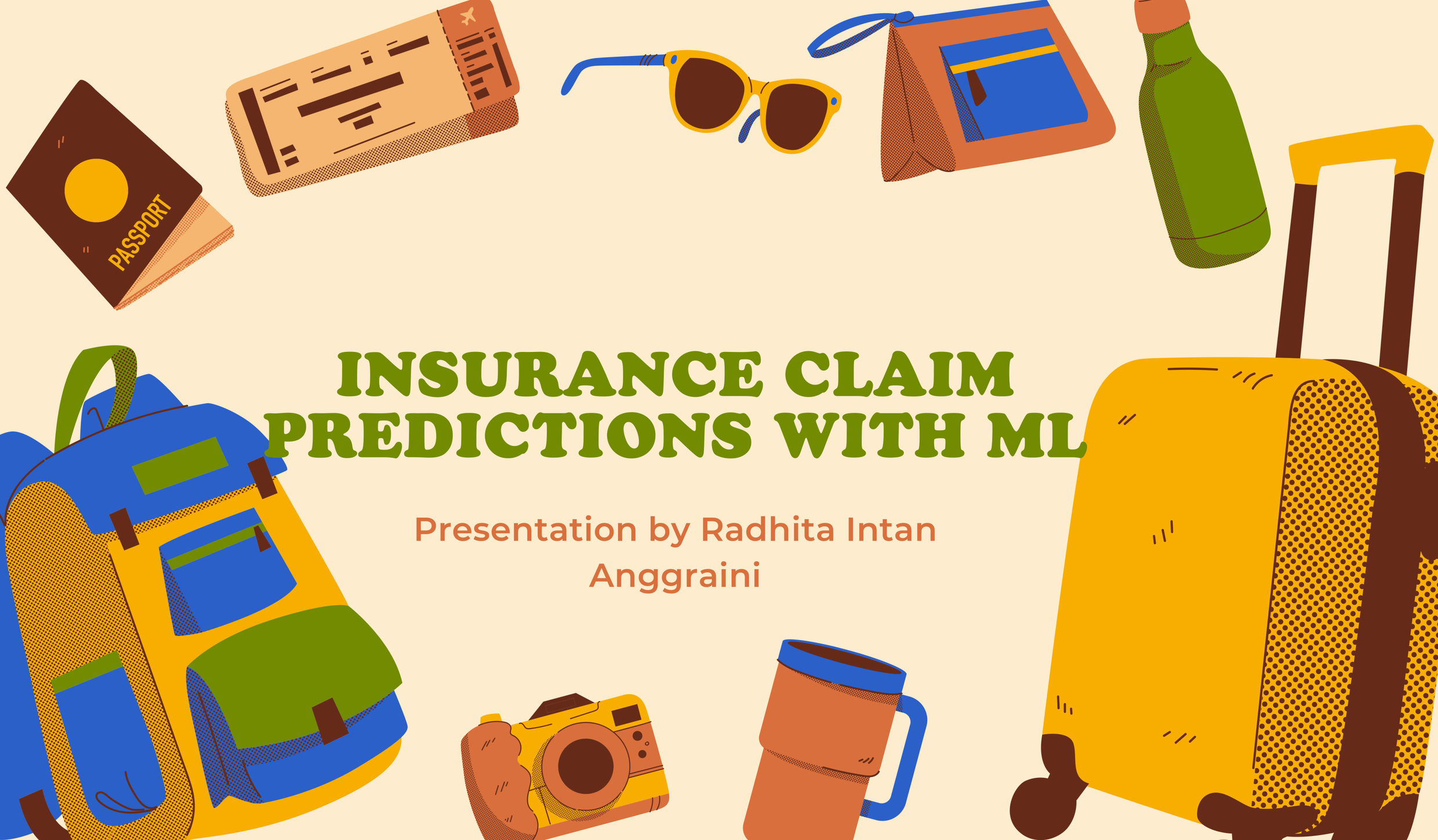


INSURANCE CLAIM PREDICTIONS WITH ML

Presentation by Radhita Intan Anggraini

Problems	Condition	Treatment
Duplicates	10,53%	drop
Null Values	Gender 69,7% missing	fillna (prefer not to say)
Outliers	Net Sales, Duration, Comission (in value), Age all have outliers	drop extreme values
Data Imbalance	98,63% No Claim 1,31% Yes Claim	Resampling (SMOTE, RUS)

