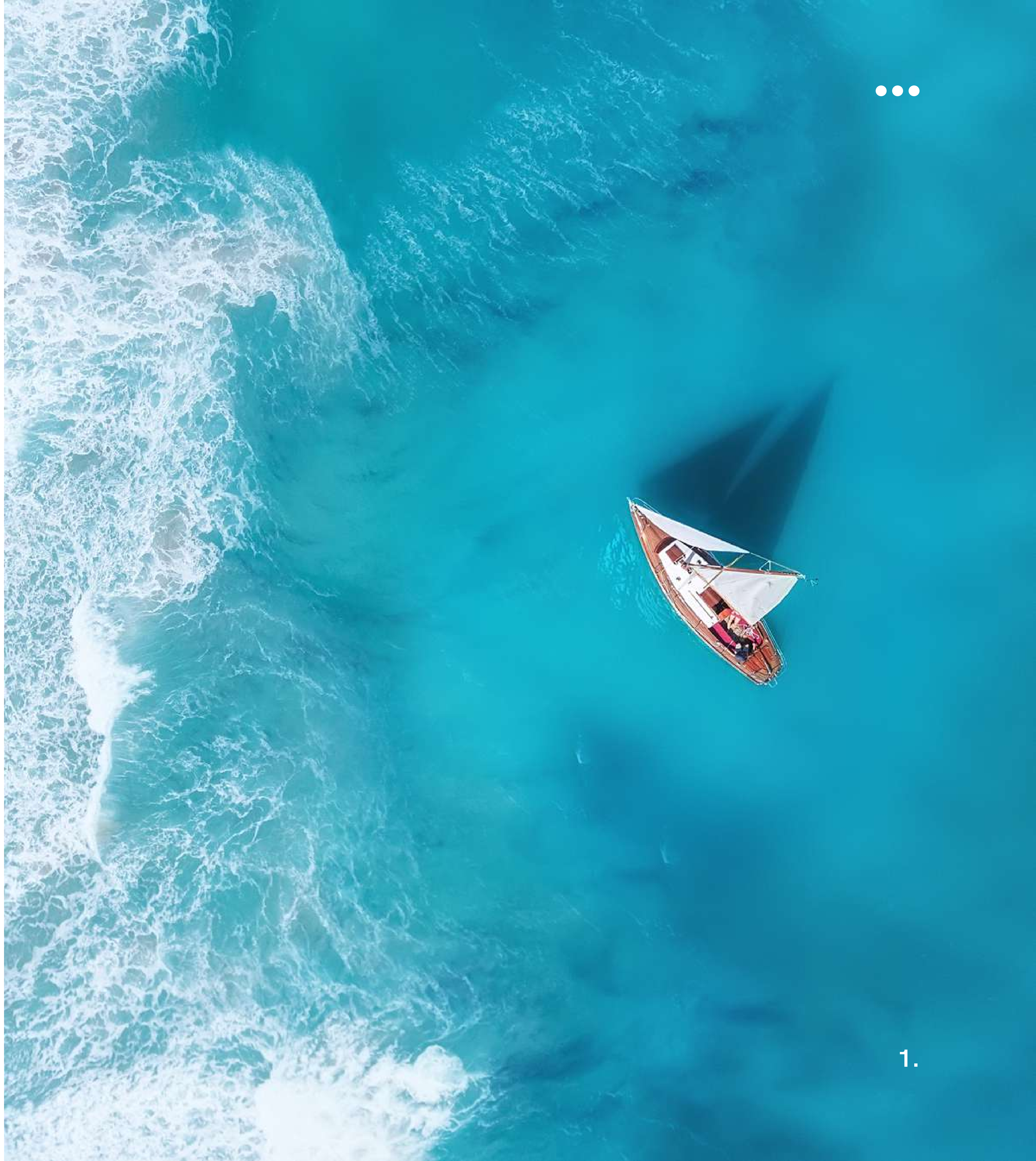
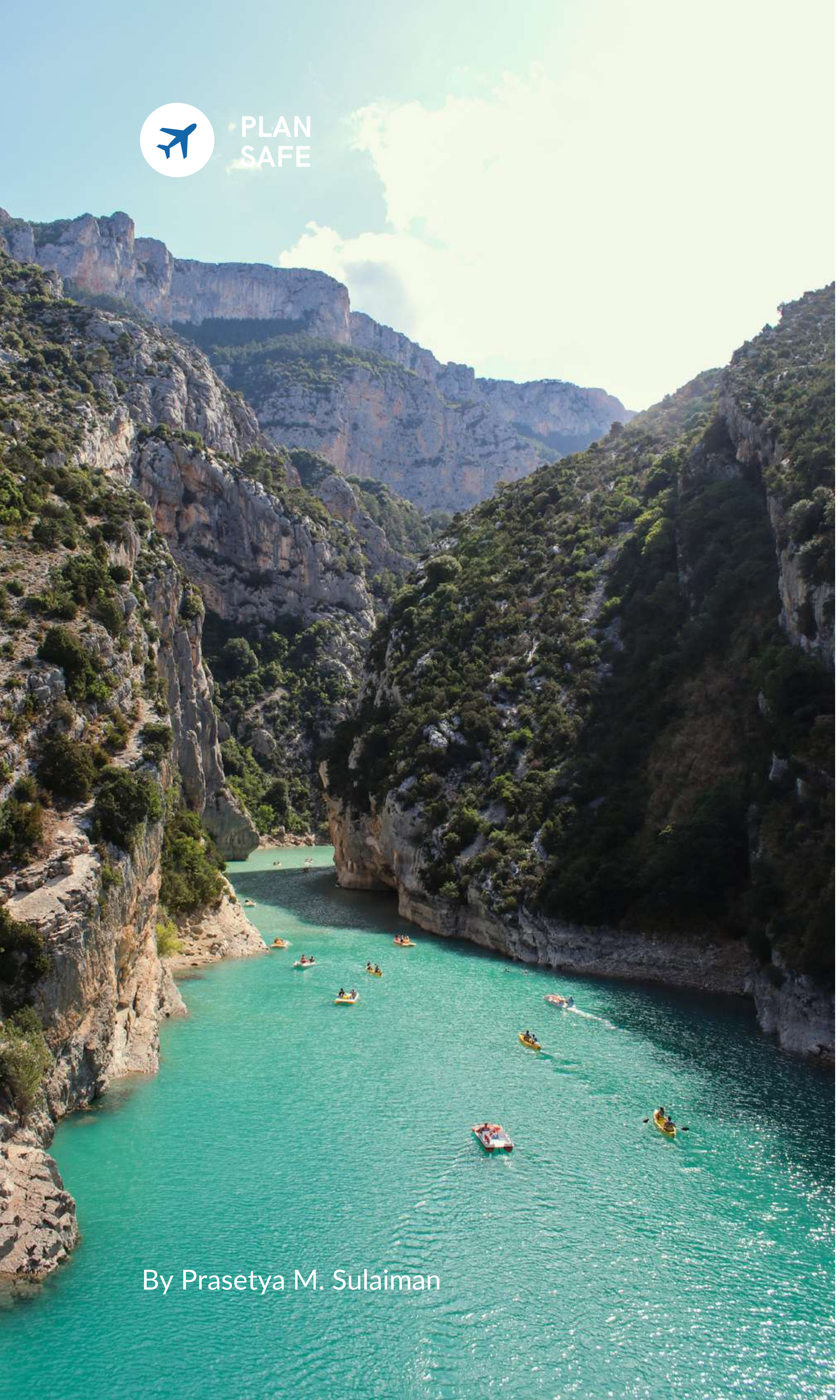
PLAN
SAFE

