Purwadhika
Digital Technology School

Travel Insurance – Claim Prediction

Capstone project presentation – Machine Learning

Created by **Harish Muhammad**

JCDS - 0308

Overview

Content of this presentation

Business Problem Understanding

Context

Travel insurance

policybazaar.com/

Financial protection

Business Problem Understanding

Problem Statement

- Maintaining sufficient balance between customer premium with claims and expenses is the crucial issue.
- The ability to predict the probability off customers who are claim and not claim will help enterprise in managing risks.

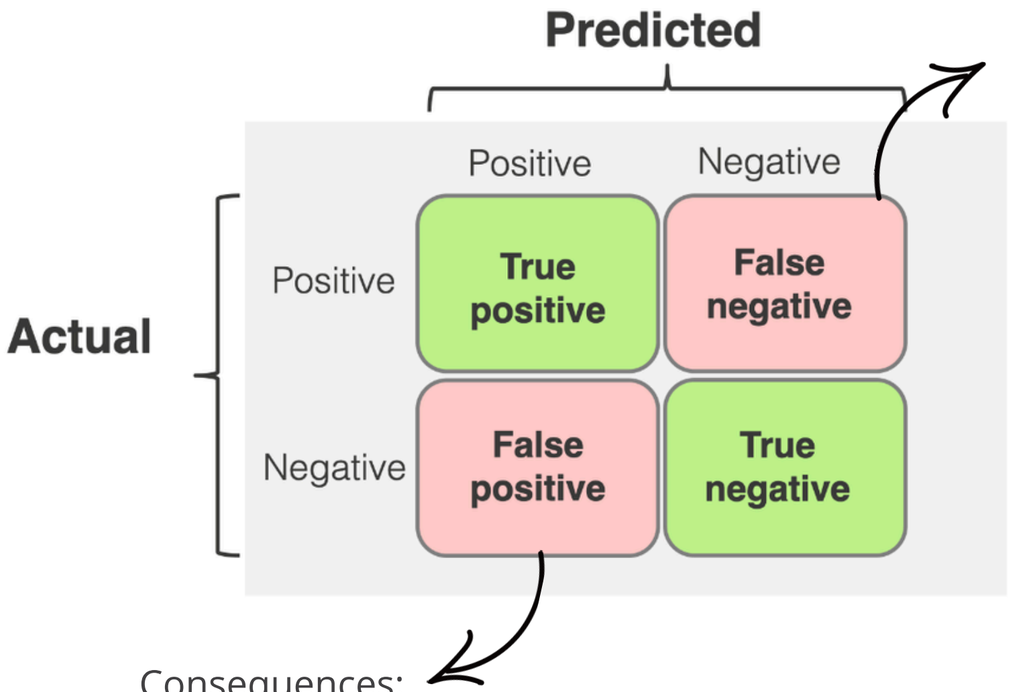
Goals

- Predictive modelling that can distinguish which customers more likely to claim and not claim.
- Which factors or features that contribute customers to claim.

Analytical Approach

- Analyzing data to learn about pattern that can differentiate customers who will claim the insurance and who will not.
- Building classification models.

Business Problem Understanding



Consequences:
insurance claim, processing and claim inspection cost (USD 200 + USD 1241)

Consequences:
processing and claim inspection cost (USD 200)

Metric Evaluation

- Highly imbalanced dataset: 98.5% non-claim & 1.5% claim
- Only accuracy --> not accurate
- ROC-AUC score --> more comprehensive evaluation

Data Understanding

Attribute	Data Type	Description
Agency	Object	Name of agency or insurance providers
Agency Type	Object	Type of travel insurance agencies
Distribution Channel	Object	Channel of travel insurance agencies
Product Name	Object	Name of the travel insurance products
Gender	Object	The gender of insured
Duration	Integer	Duration of travel
Destination	Object	Destination of travel
Net Sales	Float	Amount of sales of travel insurance policies
Commission (in value)	Float	Commission received for travel insurance agency
Age	Integer	Age of insured customers
Claim	Object	Claim status

Total data : 44328
Columns: 11

Data distribution:
Not normally distributed

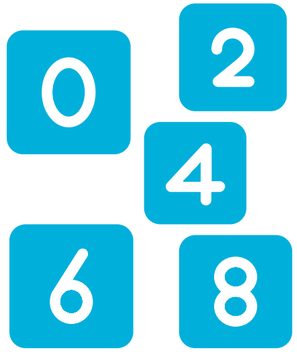
Data Cleaning

- Handling duplicates
- Handling missing values
- Removing irrelevant features
- Handling anomalies
- Handling outliers