Data Analysis Finding

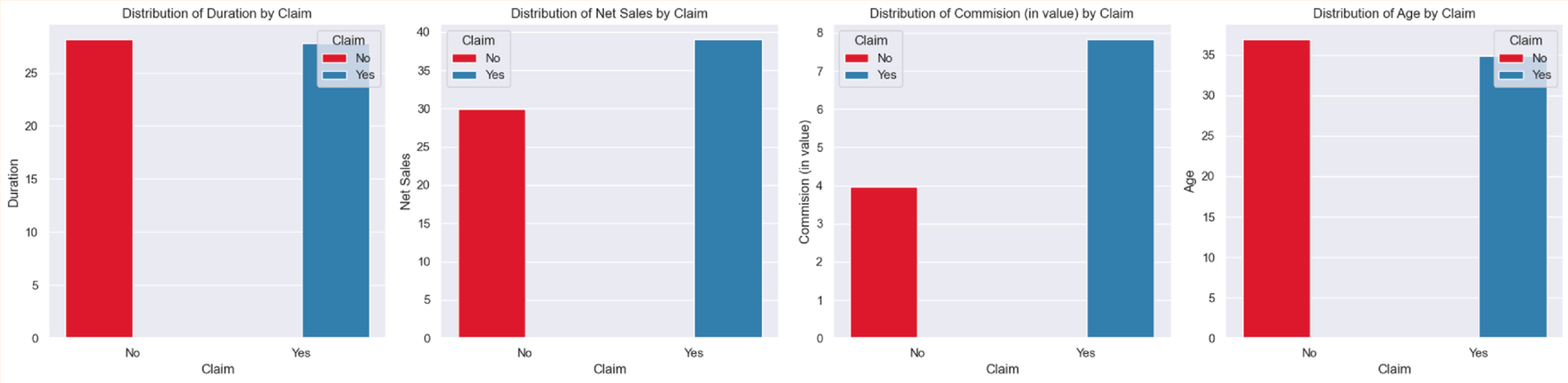
Numerical

Duration

Net Sales

Comission(in value)

Age



overall insight:

Policies that lead to claims tend to have longer durations, higher net sales, and generate more commission. on average younger travelers appear to make more claims than older travelers.

Data Analysis Finding

Categorical

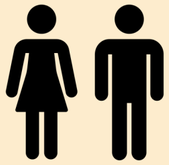
Agency

Agency Type

Distribution Channel

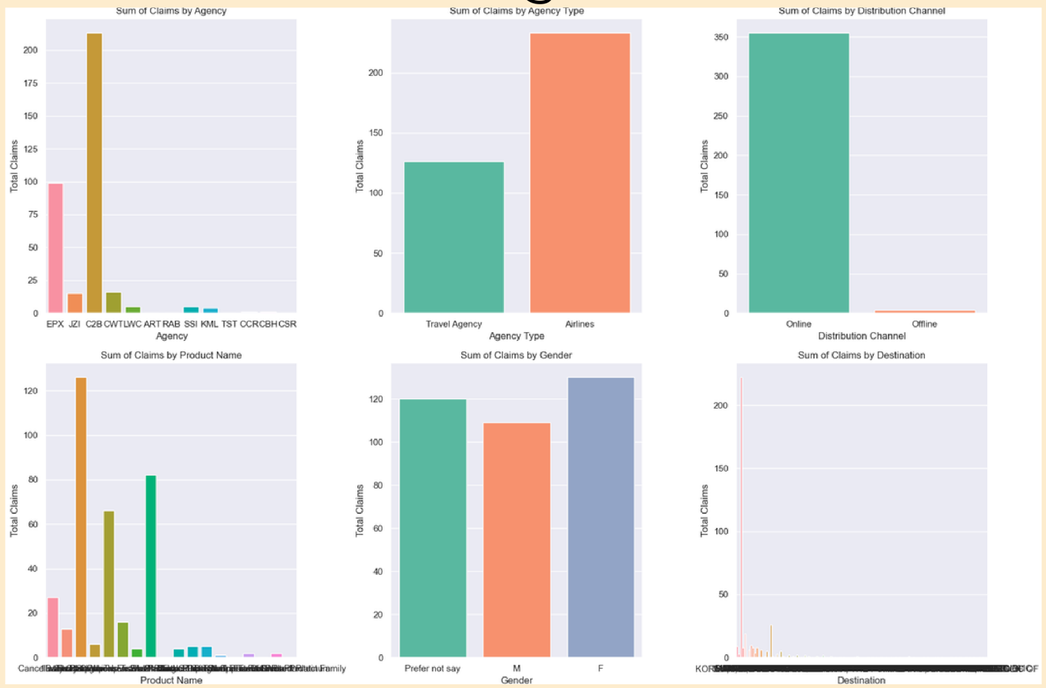
Product Name

Gender



Destination

insight



Business Problem

Travel Insurance Claims Paid Out 6X Policy Premium in 2023

ST. PETERSBURG, Fla., March 21, 2024 - New data reveals a 30% increase in paid travel insurance claims filed in 2023. Squaremouth.com, the

Squaremouth Press Room / Apr 2

Background

Predicting insurance claims accurately is vital in the insurance industry for risk management, operational efficiency, and customer satisfaction. Companies aim to forecast claim likelihood using customer data like travel duration, net sales, commission, and age to enhance risk assessment, policy pricing, and fraud detection.

