PREDICTING TRAVEL INSURANCE CLAIMS: A MACHINE LEARNING APPROACH WITH COST ANALYSIS

Naufal Fajar Revanda JCDS 0808-009

OUTLINE

1 **Background**

Discuss the business context and the importance of accurately predicting travel insurance claims.

2 **Exploratory Data Analysis**

Analyzing historical data to uncover key factors and patterns influencing customer claims.

3 **Modeling**

Building a predictive model to identify high-risk customers, with a focus on mitigating the high cost of a False Negative

4 **Conclusion**

Summarizing the model's performance and the financial impact of our findings on claim prediction.

5 **Recommendation**

Providing actionable steps to optimize premiums, improve customer segmentation, and enhance business efficiency.

EXECUTIVE SUMMARY

This project developed a machine learning solution to predict the likelihood of travel insurance claims. After evaluating several models, **Logistic Regression** was selected as the final approach, achieving the best balance between **recall (0.81)** and **financial cost efficiency**.

The model effectively minimizes high-cost **false negatives** (unpredicted claims) while controlling for overall risk.

By leveraging this model, the company can:

- Implement **risk-based premium pricing**
- Improve **customer segmentation**
- Enhance **financial stability** in a competitive market.

BUSINESS OVERVIEW

Travel insurance is a vital protection for travelers, with premiums based on coverage, trip duration, and destination. Our company seeks to accurately identify which policyholders, specifically from the economic class, are likely to submit an insurance claim.

Using our historical policyholder data, we'll build a predictive model with the following target:

- **1** : Customer files a claim.
- **0** : Customer does not file a claim.

This will allow us to move beyond general risk rules and use data to better manage our business.

PROBLEM STATMENT

Our travel insurance company currently **relies on general rules** to assess customer risk (e.g., travel duration, destination). This approach is inaccurate and leads to two major issues:

1. **Financial Loss**: We risk setting premiums too low for high-risk customers, leading to unanticipated claim costs.
2. **Lost Opportunities**: We may set premiums too high for low-risk customers, causing them to choose a competitor instead.

To address this, we need a **data-driven solution**. Our goal is to use a machine learning model to accurately identify high-risk customers, enabling us to set fair premiums, **ensure the company's financial health**, and pay all legitimate claims smoothly.

GOALS

- **Build a predictive model** to accurately identify policyholders who are likely to file a claim.
- **Identify key factors** influencing a claim to improve our understanding of customer risk patterns.
- **Support business decisions** with data-driven insights for more accurate premium determination and risk management.
- **Increase operational efficiency** by reducing unexpected claim payouts and maximizing profitability.

BUSINESS METRICS

	Actual: Claim	Actual: No Claim
Predicted: Claim	**True Positive (TP)** Correctly predicts a claim. Premium is adjusted higher, so loss is controlled Company can manage risk effectively.	**False Positive (FP)** Predicted a claim but no claim occurs. Company must set aside $65,000 as standby funds Opportunity cost = $3,250/year.
Predicted: No Claim	**False Negative (FN)** Predicted no claim but a claim occurs. Biggest risk: company bears an unanticipated cost. Cost up to $65,000 per claim. FN is ~20x more costly than FP.	**True Negative (TN)** Correctly predicts no claim. Standard premium applies. Stable and profitable outcome.

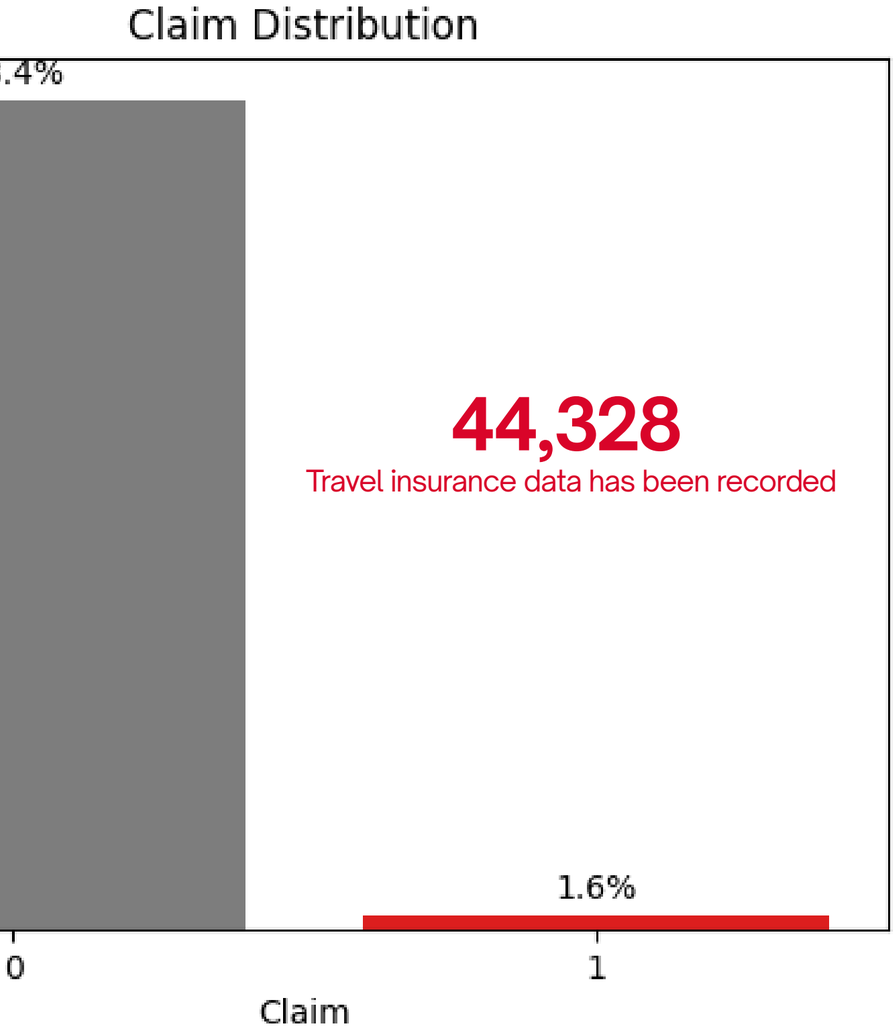
DATA UNDERSTANDING

Attributes	Description	Categorize
Gender	Gender of insured	Policyholder information
Age	Age of insured	
Product Name	Name of the travel insurance products	
Claim	Claim status	
Agency	Name of agency	Sales & agency information
Agency Type	Type of travel insurance agencies	
Distribution Channel	Channel of travel insurance agencies	
Net Sales	Amount of sales of travel insurance policies	
Commission (in value)	Commission received for travel insurance agency	
Duration	Duration of travel	Trip information
Destination	Destination of travel	

Each row in dataset represent unique travel insurance

CLAIM DISTRIBUTION

Only 1.6% of 44,328 file their claim, good news or bad news?