Travel Insurance

Data Science



By Prasetya M. Sulaiman

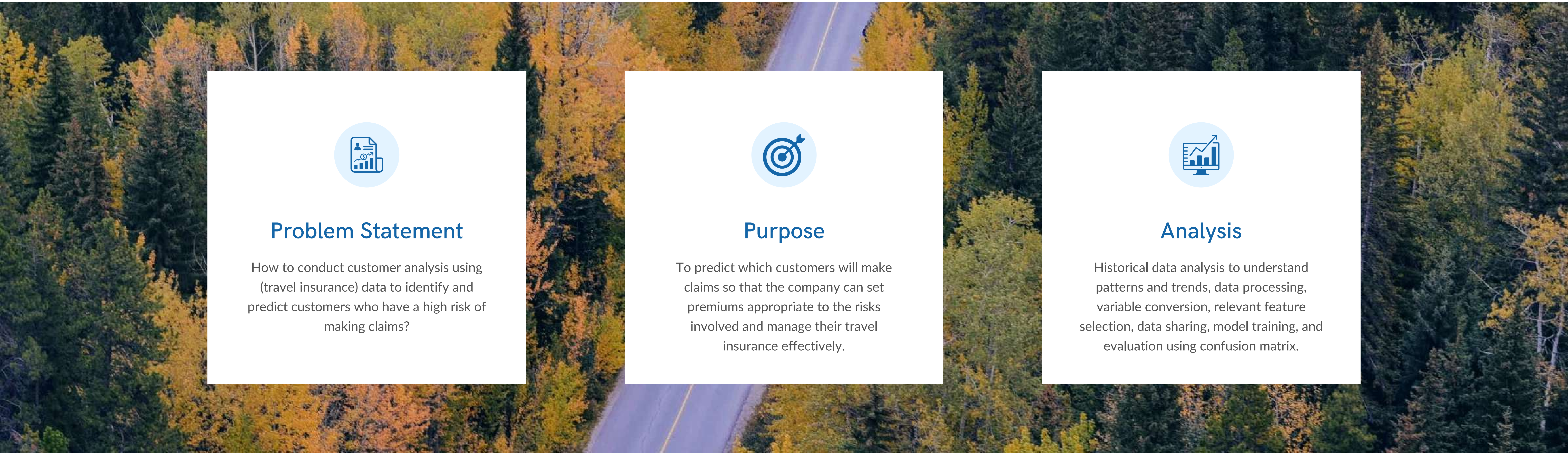


Context

Travel insurance is a type of insurance that provides protection as long as we travel both domestically and abroad. Several countries have even required travelers to have travel insurance, for example, countries in Europe and America. The amount of premium depends on the coverage desired, the length of the trip, and the purpose of the trip. A company engaged in travel insurance wants to know the policyholder who will submit an insurance claim for coverage. Policyholder data at insurance companies is historical data consisting of destinations, insurance products, and so on.

PLAN
SAFE

Business Problem Understanding



Problem Statement

How to conduct customer analysis using (travel insurance) data to identify and predict customers who have a high risk of making claims?

Purpose

To predict which customers will make claims so that the company can set premiums appropriate to the risks involved and manage their travel insurance effectively.

Analysis

Historical data analysis to understand patterns and trends, data processing, variable conversion, relevant feature selection, data sharing, model training, and evaluation using confusion matrix.

PLAN
SAFE

Metric Evaluation

Type 1 error : False Positive

Type 1 Error (False Positive): An error in classifying something as positive (Yes/1) when it is not (No/0). This means that a claim was made (1) when in fact no claim was made (0).