Problem Statement

The current insurance claims predictive model needs improvement to reduce missed claims, focusing on maximizing recall for better risk management, customer trust, and regulatory compliance.

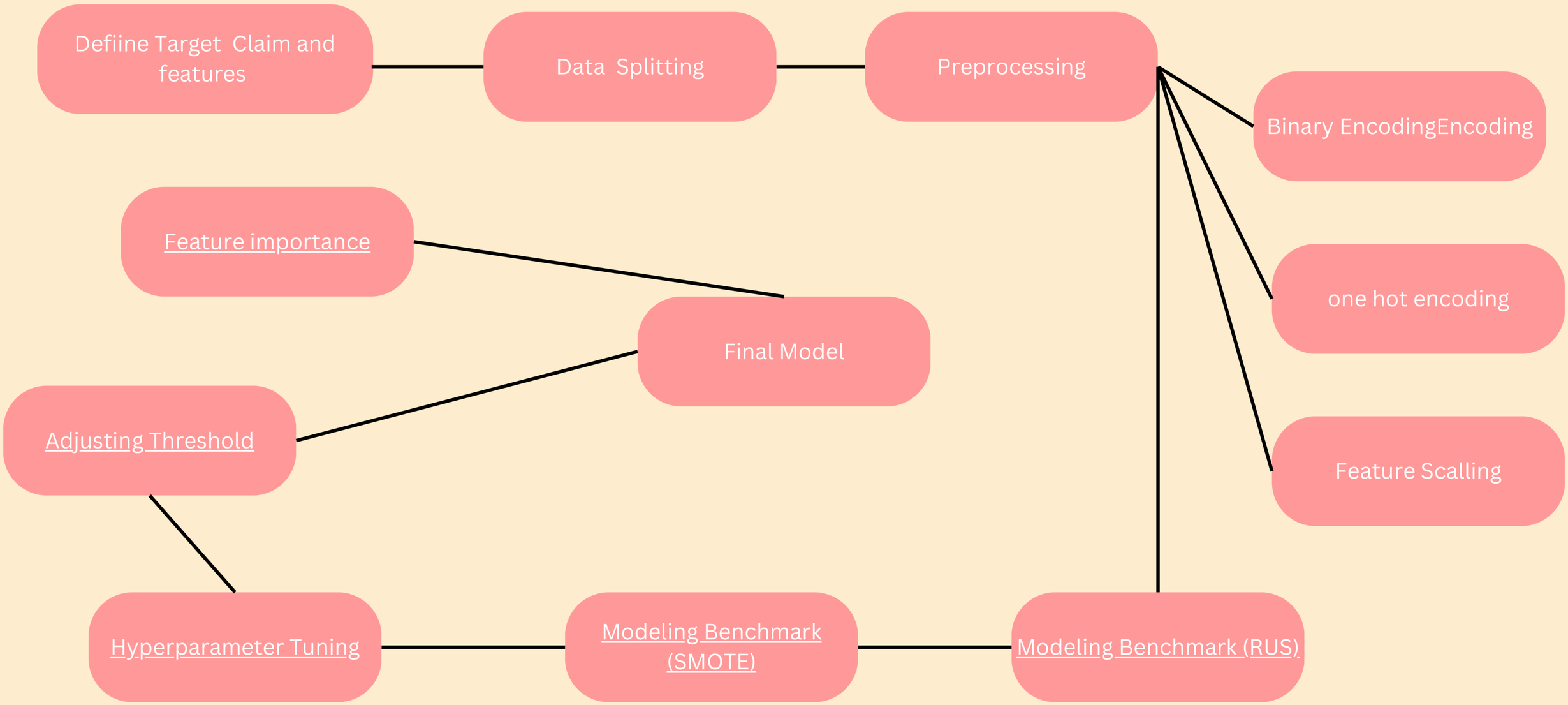
Objectives

Develop Classification model to predict insurance claims based on travel and demographic data. Aim to maximize recall, address class imbalance with techniques like SMOTE, select important features, and optimize the decision threshold for better sensitivity in detecting claims.

Goals

Improving risk management by predicting claims accurately, enhancing customer satisfaction, ensuring regulatory compliance, and optimizing operational efficiency by reducing false positives through model threshold adjustments.