44,328

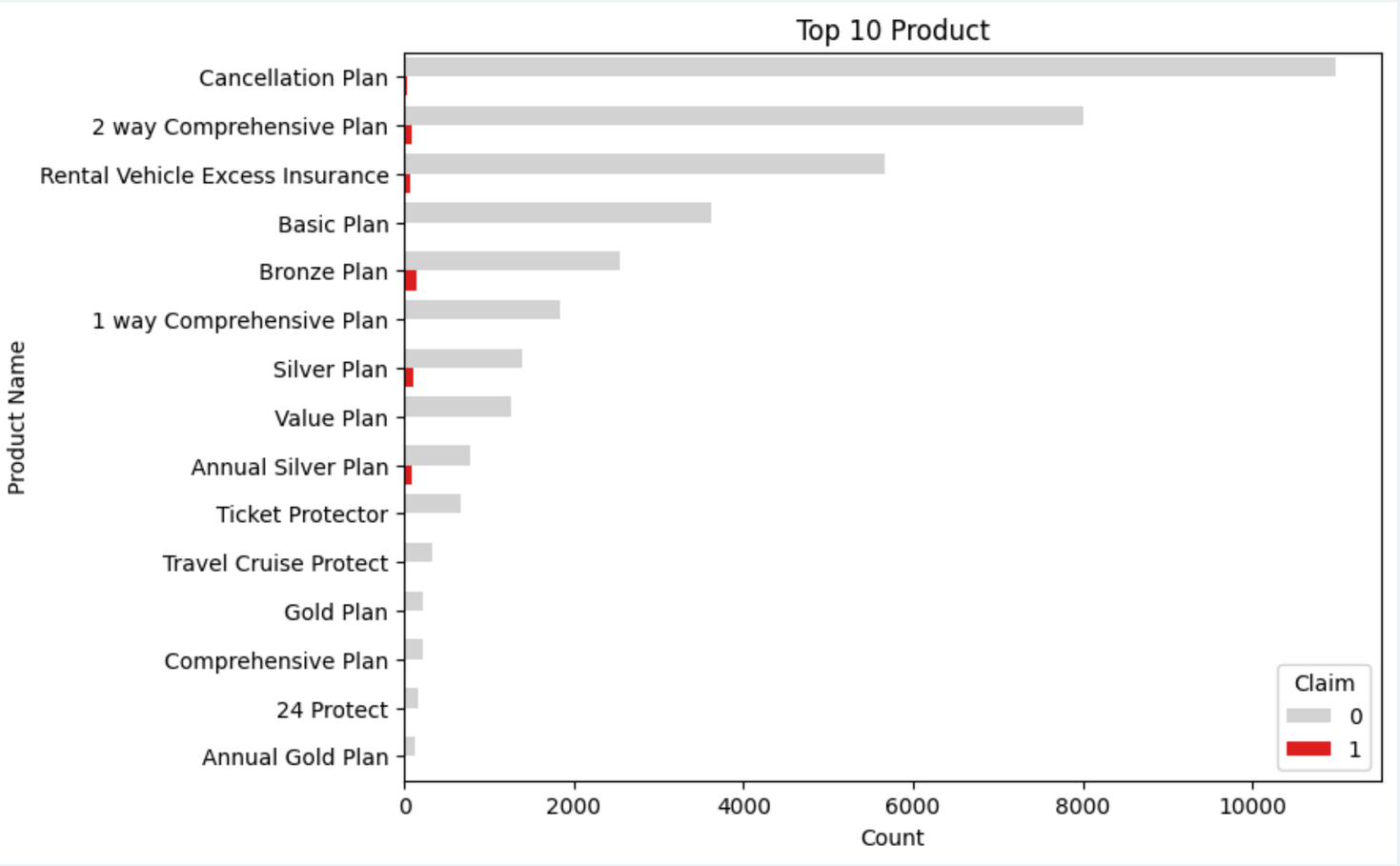
Travel insurance data has been recorded

Critical Insight

The claim proportion is severly imbalance for modeling

BEST-SELLING PLANS, BUT WHICH ONES DRIVE CLAIMS?

despite the fact that only 146 people filed the claim, the Bronze Plan stands out as the most claimed product.