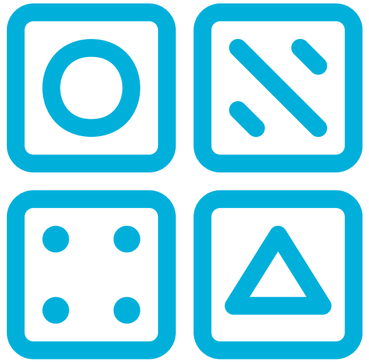
Duplicated data: 11.4%
Missing values in one column: 70.45%

Exploratory Data Analysis (EDA)

Numerical vs Target



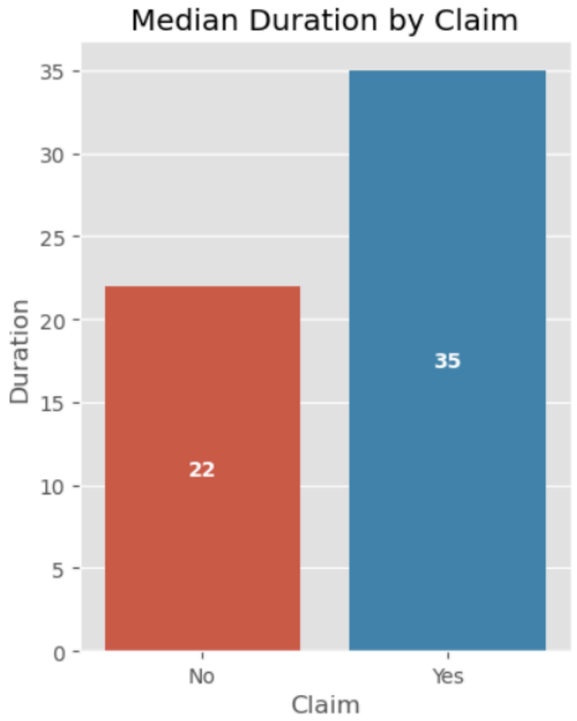
Categorical vs Target

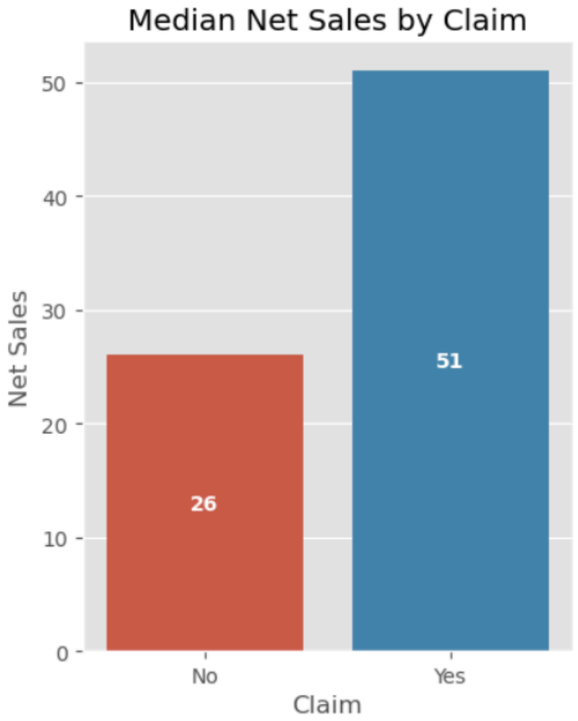


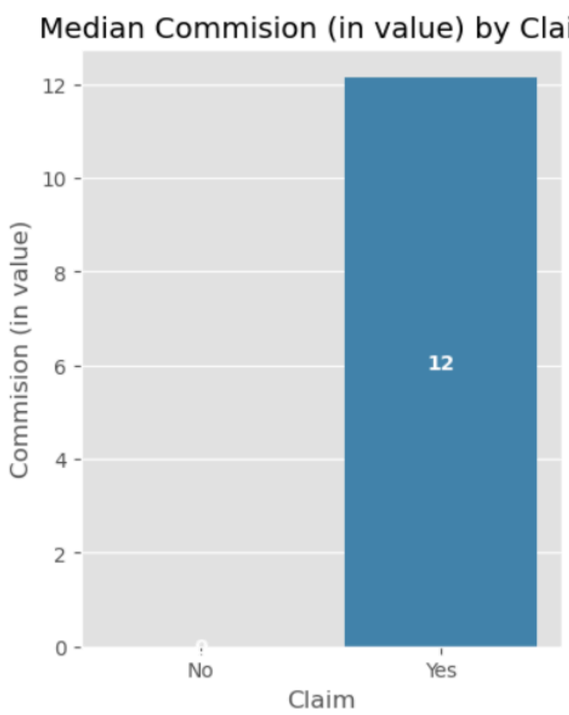
Numerical vs Target

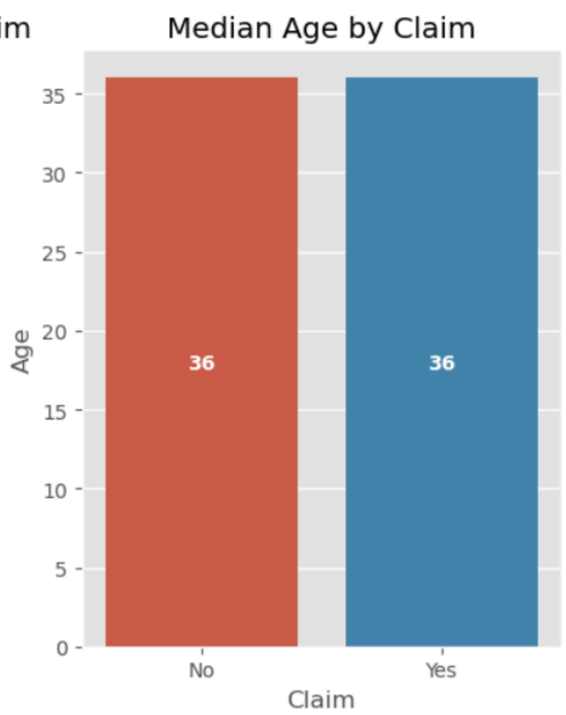