Data Processing and Modeling Chart Flow



Modeling Benchmark (RUS)

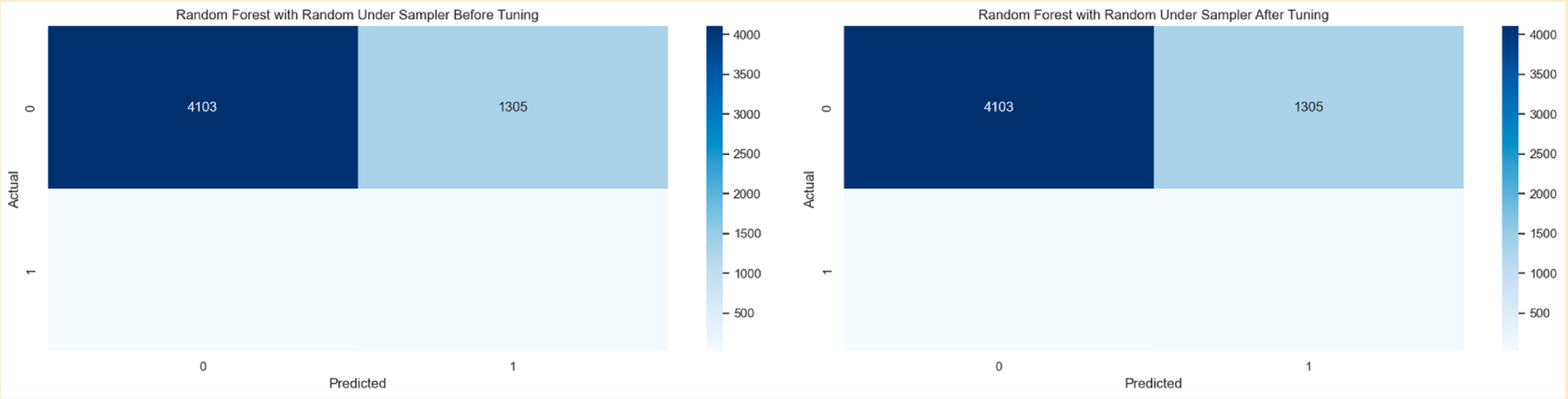
Model	Test Score	Train Score	Train Std Dev	Difference
RandomForestClassifier	0.791667	0.703811	0.076668	0.087855
LogisticRegression	0.722222	0.689837	0.077842	0.032386
AdaBoostClassifier	0.708333	0.637629	0.042008	0.070705
KNeighborsClassifier	0.694444	0.710768	0.067216	0.016324
GradientBoostingClassifier	0.694444	0.686509	0.065898	0.007935
DecisionTreeClassifier	0.680556	0.620448	0.046612	0.060108
XGBClassifier	0.680556	0.637629	0.033899	0.042927

Modeling Benchmark (SMOTE)

Model	Test Score	Test Std Dev	Train Score	Train Std Dev	Difference
LogisticRegression	0.736111	0	0.630672	0.079116	0.105440
GradientBoostingClassifier	0.569444	0	0.491470	0.091565	0.077974
KNeighborsClassifier	0.305556	0	0.285602	0.069980	0.019954
DecisionTreeClassifier	0.125000	0	0.076891	0.039629	0.048109
XGBClassifier	0.125000	0	0.094253	0.045427	0.030747
RandomForestClassifier	0.097222	0	0.090563	0.025476	0.006660
AdaBoostClassifier	0.097222	0	0.073200	0.023355	0.024022

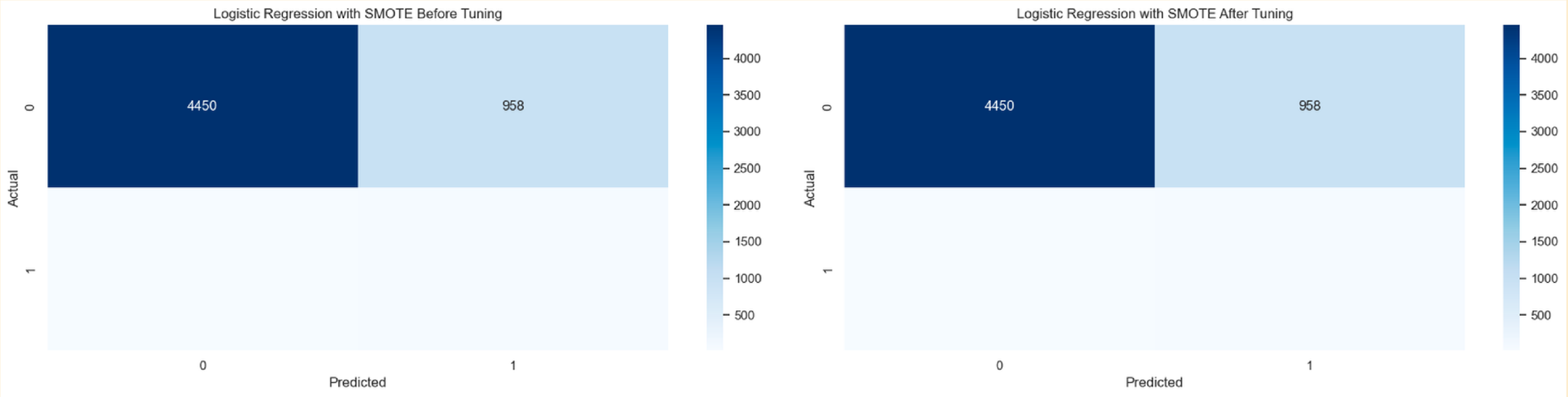
Hyperparameter Tuning (RUS)

```
Random Forest with Random Under Sampler - Test Set Evaluation
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.76      0.86      5408
           1       0.04      0.74      0.07        72

    accuracy                           0.76      5480
   macro avg       0.52      0.75      0.47      5480
weighted avg       0.98      0.76      0.85      5480

Confusion Matrix:
[[4103 1305]
 [  19   53]]
```