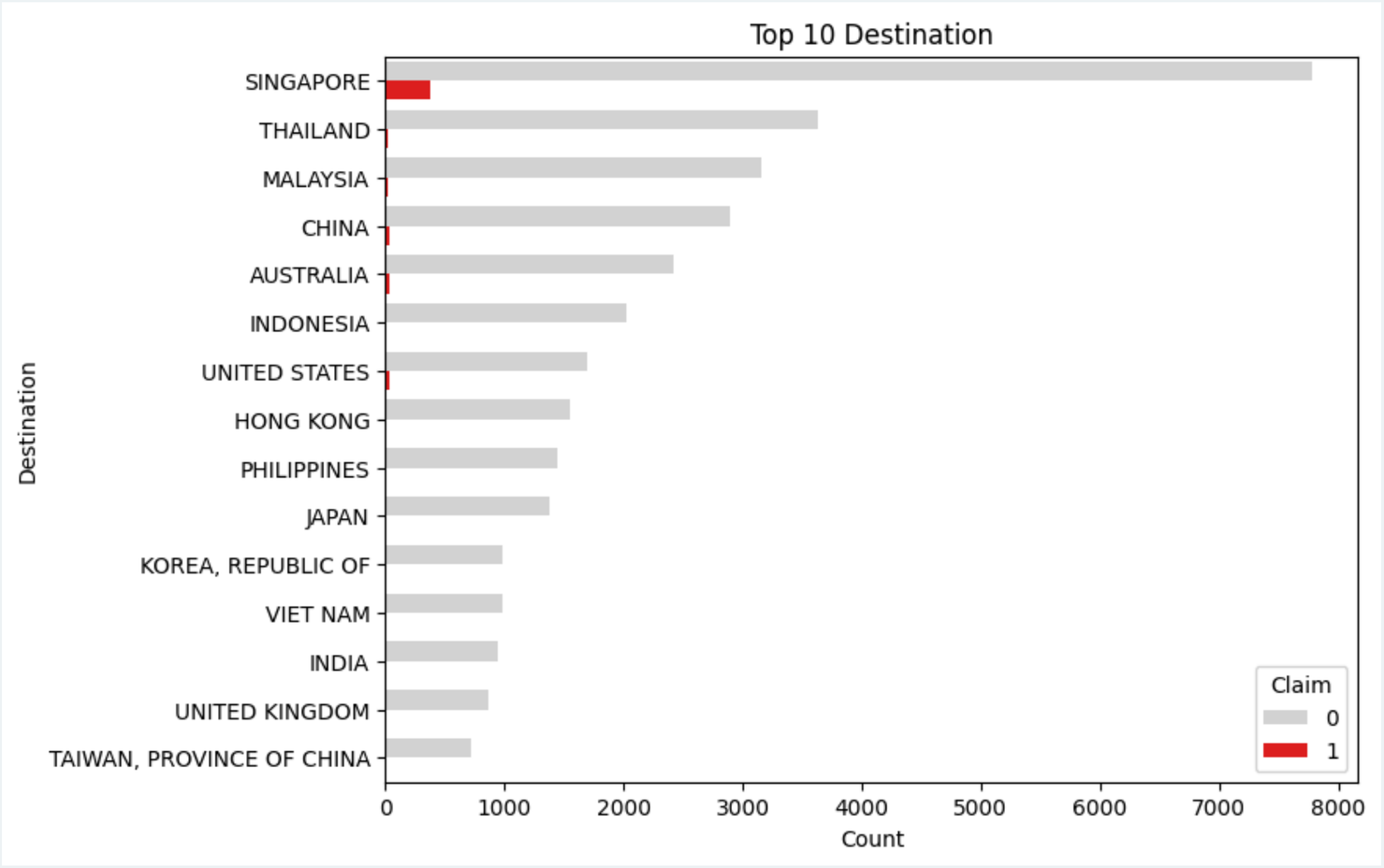
Critical Insight

High sales don’t always mean high claim risk

Annual Silver & Annual Gold Plans have the highest claim rates (⩾10%

DESTINATION SINGAPORE: TRAVEL LEADER, CLAIM LEADER

High travel volumes make Singapore the top destination but also the top source of claims



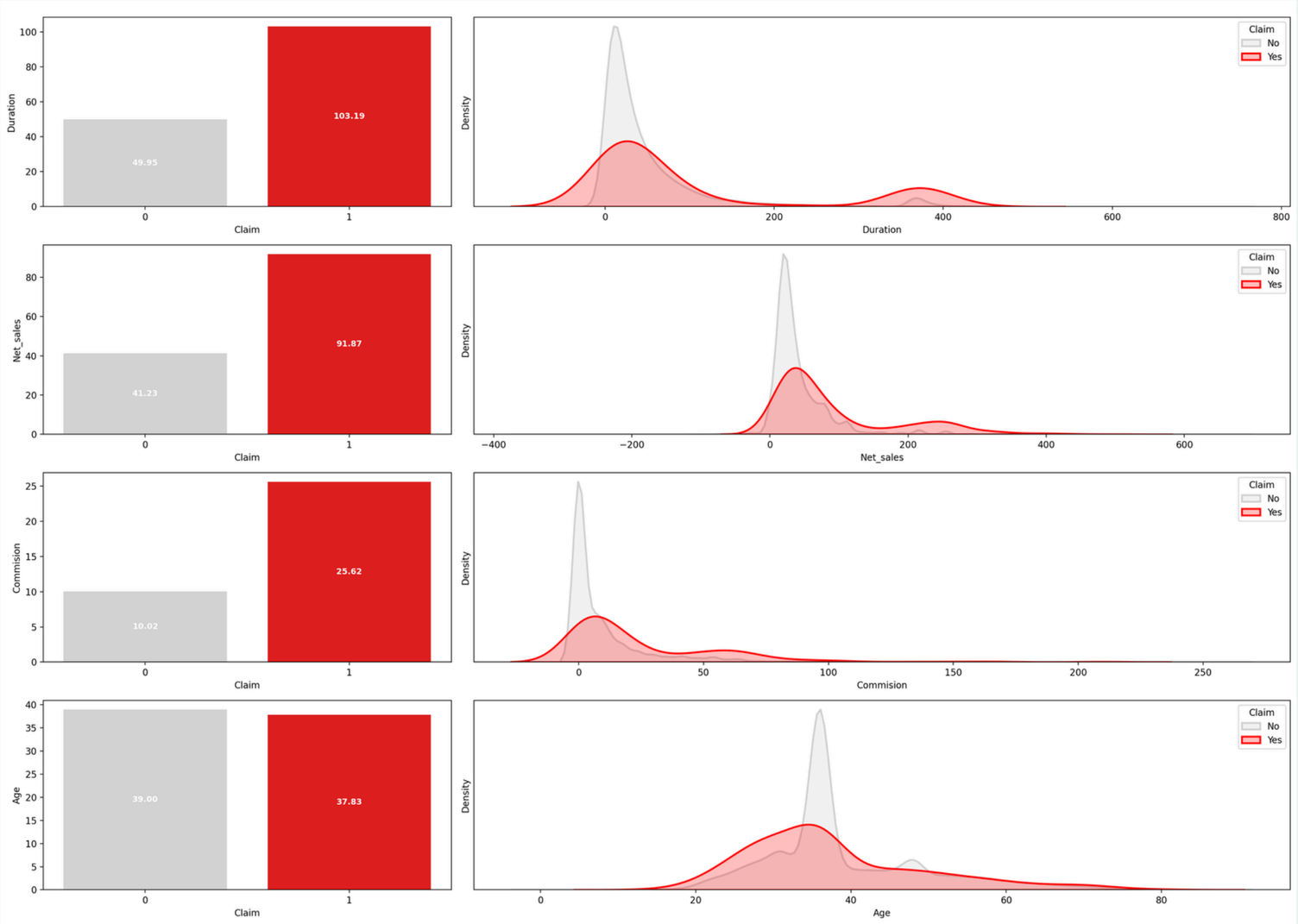
Critical Insight

Singapore stands out as the destination with the highest absolute number of claims (379), far exceeding other countries.

Other destinations like the United States (1.7% claim rate) and China (1.0%) show higher relative risk compared to Thailand and Malaysia (<1%), indicating claim patterns vary by destination.