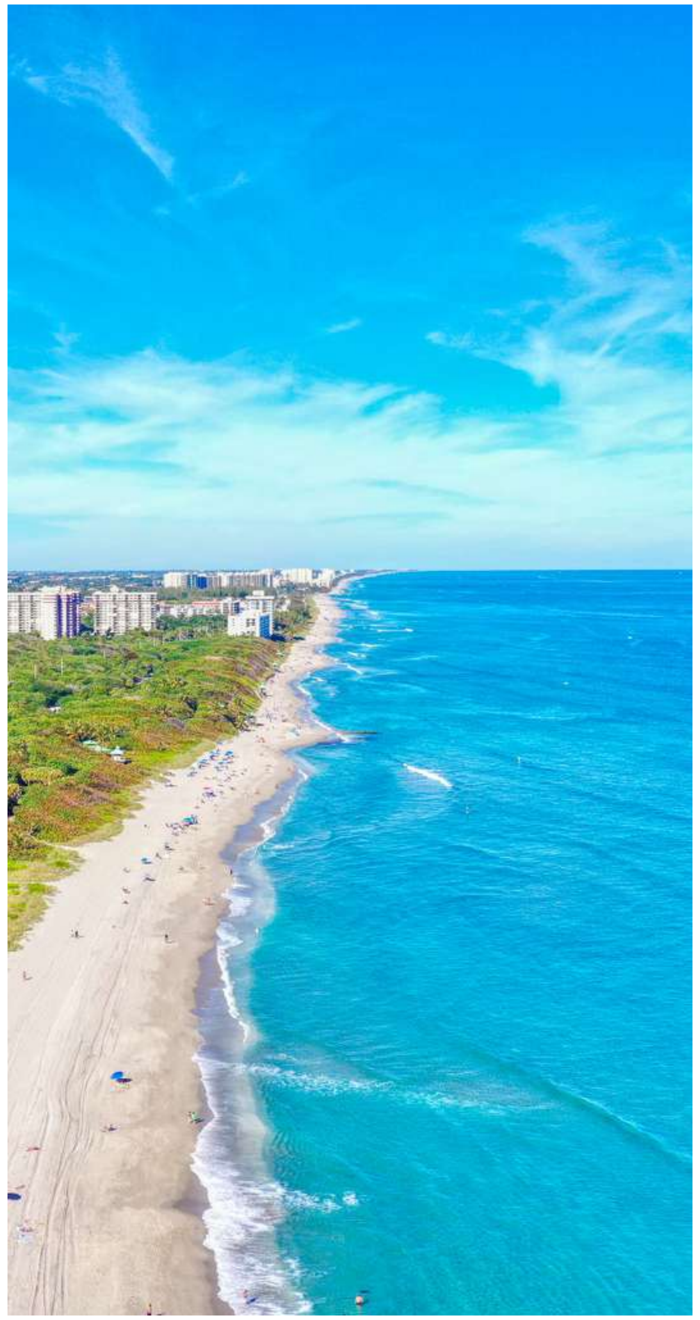
Type 2 error : False Negative

Type 2 Error (False Negative): An error in classifying something as negative (No/0) when it is actually positive (Yes/1). This means that no claim was made (0) when in fact a claim was made (1).

By Prasetya M. Sulaiman

PLAN
SAFE

Step by Step



Data Understanding ....
Data Cleaning ....
Data Analysis ....
Data Preparation ....
Modeling and Evaluation ....
Conclusion ....
Recomendation ....

By Prasetya M. Sulaiman

Data Understanding

By Prasetya M. Sulaiman

AGENCY

AGENCY TYPE

DISTRIBUTION CHANNEL

PRODUCT NAME

GENDER

DURATION

DESTINATION

NET SALES

COMMISSION (IN VALUE)