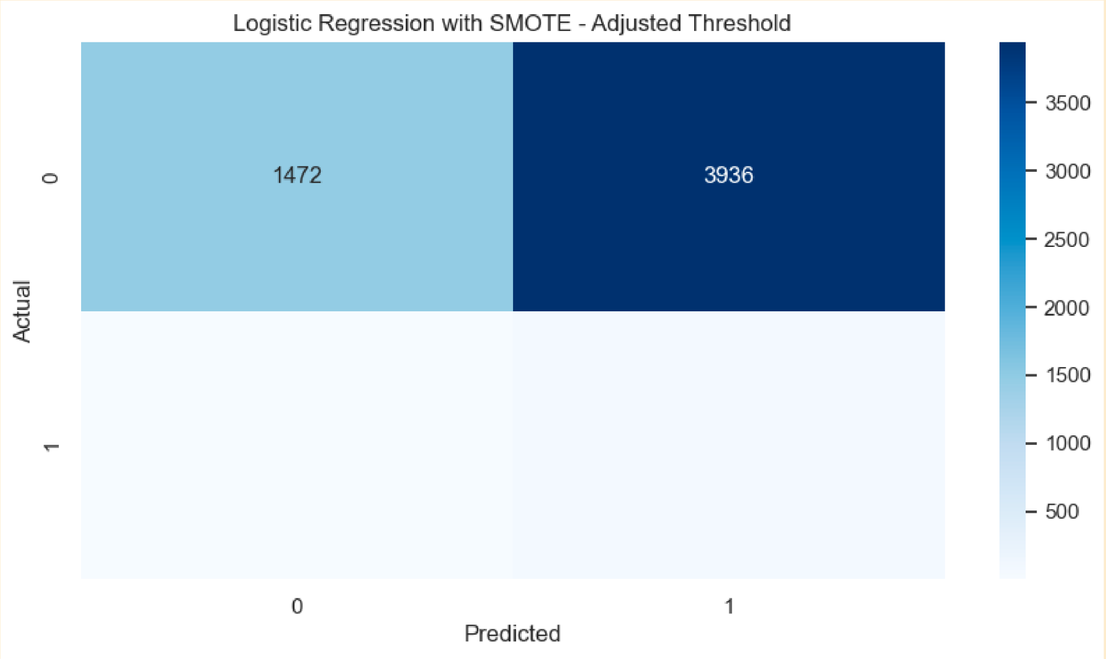
Hyperparameter Tuning (SMOTE)

```
Logistic Regression with SMOTE – Test Set Evaluation
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.82      0.90      5408
           1       0.05      0.74      0.10        72

    accuracy                           0.82      5480
   macro avg       0.52      0.78      0.50      5480
weighted avg       0.98      0.82      0.89      5480

Confusion Matrix:
[[4450  958]
 [  19   53]]
```