CLAIMS FOLLOW THE MONEY, NOT THE AGE

Sales-related features stand out as key risk indicators, not demographics



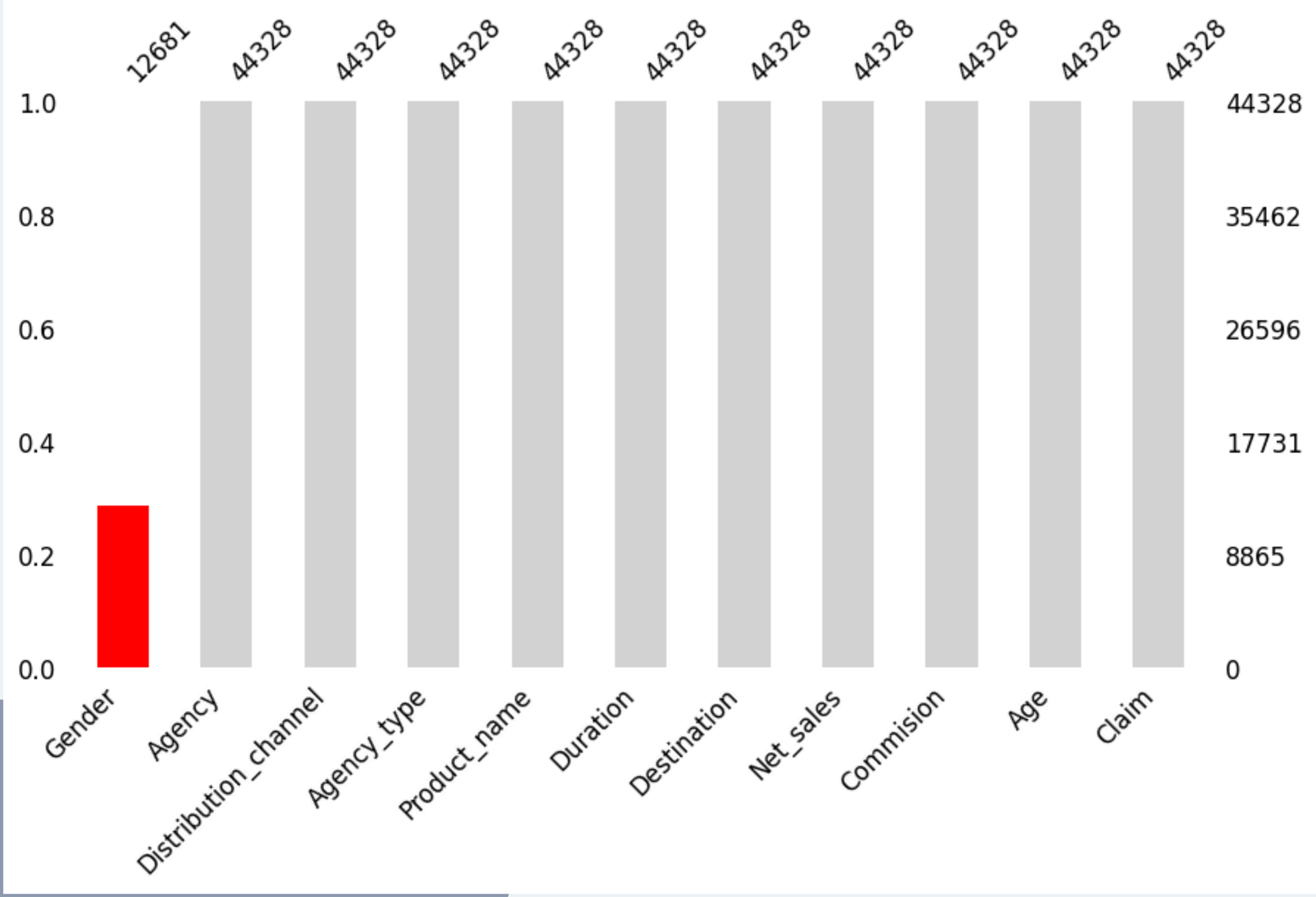
Critical Insight

Claims are linked to higher duration, sales, and commission values.

Age shows little difference between claim and non-claim groups.

DATA CLEANING: MISSING VALUE

Only Gender shows missing values. Other features are fully available



Critical Insight

70% of Age colums is missing

Age values above 117 are unrealistic, since the verified human lifespan record is 117 years and Europe's average life expectancy is ~80. (Guiness World Records, 2019))

Such values likely indicate data entry errors or anomalies rather than valid observations

Moreover, taking ethical considerations into account, the gender feature can be dropped

DATA CLEANING: HANDLING DUPLICATE

Identical feature, different target can make the model confused

10%

of total rows are duplicate

→ Drop all sets of rows that **have identical features but conflicting target**

→ drop any remaining rows that are perfect duplicates

44234 rows → 39579 rows

FEATURE TRANSFORMATION

Scaler

Robust Scaller

was used on numerical features to normalize the data while reducing the influence of extreme outliers. This ensures that numerical variables remain comparable without being distorted by skewed distributions

Encoder

One-hot Encoding

was applied to low-cardinality categorical features (Agency_type, Distribution_channel) to create clear binary indicators for each category. This method avoids introducing bias and is computationally efficient when categories are limited

Binary Encoding

Binary encoding was chosen for high-cardinality features (Agency, Product_name, Destination) to reduce dimensionality. Unlike one-hot, it transforms categories into binary codes, balancing interpretability with efficiency

MODEL SELECTION

Using default config from each alogrithm

Model	F2 Mean	F2 Std	Recall Mean	Recall Std	Precision Mean	Precision Std
Decision Tree Classifier	0.0669	0.0394	0.0717	0.043	0.0533	0.0308
Random Forest Classifier	0.0215	0.0166	0.0179	0.0139	0.1129	0.0896
KNeighbors Classifier	0.0049	0.0098	0.004	0.008	0.1167	0.2986
Gradient Boosting Classifier	0.005	0.0099	0.004	0.008	0.1	0.2
LGBM Classifier	0.0049	0.0098	0.004	0.0079	0.1333	0.3055
XGB Classifier	0.0024	0.0072	0.002	0.0059	0.025	0.075
Logistic Regression	0	0	0	0	0	0