AGE

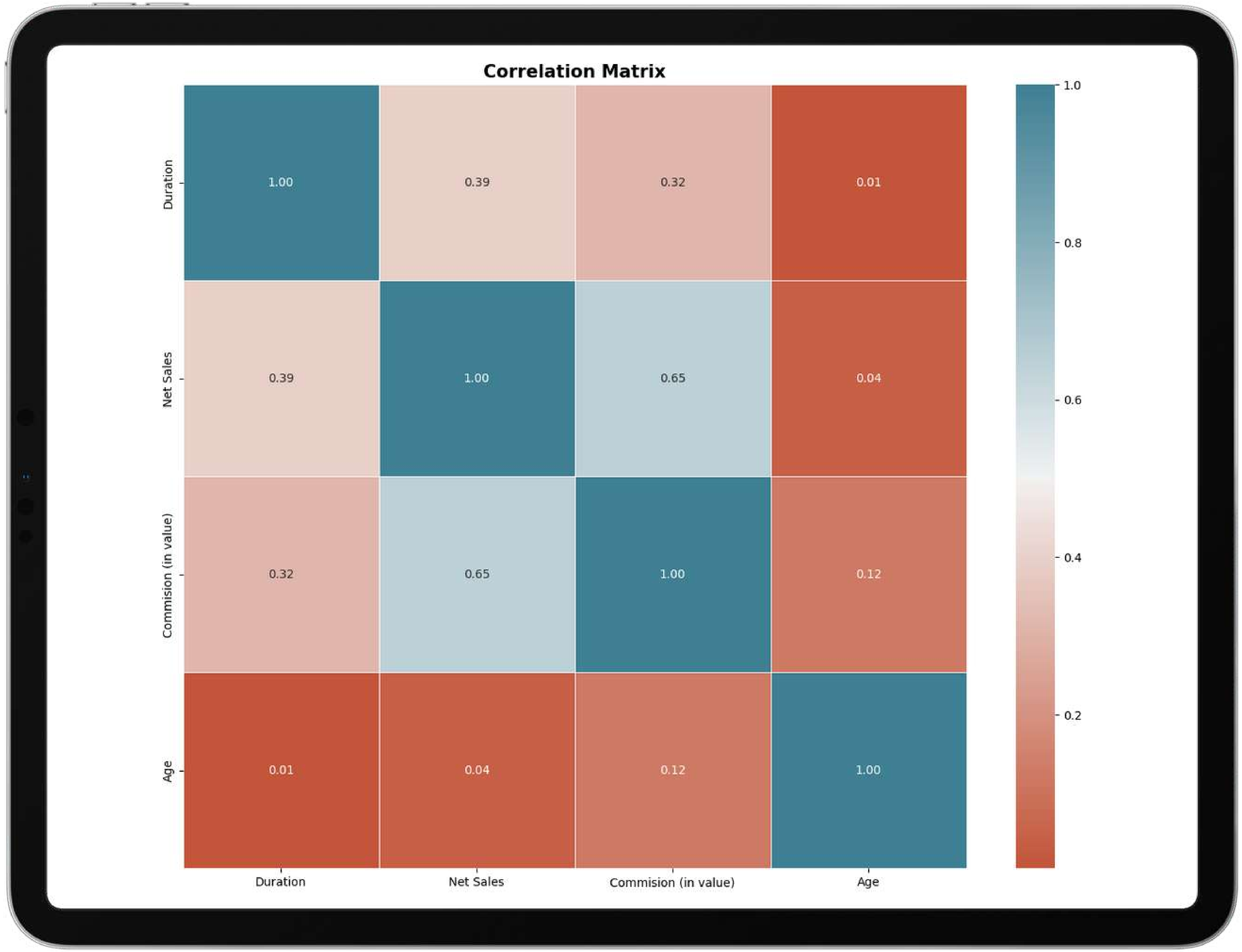
CLAIM

Data Cleaning

By Prasetya M. Sulaiman

Correlation Matrix

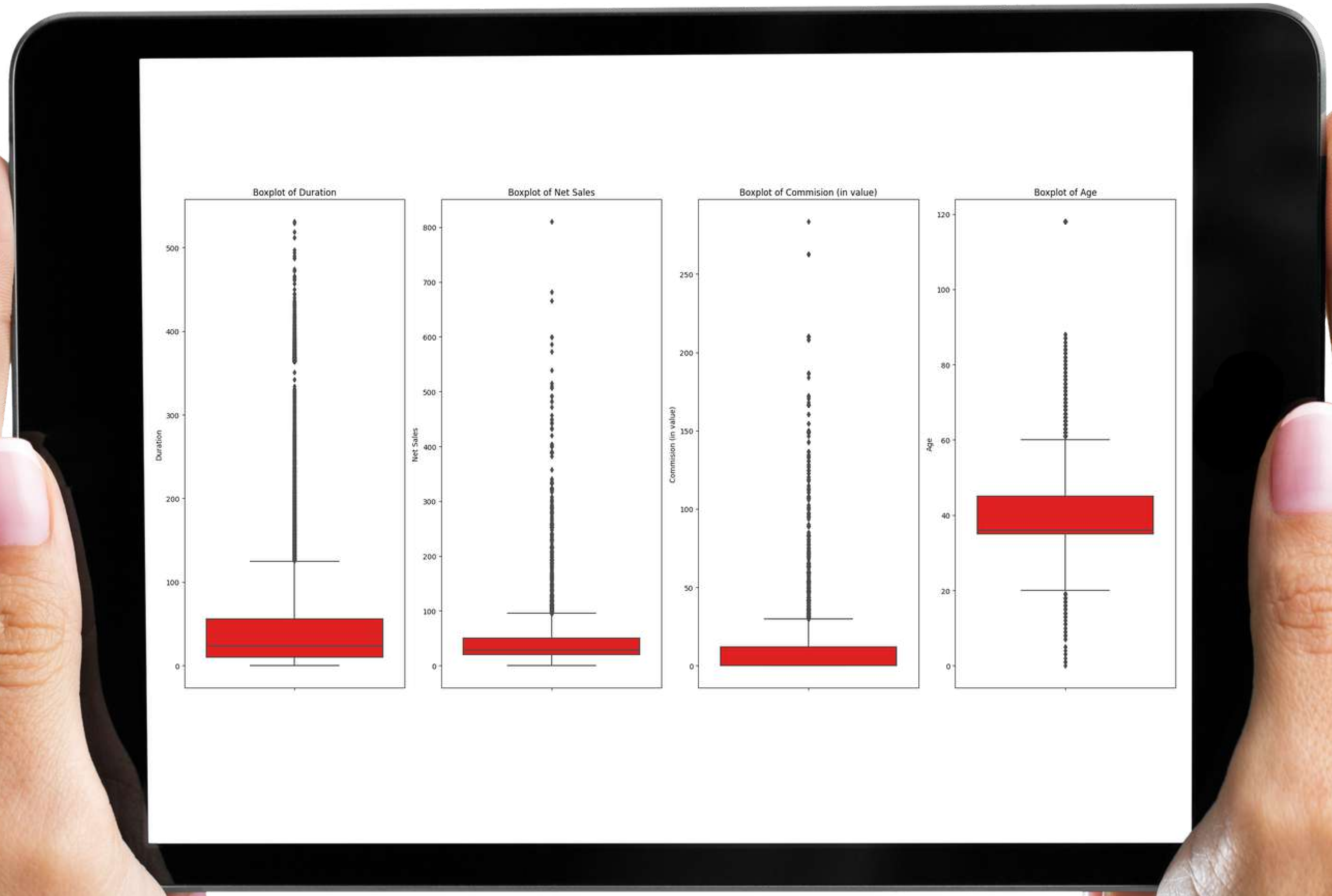
In this correlation matrix, insurance policy duration (Duration) has a moderate positive correlation with net sales and commission in value, and a weak correlation with age. Net sales and commission in value also have a moderate positive correlation. Age has a weak correlation with other variables. In conclusion, the longer the insurance policy duration, the higher the net sales and commission values, while the correlation with age is weak. However, it is important to remember that correlations do not indicate a cause-and-effect relationship, and further analyses are required for a more in-depth understanding.