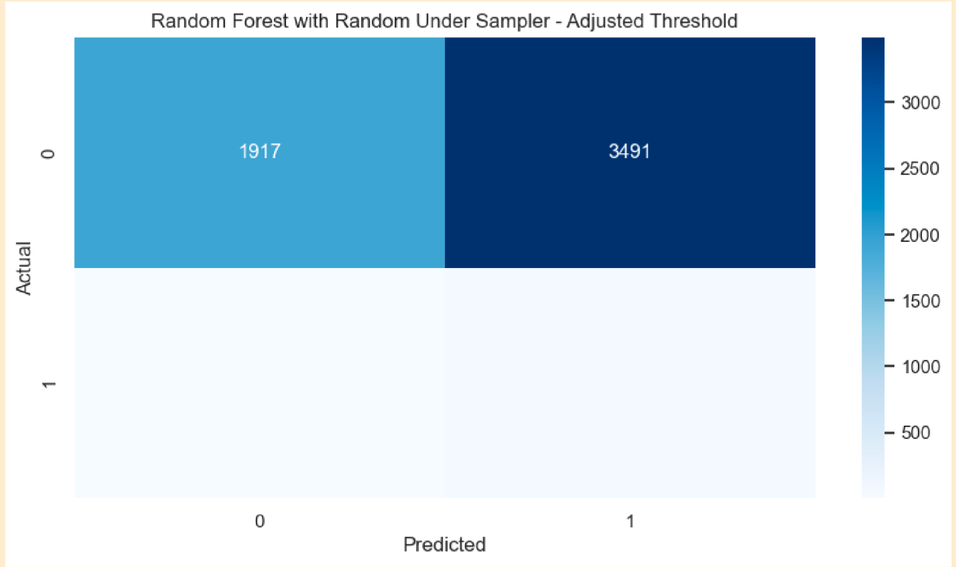
Adjusting Threshold

```
Logistic Regression with SMOTE - Adjusted Threshold Evaluation
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.27      0.43      5408
           1       0.02      0.93      0.03        72

    accuracy                           0.28      5480
   macro avg       0.51      0.60      0.23      5480
weighted avg       0.98      0.28      0.42      5480

Confusion Matrix:
[[1472 3936]
 [   5   67]]
```



Adjusting Threshold

```
Random Forest with Random Under Sampler - Adjusted Threshold Evaluation
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.35      0.52      5408
           1       0.02      0.92      0.04        72

    accuracy                           0.36      5480
   macro avg       0.51      0.64      0.28      5480
weighted avg       0.98      0.36      0.52      5480

Confusion Matrix:
[[1917 3491]
 [   6   66]]
```