MODEL SELECTION

Using class_weight parameter for each model

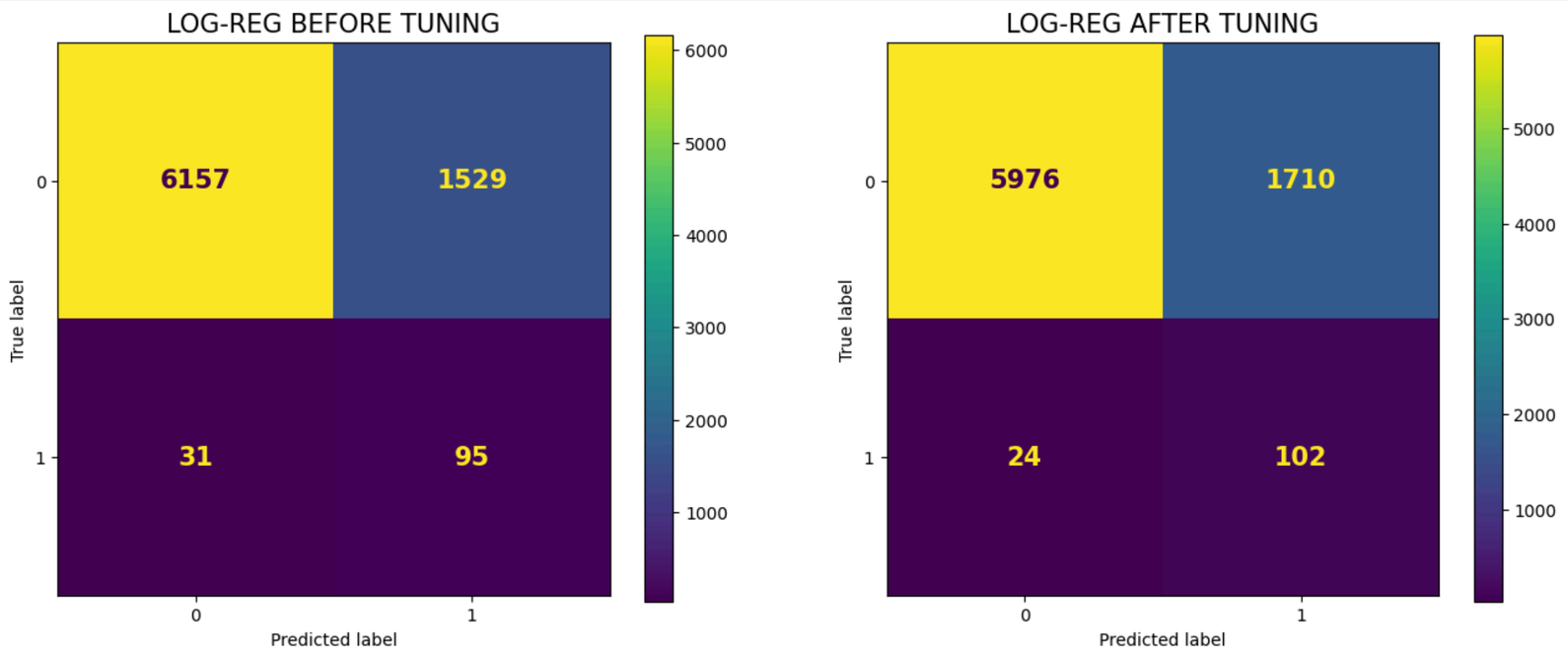
Model	F2 Mean	F2 Std	Recall Mean	Recall Std	Precision Mean	Precision Std
Logistic Regression	0.2019	0.0135	0.6813	0.0339	0.053	0.0041
LGBM Classifier	0.211	0.0124	0.4861	0.0295	0.0647	0.0045
XGB Classifier	0.1774	0.0321	0.273	0.0534	0.0739	0.0123
Decision Tree Classifier	0.0692	0.0163	0.0697	0.0165	0.0672	0.0156
Random Forest Classifier	0.0096	0.009	0.008	0.0074	0.0617	0.0577

MODEL SELECTION

Using resampling method for each model

Model	Resampler	F2 Mean	F2 Std	Recall Mean	Recall Std	Precision Mean	Precision Std
LGBM Classifier	RandomUnderSampler	0.1614	0.0064	0.7112	0.0269	0.0395	0.0017
Random Forest Classifier	RandomUnderSampler	0.1696	0.0117	0.7052	0.0393	0.042	0.0032
XGB Classifier	RandomUnderSampler	0.1543	0.0059	0.7013	0.0324	0.0374	0.0014
Logistic Regression	RandomUnderSampler	0.1915	0.0188	0.6952	0.0273	0.0493	0.0059

MODEL EVALUATION : LOGISTIC REGRESSION

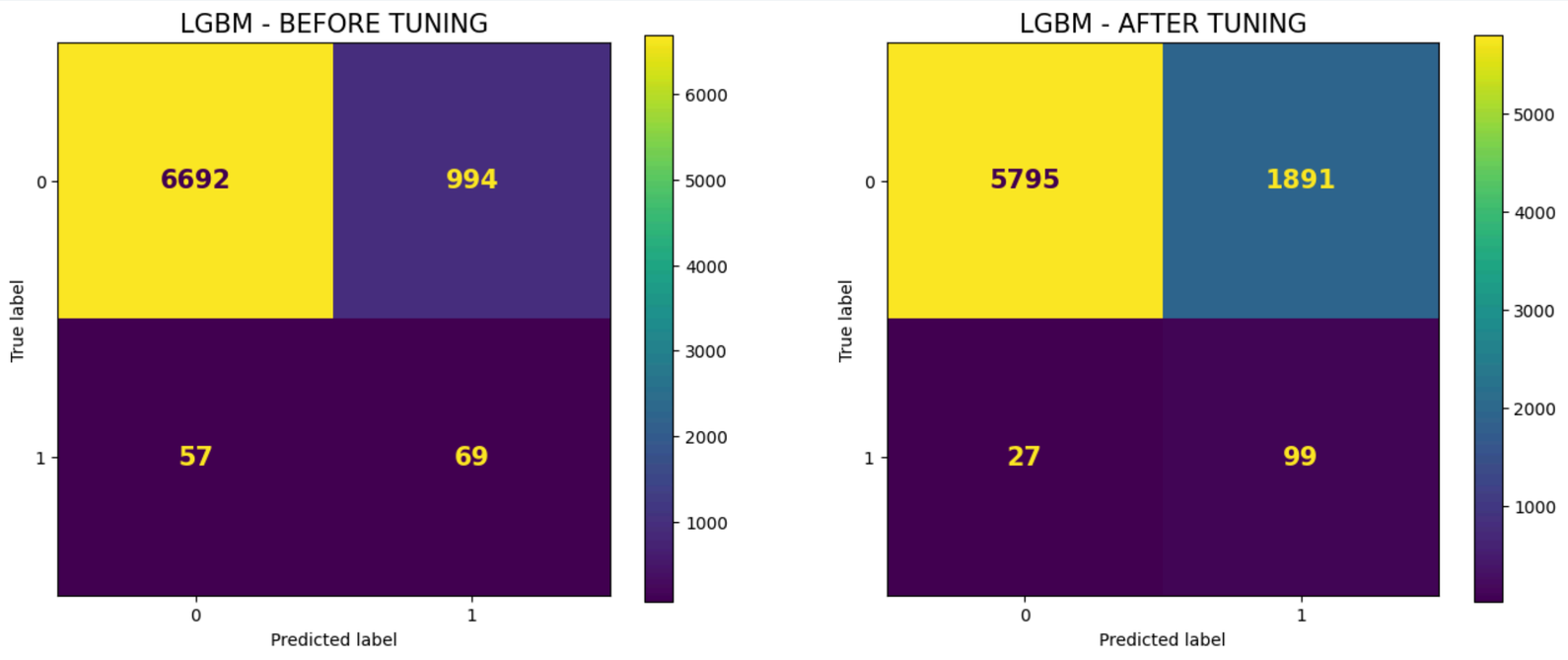


- **Before tuning**: Total cost = **6,984,250**

- **After tuning**: Total cost = **7,117,500**

- Effect: Recall improved (**Train: 0.71, Test: 0.81**), but cost slightly increased due to more False Positives.