Correlation Matrix

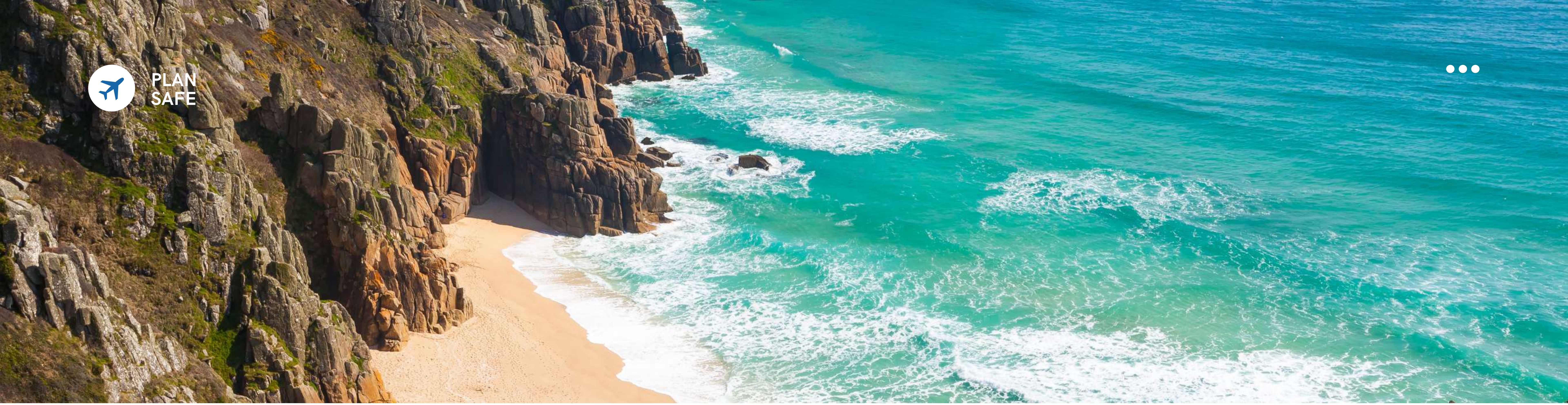
	Duration	Net Sales	Commision (in value)	Age
Duration	1.00	0.39	0.32	0.01
Net Sales	0.39	1.00	0.65	0.04
Commision (in value)	0.32	0.65	1.00	0.12
Age	0.01	0.04	0.12	1.00

PLAN
SAFE



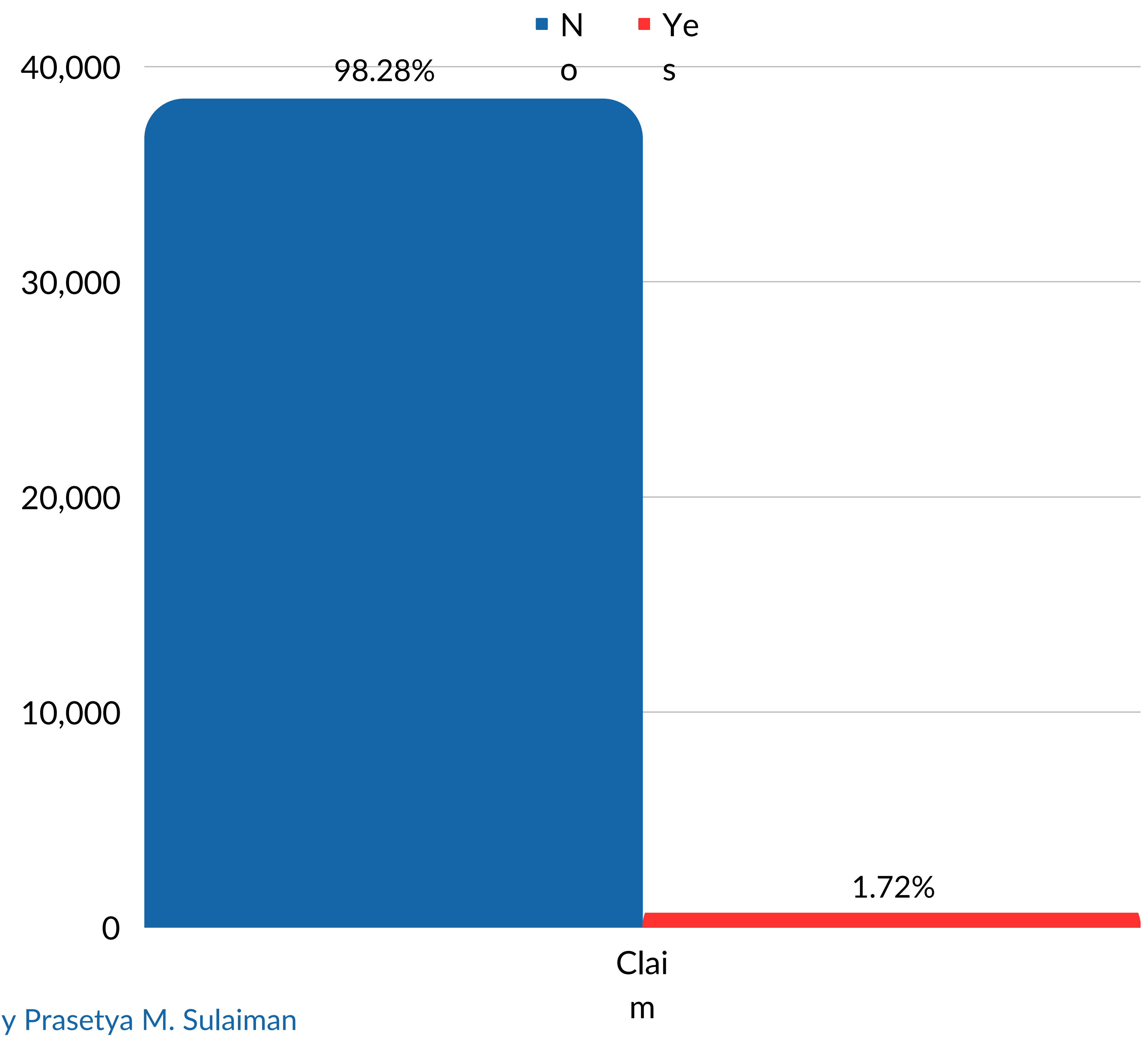
Boxplot

By Prasetya M. Sulaiman



Data Analysis

By Prasetya M. Sulaiman



Count Plot of Claim

From the count plot for the variable 'Claim', it can be seen that there are two main categories, '0' and '1', which indicate whether a claim was filed (1) or not (0).

In this data, there are 38,491 data that did not file a claim (category '0'), while only 675 data that filed a claim (category '1').