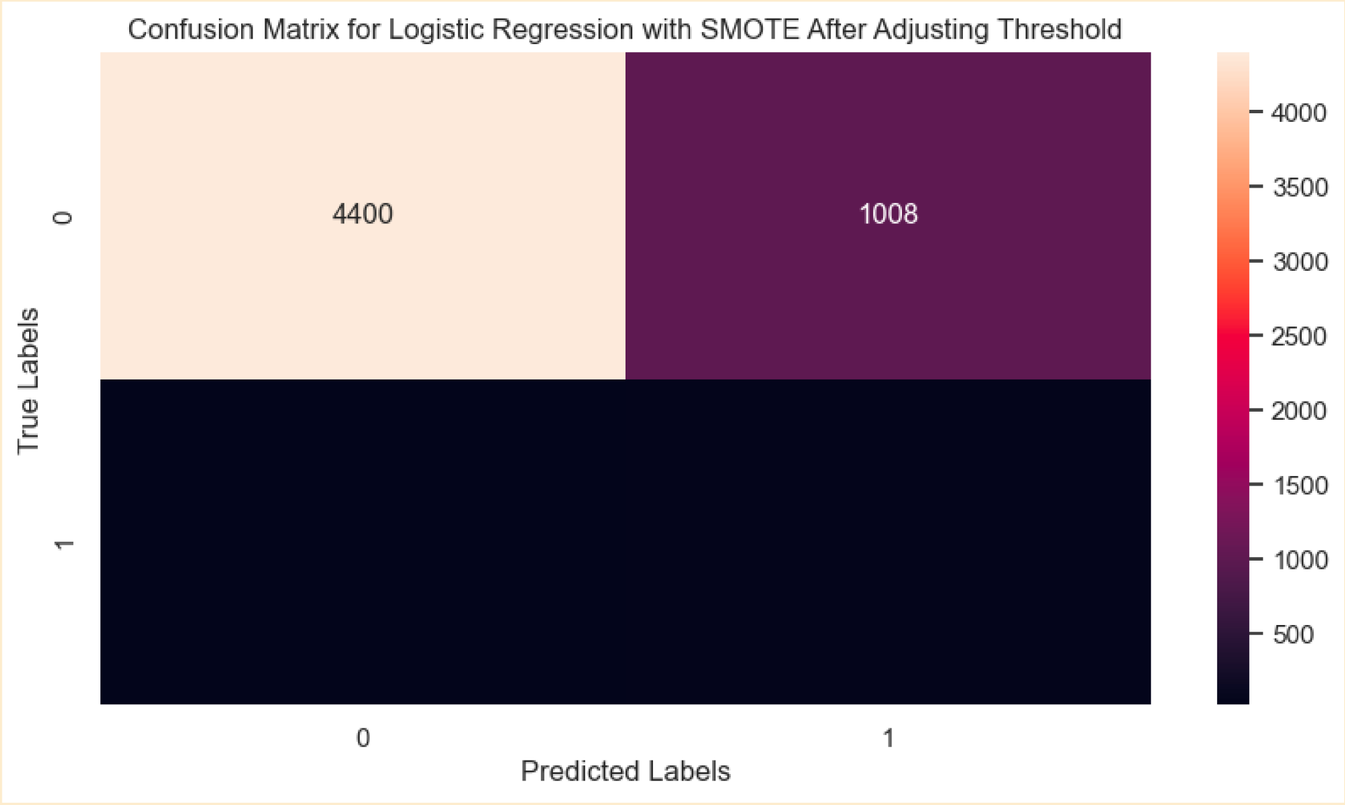


Top Model

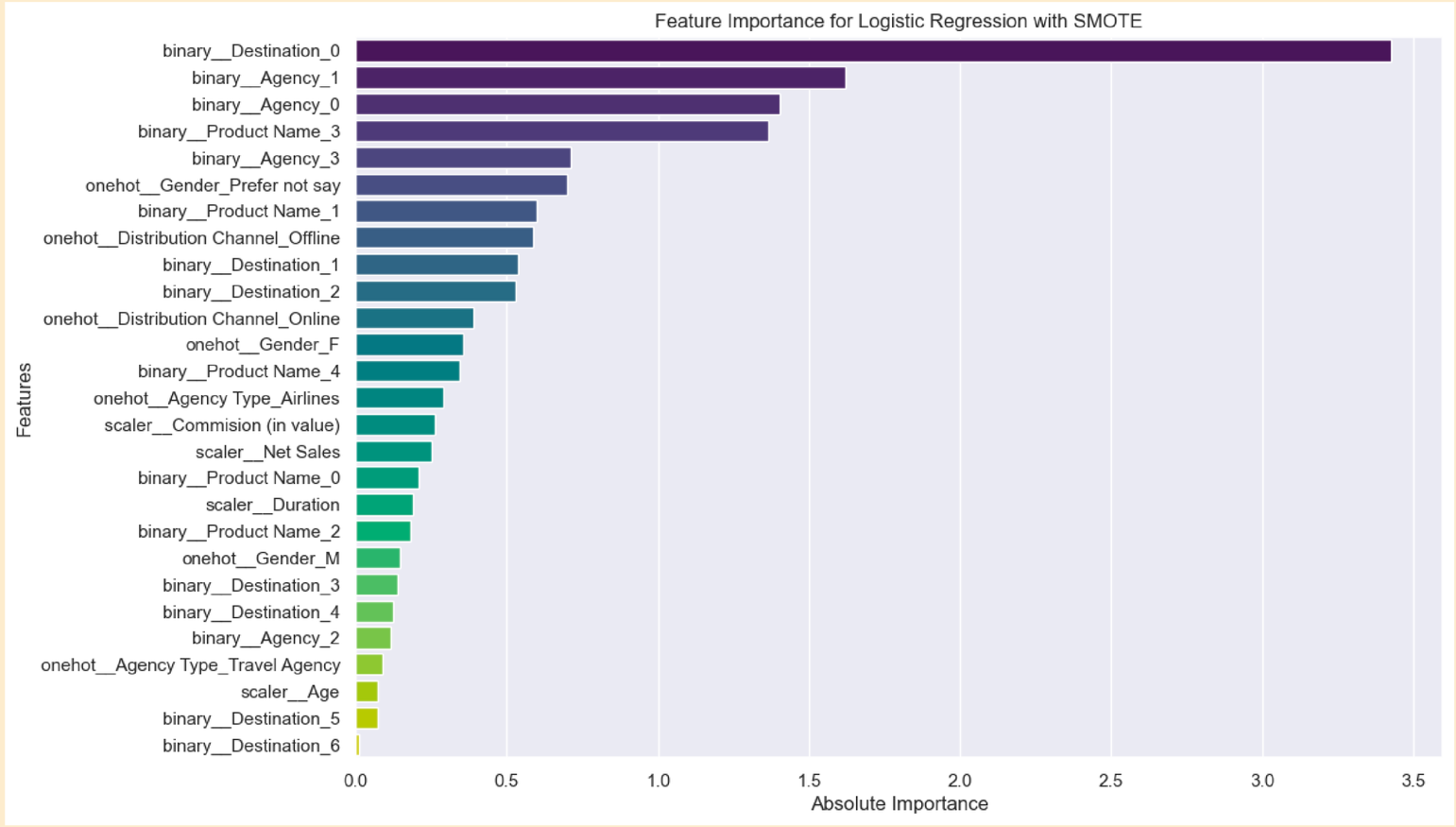
	Metric	Logistic Regression	Random Forest
0	Precision	0.016737	0.018555
1	Recall	0.930556	0.916667
2	F1-Score	0.032883	0.036374