MODEL EVALUATION : LIGHT GBM

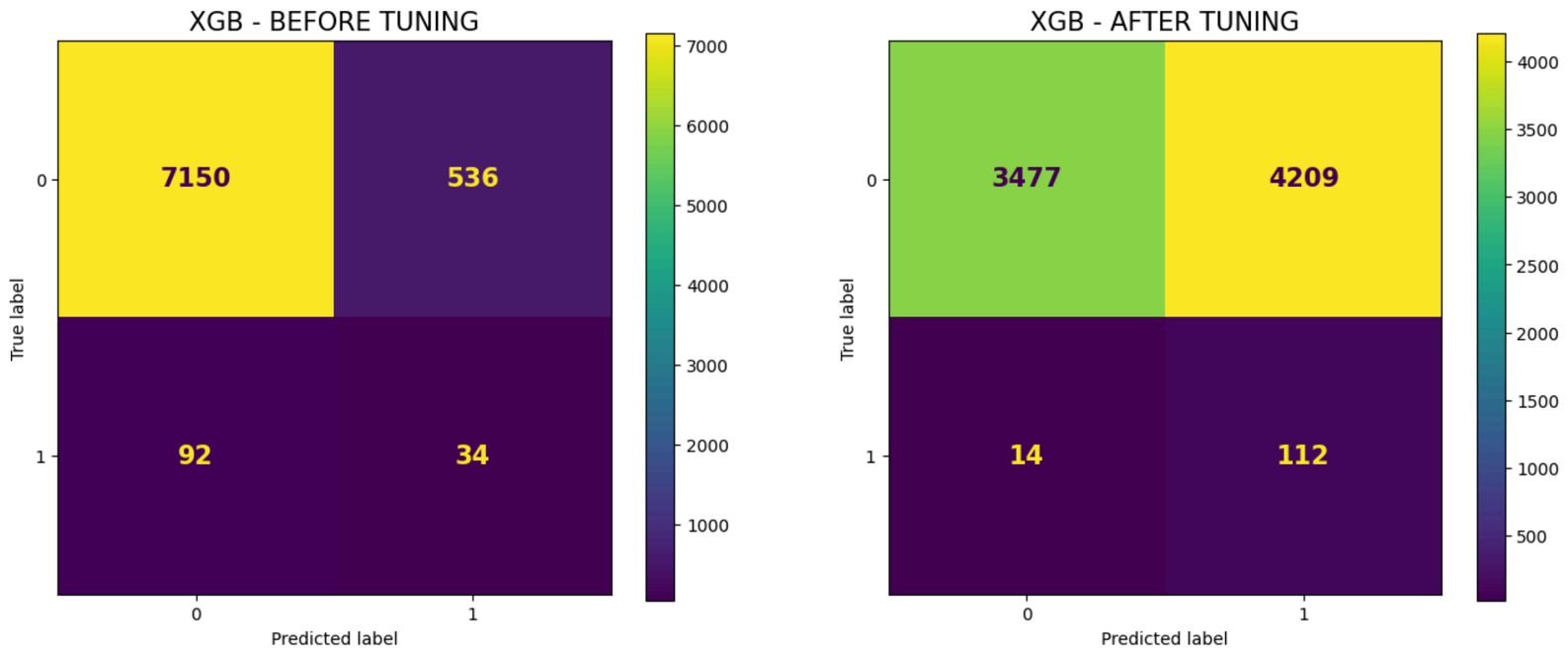


- **Before tuning**: Total cost = **4,010,500**

- **After tuning**: Total cost = **7,900,750**

- Effect: Recall improved (**Train: 0.80, Test: 0.79**), False Negatives reduced, but False Positives increased sharply -> much higher total cost.

MODEL EVALUATION : XGBOOST



- **Before tuning**: Total cost = **7,722,000**

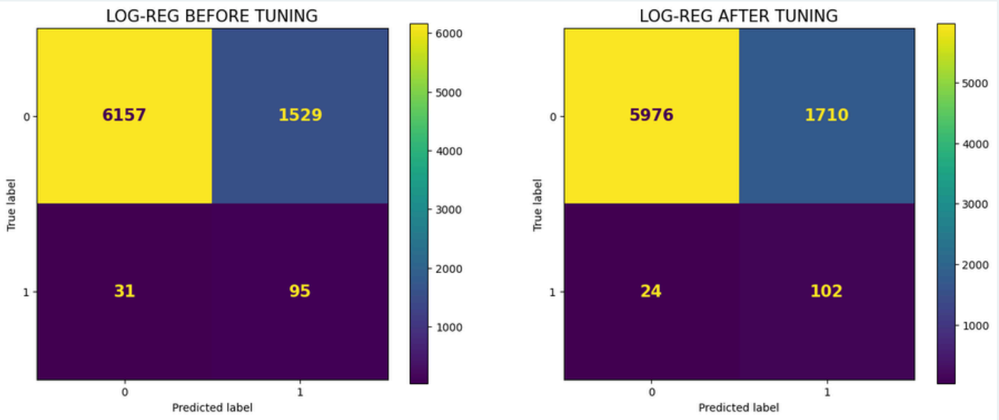
- **After tuning**: Total cost = **14,589,250**

- Effect: Recall significantly improved (**Train: 0.91, Test: 0.89**), False Negatives minimized, but False Positives exploded, leading to the highest cost overall.