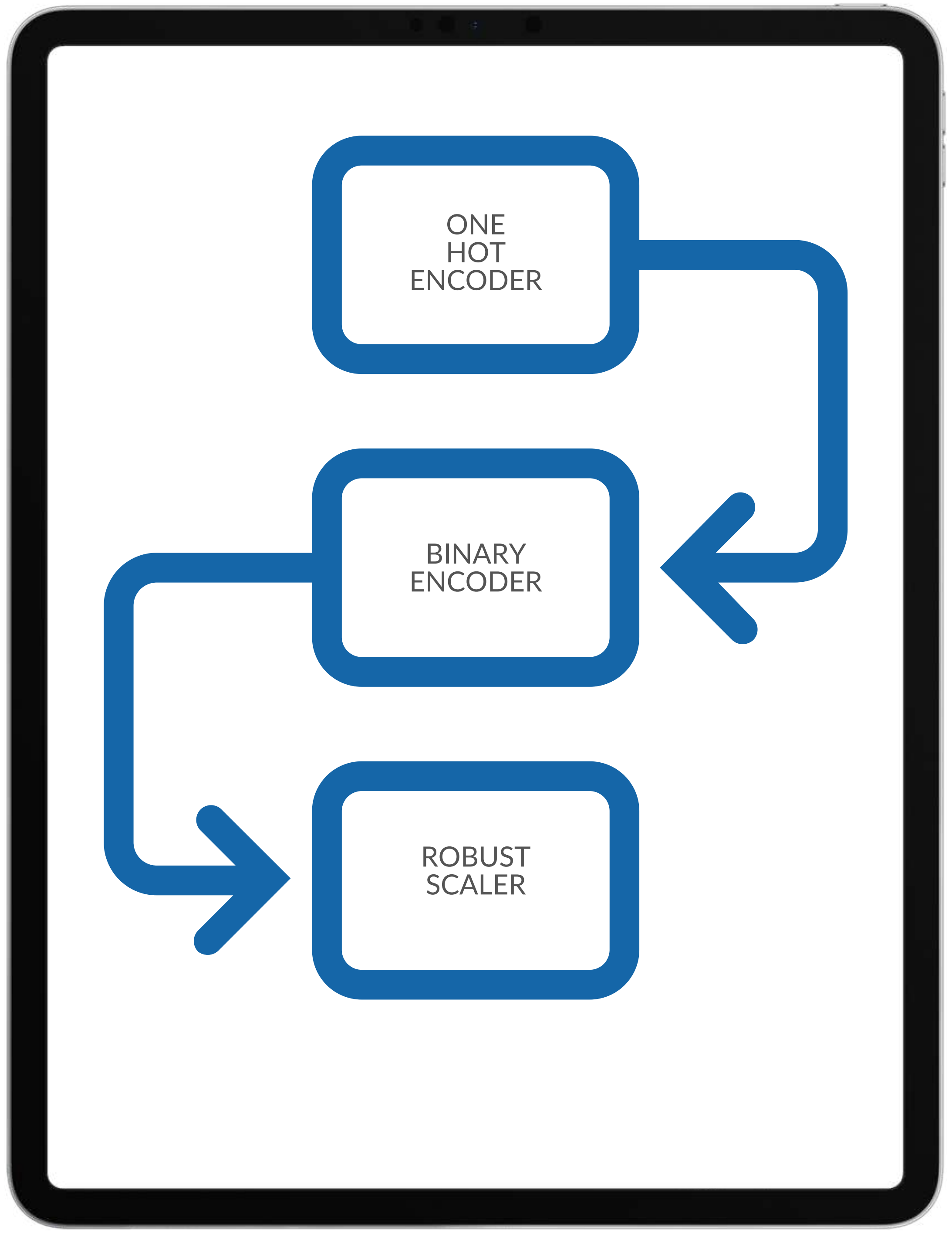
Thus, from this information, the number of data that did not submit a claim is much greater than the number of data that did. This indicates a significant imbalance in the distribution of the 'Claim' variable. Therefore, when conducting further analysis and modelling, it is necessary to consider strategies to deal with this imbalance, such as oversampling or undersampling the less representative categories.