Final Model

```
Logistic Regression with SMOTE - Initial Evaluation
              precision    recall  f1-score   support

           0       1.00      0.82      0.90      5408
           1       0.05      0.74      0.10        72

    accuracy                           0.82      5480
   macro avg       0.52      0.78      0.50      5480
weighted avg       0.98      0.82      0.89      5480

Confusion Matrix:
[[4451  957]
 [  19   53]]
Logistic Regression with SMOTE - Test Set Evaluation (Adjusted Threshold)
              precision    recall  f1-score   support

           0       1.00      0.81      0.90      5408
           1       0.05      0.74      0.09        72

    accuracy                           0.81      5480
   macro avg       0.52      0.77      0.49      5480
weighted avg       0.98      0.81      0.88      5480

Confusion Matrix:
[[4400 1008]
 [  19   53]]
```

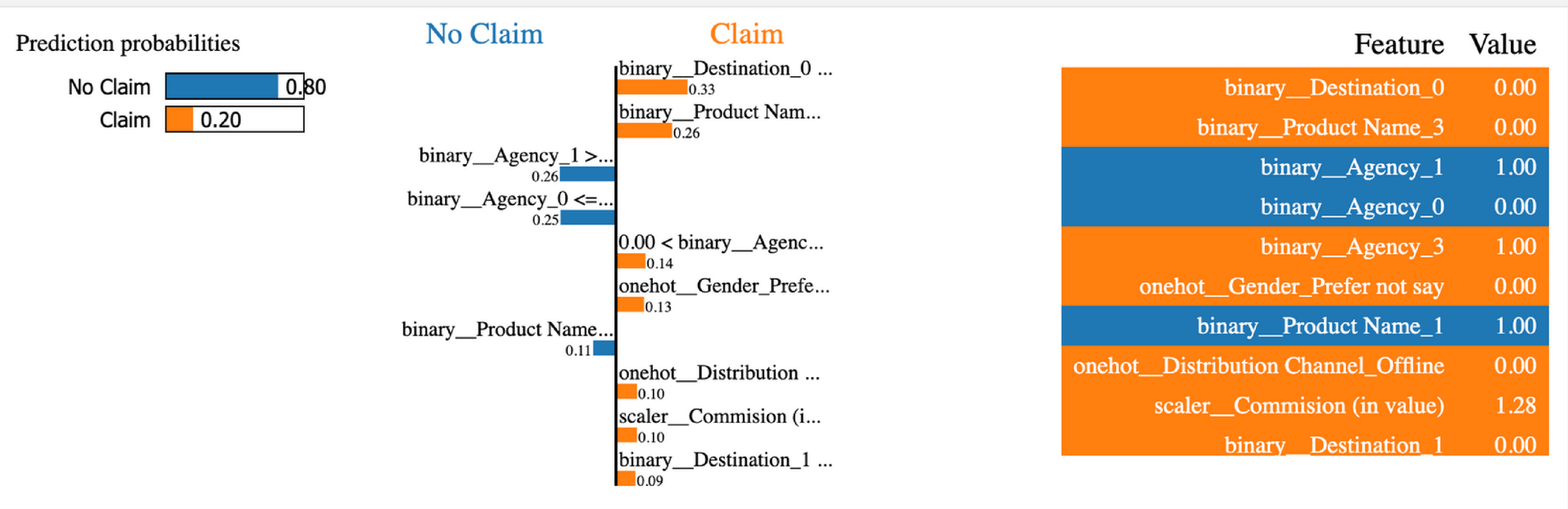


Feature Importance



The model's top features include binary-encoded destination and agency features, indicating predictive claim factors. Gender, distribution channels, commission, and net sales are also significant, reflecting demographic, purchasing mode, and economic influences on claim likelihood.

Explainable ai



The content provides prediction probabilities and feature importance in a visualization, showing how each feature contributes to the prediction decision. Features like binary_Agency_1 and binary_Agency_3 have high influence on predicting "Claim," while binary_Product_Name_1 and scaler_Commission also play significant roles. Other features like onehot_Gender_Prefer not say and onehot_Distribution Channel_Offline have low contributions.

Business Summary

Chosen Model: Logistic Regression with SMOTE

Strength:

- High Recall
- Moderate Precision:

Weaknesses:

- Low Precision for Claims:

Management implication

- Enhanced Risk Assessment:

Accurate claim predictions help in assessing policy risks better, allowing for more tailored risk management strategies

- Improved Resource Allocation:

By predicting which claims are likely, resources can be optimally distributed towards high-risk claims, improving efficiency.

Thank You