MODEL EVALUATION : THE BEST MODEL FOR THE CASE

Logistic Regression

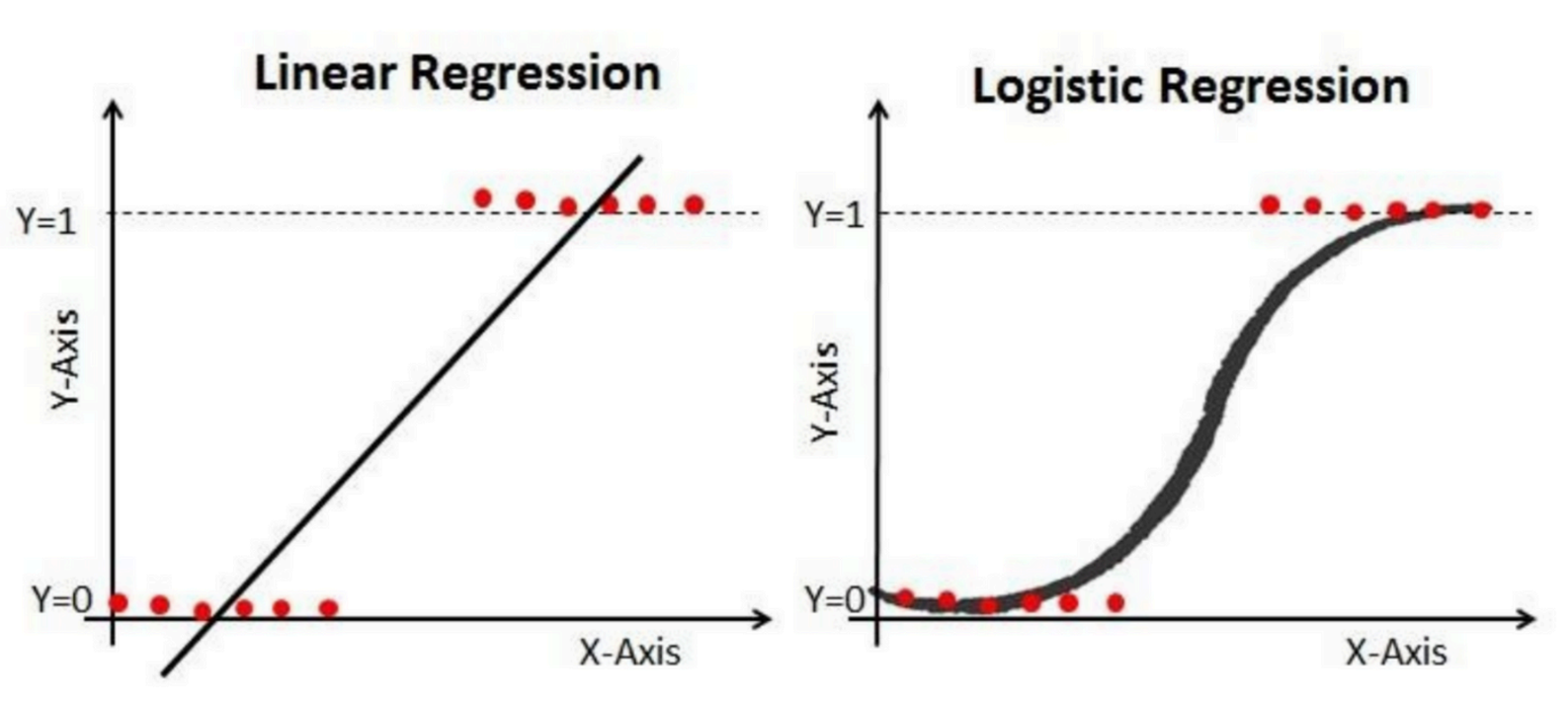
Why?



- **Lowest overall cost** while keeping solid recall → most cost-efficient for extreme imbalance (98.4% no-claim vs 1.6% claim).
- **Better balance**: detects claims without excessive false positives (unlike XGBoost & LightGBM)
- **Controls operational expense**s by avoiding overcompensation for the minority class.
- **Simplicity & interpretability** -> critical for insurance: transparency, explainability, and regulatory compliance.

THE WORKING MECHANISM OF LOGISTIC REGRESSION

Logistic Regression is a **classification** method that models the relationship between a set of **independent variables** (features) and a binary **dependent variable** (e.g., claim vs. no claim). Unlike Linear Regression, which produces continuous outputs, Logistic Regression applies the logistic (**sigmoid**) function to map the linear combination of features into probabilities between 0 and 1.