Data Preparation

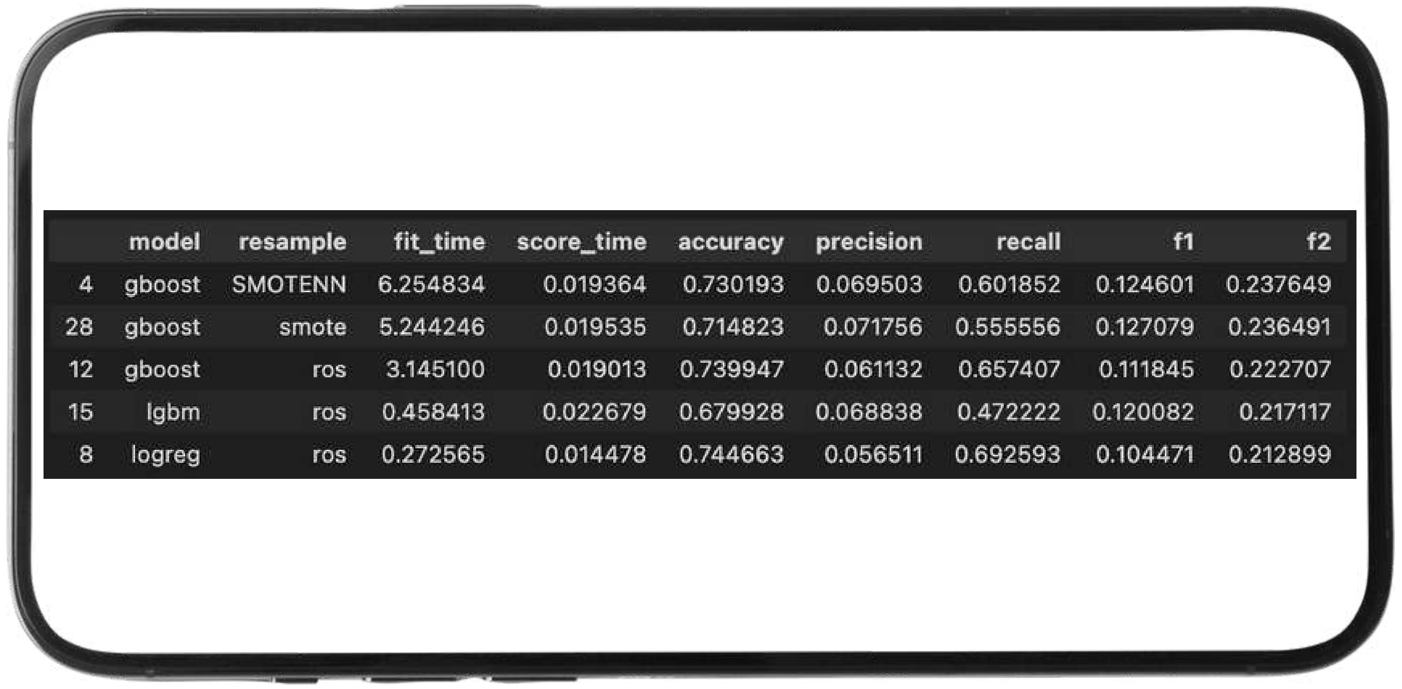
By Prasetya M. Sulaiman

PLAN
SAFE

Transformer


ONE HOT ENCODER → BINARY ENCODER → ROBUST SCALER

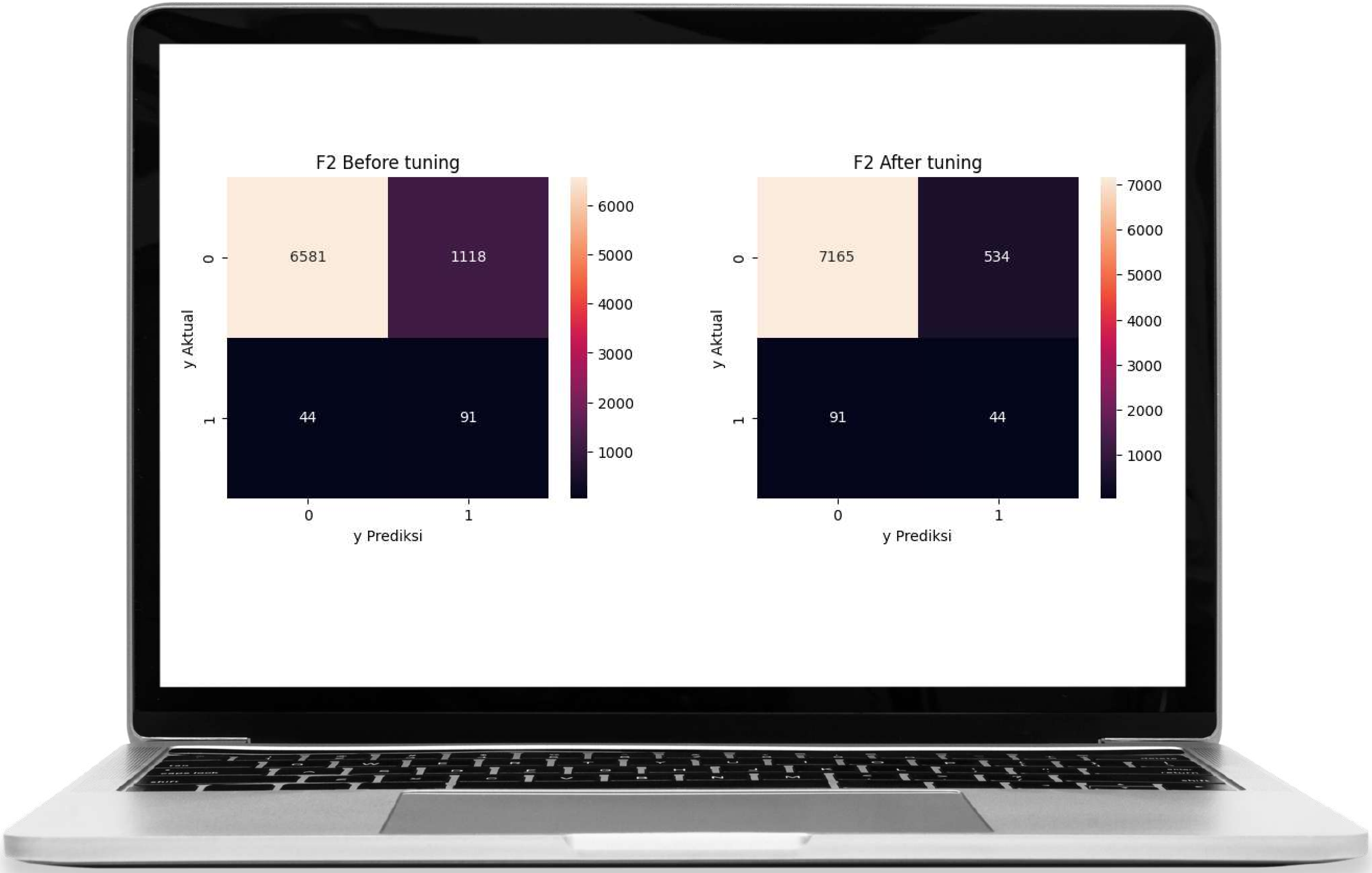
Modeling



	model	resample	fit_time	score_time	accuracy	precision	recall	f1	f2
4	gboost	SMOTENN	6.254834	0.019364	0.730193	0.069503	0.601852	0.124601	0.237649
28	gboost	smote	5.244246	0.019535	0.714823	0.071756	0.555556	0.127079	0.236491
12	gboost	ros	3.145100	0.019013	0.739947	0.061132	0.657407	0.111845	0.222707
15	lgbm	ros	0.458413	0.022679	0.679928	0.068838	0.472222	0.120082	0.217117
8	logreg	ros	0.272565	0.014478	0.744663	0.056511	0.692593	0.104471	0.212899

