

| | | | | | | | Occupation = Accountants > 0.500: true {false=157, true=182}

| | | | | | | | Occupation = Accountants ≤ 0.500: false {false=3003, true=2804}

| | | | | Accident Date ≤ 13/01/2010

| | | | | | VehiclePrice = 20000 to 29000 > 0.500

| | | | | | | Repair amount > 4840.500

| | | | | | | | Holiday Weeks = No: false {false=2, true=0}

| | | | | | | | Holiday Weeks = Yes: true {false=0, true=1}

| | | | | | | Repair amount ≤ 4840.500: false {false=3, true=0}

| | | | | | VehiclePrice = 20000 to 29000 ≤ 0.500: true {false=0, true=11}

| | | Repair amount ≤ 200.500: true {false=0, true=2}

| | Accident Date ≤ 05/01/2010

| | | Repair amount > 4606: true {false=0, true=3}

| | | Repair amount ≤ 4606: false {false=13, true=0}

1. **Deliverable: Interpretation of the results of models**

Explain the choice and impact of the most decisive predictors in the above models **(max. 100 word**s in total)

Random forest tree makes it easier to evaluate variable importance, or contribution, to the model (IBM,2021). This helps insurers in detecting suspicious claims. According to the result, the first split is physical therapist means that this occupation is more likely associated with suspicious claim patterns.

Logistic regression is a supervised machine learning that predict shows that variables such as **Saab car, Queensland and Physical Therapist** have **low p-values**. This means they are statistically significant and strong predictors of fraud. Additionally, variables like **sum insured, accident date**, and **claim date** show **lowest standard error**, indicating precise and influential estimates in identifying fraudulent activity.

**3.2 Model Comparison**.

* + 1. Deliverable: Screenshots of Results of the two models that you chose.

Logistic regression model

A screenshot of a graph

Description automatically generated

Random forest

A screenshot of a graph

Description automatically generated

* + 1. Deliverable: Interpretation and justification of model performances

Describe the model's performance and which model is better for this analytical task. Justify your preference. You may also refer to the prediction and classification, model selections, and other related academic articles, blogs, and other publications to justify the model selections by following the academic style of writing and referencing (**max. 150 words in total**).

The logistic regression confusion matrix reveals a weighted mean recall of 52.07% and it outperformed **Random Forest** with 51.87%. The model correctly identifies 58.27% of fraud case and only 45.84% of true negatives which are not fraudulent. This logistic model shows a stronger ability to identify positive cases compared to Random Forest of 56.42%. Precision is similar at both models, with 56.33% for false and 47.80% for true predictions, indicating a high false positive rate. Logistic regression misses about 41.73% of actual fraud cases, while Random Forest misses slightly more around 43.58%.

As a result, both model shows the similar precision-recall, however, logistic are more suitable for the analytics task because it missed less actual fraud case. In fraud detection, a false negative error is usually more costly than a false positive error (Phua et al., 2010). Therefore, a logistic regression for fraud organisation in identify high-risk claims early and improve fraud detection efficiency

**3.3 Generate Business Insights and Strategize**

The result above has revealed several of rich information on which factors affect and may be potential fraudulent claim such as occupation, vehicle style and state. Random Forest revealed variable importance hierarchies, helping insurers prioritize investigation on high-impact features. While precision remains modest, false negatives are costlier (Phua et al., 2010), making **logistic regression preferable** for early identification. A fraud control framework is needed to monitor the high-risk pattern to guide claim assessment. Based on the Three-line model, in the first line monitors claims and flags risks factors then creates policies and manages detection technologies, and the third line makes sure that independent audits and verifications of the efficacy of fraud controls are conducted (Services Australia, 2025).

In this digital area, digital information can significantly reduce the risk of fraudulent claims occurring access and analyse data in real time (TAPsDIGITal, 2025). Company can consider are designing and distilling an enterprise technology and operational strategy by defining guiding principles and designing a lean and agile operating model. For example, DMAIC methodology to identity level of fraud, analyse and improve from it (Six Sigma, 2020). For deeper insights, additional data such as **claimant history, IP/geolocation data,** and **external databases can help model to improve its accuracy.**

The application of predictive analytics in insurance fraud detection presents important ethical challenges. Especially, when insurer adopts an algorithm in support of fraud detection, there is a risk that they engage in illegitimate profiling (Timmermans et al., 2021). A responsible and moral use of data by establishing a data ethics framework that incorporates impartial audits and fairness checks will improve the ethical problem for insurance company . (EBDF, 2023).