$$P(y=1|X) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}$$

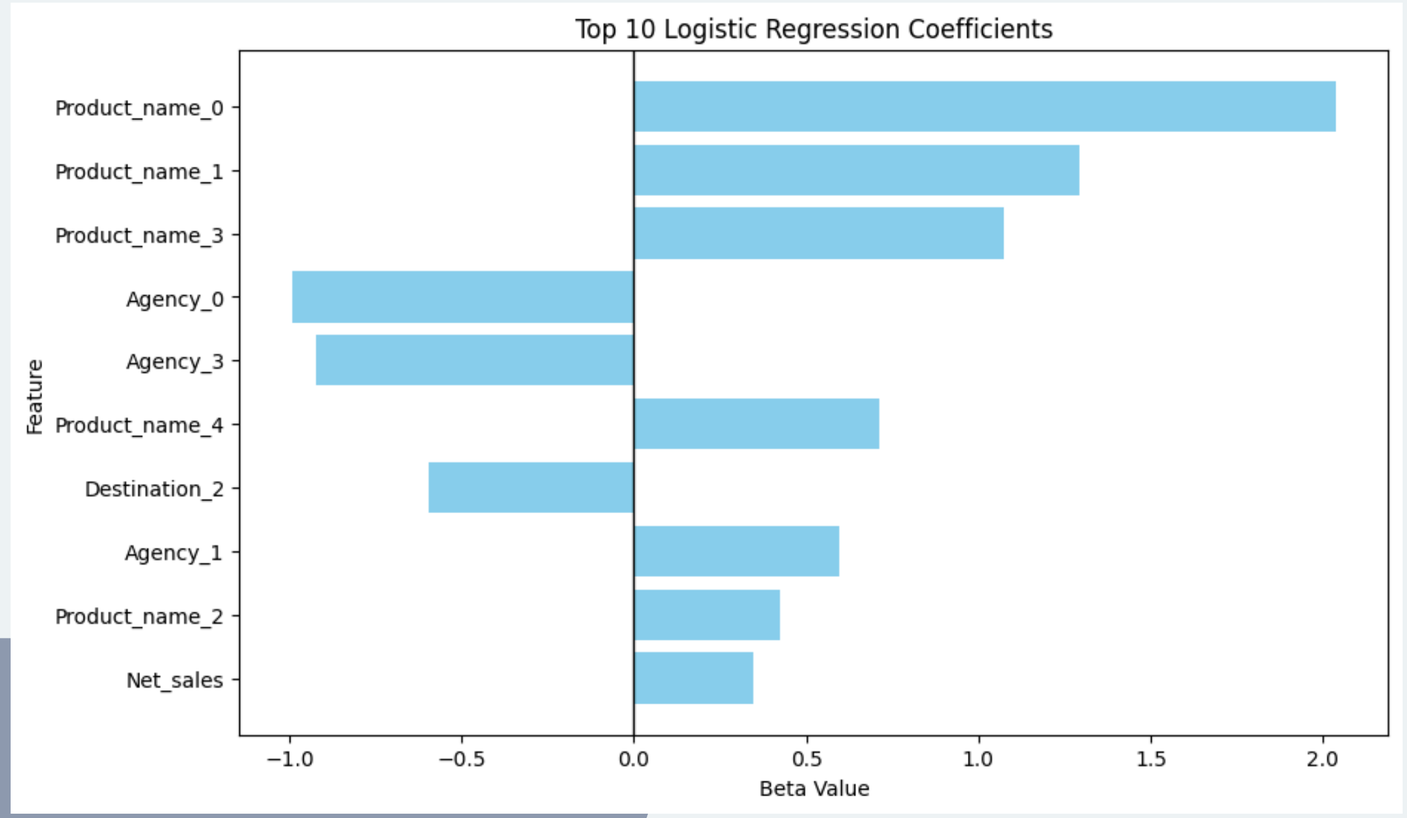
EVALUATION

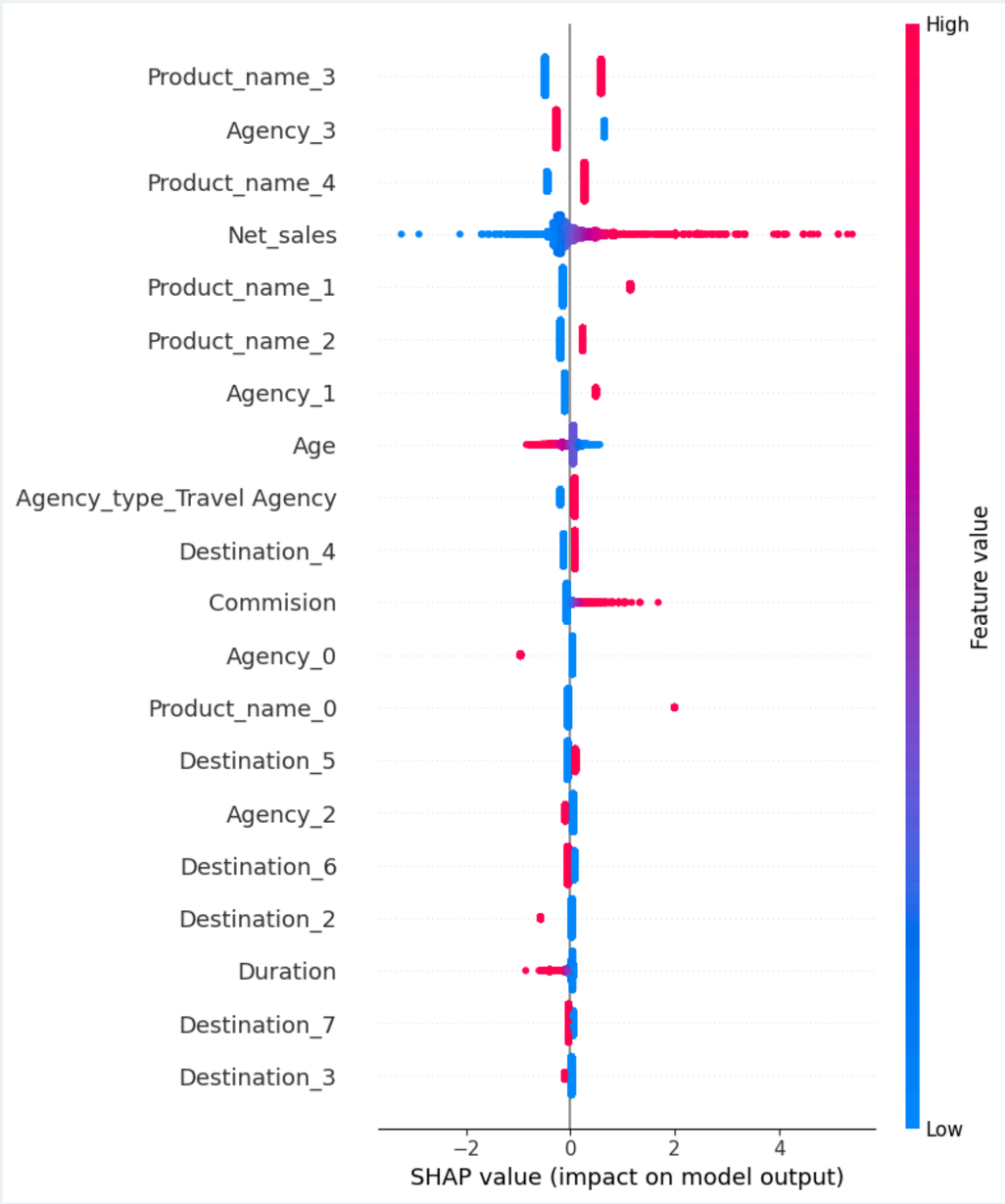
Beta score

- The model is heavily influenced by specific product names (e.g., `Product_name_0`, `Product_name_1`) and certain agency categories*(`Agency_0`, `Agency_3`).
- `Product_name_0` stands out as the single most impactful feature.

Shap

The model is strongly influenced by agency-related variables and Net_sales, with product categories also playing a significant role. Lower-ranked features have relatively minor contributions.





CONCLUSION

- Built a machine learning model to predict travel insurance claims.
- Logistic Regression chosen as final model-> best trade-off between recall (0.81) and cost efficiency.
- Ensemble models had higher recall but too many false positives → drove costs up.
- Cost analysis: minimizing False Negatives (missed claims) is more critical than reducing False Positives.

Implement Risk-Based Premium Pricing

- High-Risk Customers: Apply an adjusted, higher premium to cover the anticipated risk and maintain profitability.
- Low-Risk Customers: Offer more competitive pricing to attract a larger customer base and increase market share.

Enhance Customer Segmentation

Use the model's output to segment policyholders by their predicted claim probability. This enables more targeted strategies for growth and risk management.

Optimize Operational Risk Management

Align the company's financial reserves with the model's predictions. By proactively identifying high-risk policies, the company can allocate standby funds more accurately and reduce financial shocks from unexpected claims.

For Business Strategy

RECOMMENDATIONS

RECOMMENDATIONS

For the Model Lifecycle

Continuous Monitoring and Maintenance

Continuously track the model's performance in production, focusing on Recall, False Negative counts, and the overall business cost metric. The model should be periodically retrained with new data to adapt to evolving customer behaviors and market trends.

Future Enhancements

- Explore advanced techniques such as cost-sensitive learning to optimize the model directly for financial outcomes.
- Enrich the dataset with new features (e.g., trip purpose, travel frequency) to potentially improve predictive power.
- Experiment with hybrid models (e.g., stacking Logistic Regression with ensemble methods) to further refine risk calibration.

Several ways to validate real claims:

1. Documentation & evidence: by verifying tickets, visa, hospital/police reports, payment receipt

2. Cross check: cross check with airlines, immigrations, or hospitals

3. Chek their behavior: review customer's claim history and any suspicious behavior

4. Direct investigation: high-risk claims that tend to be fraud should be investigated by professional investigators

For the Claim Validation Process

Note: Any enhanced validation process must be conducted with high regard for customer privacy and regulatory compliance.

RECOMMENDATIONS

THANK YOU