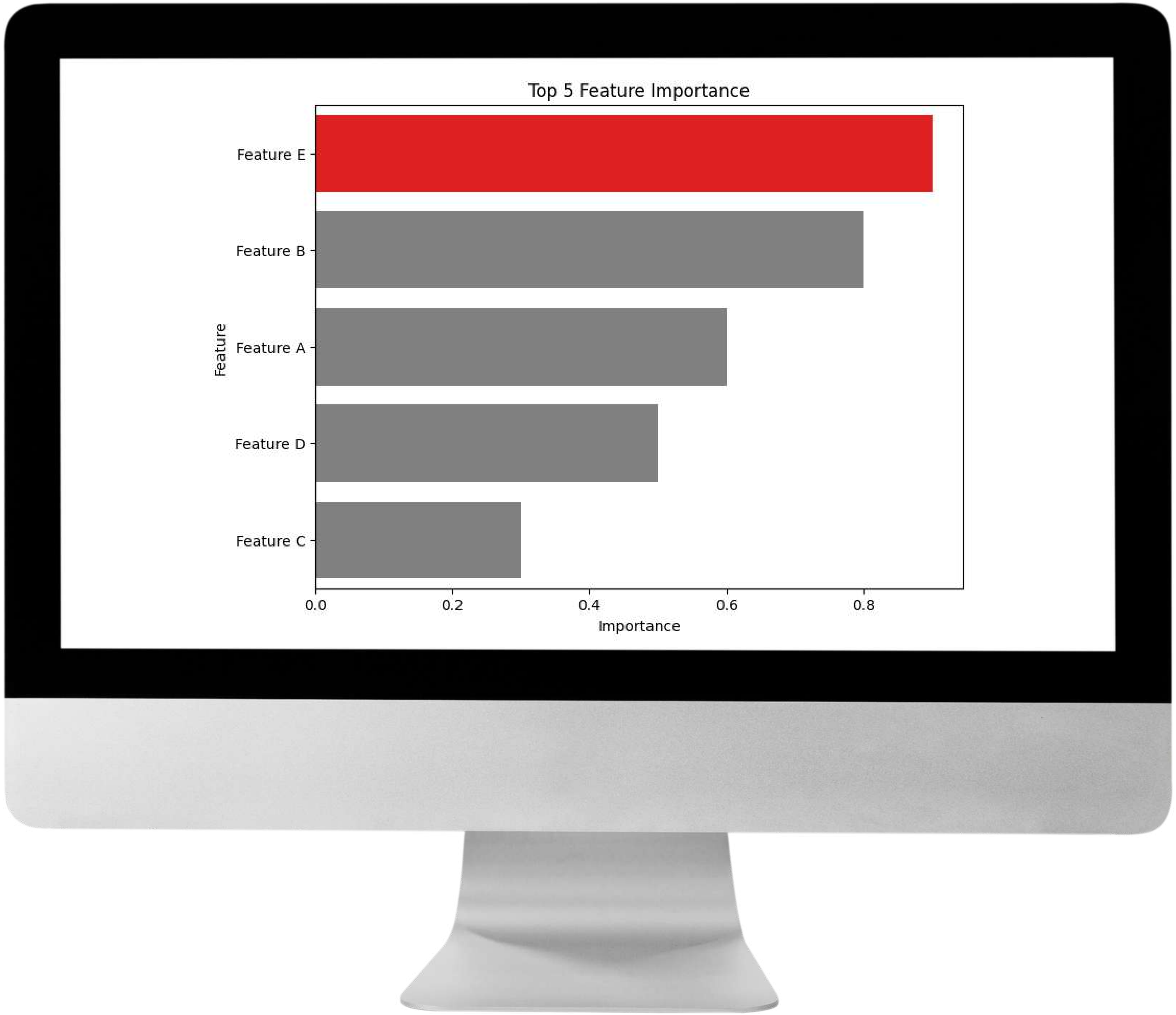
By Prasetya M. Sulaiman

Before & After Tuning



PLAN
SAFE

Feature Importance



By Prasetya M. Sulaiman

Conclusion

Metrix Result:

Based on the results of the confusion matrices before and after tuning, as well as the given DataFrames, we can make some conclusions:

1. Before Tuning:
   - The confusion matrices before tuning show that the model has a good tendency in predicting the negative class (Actual 0) with `6581` correct predictions and `1118` incorrect predictions.
   - However, the model has a poor performance in predicting the positive class (Actual 1) with only `91` correct predictions and `44` incorrect predictions.

2. After Tuning:
   - The confusion matrix after tuning shows improvement in predicting the negative class (Actual 0) with `7172` correct predictions and `527` incorrect predictions.
   - However, the performance in predicting the positive class (Actual 1) remains poor with only `39` correct predictions and `96` incorrect predictions.

Feature Importance:

Based on the feature importance results, we can make some conclusions about the importance of the features in the model used:

1. Most important features:
   - Agency_2 has the highest importance with a value of `0.314`. This feature indicates that the Agency_2 category has a significant influence in predicting the model target or output.
   - Net Sales is also an important feature with an importance of `0.162`. This indicates that the Net Sales value has a significant contribution in influencing the prediction results.

2. Other features:
   - Besides Agency_2 and Net Sales, there are several other features that also have a significant contribution in predicting targets.
   - Duration, Age, and Product Name_3 also have high importance in the model.

3. Features with low importance:
   - Some features such as Agency_0, Destination_0, and Product Name_0 have very low importance, with almost insignificant contribution in predicting the target.

PLAN SAFE

Recomendation

Business:

Analysing insurance claims data and travel risk factors can help you in smarter pricing. By understanding different travel risks, you can adjust premiums based on factors such as travel destination, customer age, type of travel activity, and so on. Risk-adjusted pricing can help optimise revenue and minimise the risk of high claims.

Model:

Since this dataset is imbalanced, we should use additional imbalanced-learn is a Python library that provides various algorithms and methods to handle imbalanced datasets. The following are some of the algorithms provided by Imbalanced-Learn:

Oversampling:

- ADASYN (Adaptive Synthetic Sampling)
- BorderlineSMOTE
- SVMSMOTE (Support Vector Machines SMOTE)
- KMeansSMOTE
- SMOTENC (SMOTE for Nominal and Continuous features)
- RandomOverSampler

Undersampling:

- TomekLinks
- EditedNearestNeighbours
- CondensedNearestNeighbour
- NeighbourhoodCleaningRule
- OneSidedSelection
- InstanceHardnessThreshold

Combined Sampling:

- SMOTETomek (SMOTE combined with Tomek Links)

Data source: https://imbalanced-learn.org/stable/references/under_sampling.html

By Prasetya M. Sulaiman

Thank You