Also, predictive models like logistic regression and random forest are trained on historical data, which may contain chronological biases. This occurs when events and trends that will result in significant differences between the two timeframes are not taken into consideration by models (IMS,nd). For instance, changes in fraud tactics, claim reporting behaviour, or socio-economic conditions may lead to significantly different patterns over time. the models risk making inaccurate predictions, leading to both revenue loss and unjust claim denials. Insurance company requires continuous monitoring of predictive models for bias and fairness across difference demographic groups as needed (Trigyn, 2024).

When leveraging personal data in automated decision-making, transparency is essential. Consumers anticipate their claims to be transparent and unambiguous. Based on SamSung IDC report, 51% of customers report frustration over unclear information in regarding terms and claim conditions. This, in turn shows that increase in trust from customer can helps organisation trust the use of predictive model. Businesses must embrace open communication and transparent information sharing on data use and privacy procedures for all processes to implement predictive regression (Santenac et al., 2023).

# **4. References**

Please include all the references (**minimum 10 references (5 per each task**) that you have cited in your report.

1. Kamalapurkar, K., Sharma, N., & Canaan, M. (2025, April 24). Property and casualty carriers can win the fight against insurance fraud. Deloitte Insights; Deloitte. https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2025/ai-to-fight-insurance-fraud.html
2. Michel-Kerjan , E., & Serino, L. (2024, July 1). Navigating shifting risks in the insurance industry | McKinsey. Www.mckinsey.com. https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/navigating-shifting-risks-in-the-insurance-industry
3. Crime and Corruption Commission. (2018). Fraud and Corruption Control Best Practice Guide. <https://www.ccc.qld.gov.au/sites/default/files/2019-08/Fraud-and-Corruption-Control-Best-Practice-Guide-2018.pdf>
4. Six Sigma. (2020, November 30). COVID-19 Situation: Six Sigma Ongoing Training Announcements. SixSigma.us. https://www.6sigma.us/insurance/using-six-sigma-to-prevent-insurance-fraud/
5. TAPsDIGITal. (2025). LinkedIn. Linkedin.com. https://www.linkedin.com/pulse/reasons-insurers-struggling-prevent-fraud-tapsdigital/
6. Services Australia (2025, January). Fraud and Corruption Control Plan 2025-2026. https://www.servicesaustralia.gov.au/sites/default/files/2025-03/fraud-and-corruption-control-plan-2025-26.pdf
7. IBM. (2021, October 20). Random Forest. Ibm.com; IBM. https://www.ibm.com/think/topics/random-forest
8. Timmermans, M., van Praat , F., & Ruepert , A. (2021, August). The growing relevance of data ethics in insurance. Compact. https://www.compact.nl/articles/the-growing-relevance-of-data-ethics-in-insurance/
9. Enterprise Big Data Framework. (2023, April 27). Why Data Ethics Matter | Establishing a Data Ethics Framework. Enterprise Big Data Framework©. https://www.bigdataframework.org/knowledge/why-data-ethics-matters-establishing-a-framework-for-responsible-data-practices/
10. Navarro, C. L. A., Damen, J. A. A., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M., & Hooft, L. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ, 375, n2281. <https://doi.org/10.1136/bmj.n2281>
11. Intergrity Service Management (nd). Statistical Bias in Health Care Fraud Detection and Prevention Systems. https://integritym.com/wp-content/uploads/2017/08/Statistical-Bias-in-Health-Care-Fraud-Detection-and-Prevention-Systems.pdf
12. Santenac, I., Majkowski, E., Sun-Young, A., & Vermeulen|author, P. (2023, November 29). How increased trust and transparency can unlock growth. Www.ey.com. https://www.ey.com/en\_gl/insights/insurance/global-insurance-industry-trends
13. Avila, B. (2024, September 24). Why insurance transparency matters and 3 tips to improve it - Samsung Business Insights. Samsung Business Insights. https://insights.samsung.com/2024/09/24/why-insurance-transparency-matters-and-3-tips-to-improve-it/
14. Trigyn. (2024, March 6). Managing Bias in Predictive Analytics. Www.trigyn.com. https://www.trigyn.com/insights/minimizing-bias-predictive-analytics