Data Distribution

Feature	D'Agostino-Pearson Statistic	P-value	Distributed
Duration	106362.302085	0.0	Not Normally Distributed
Net Sales	32695.931743	0.0	Not Normally Distributed
Commision (in value)	37939.755693	0.0	Not Normally Distributed
Age	28763.309947	0.0	Not Normally Distributed

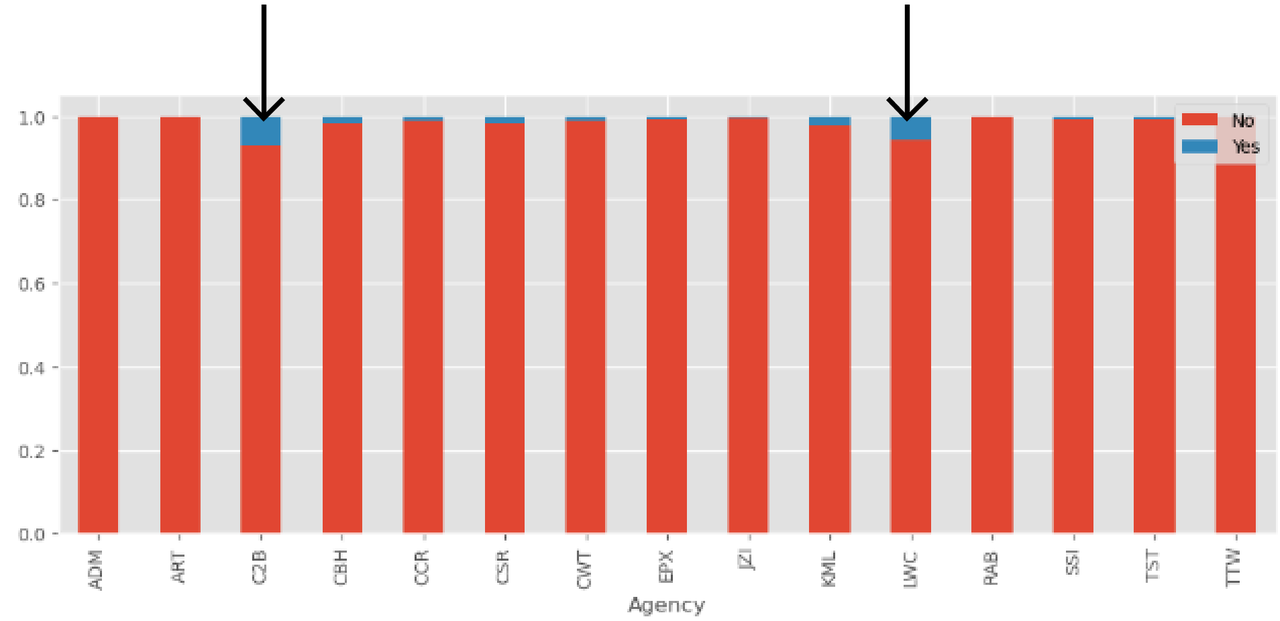


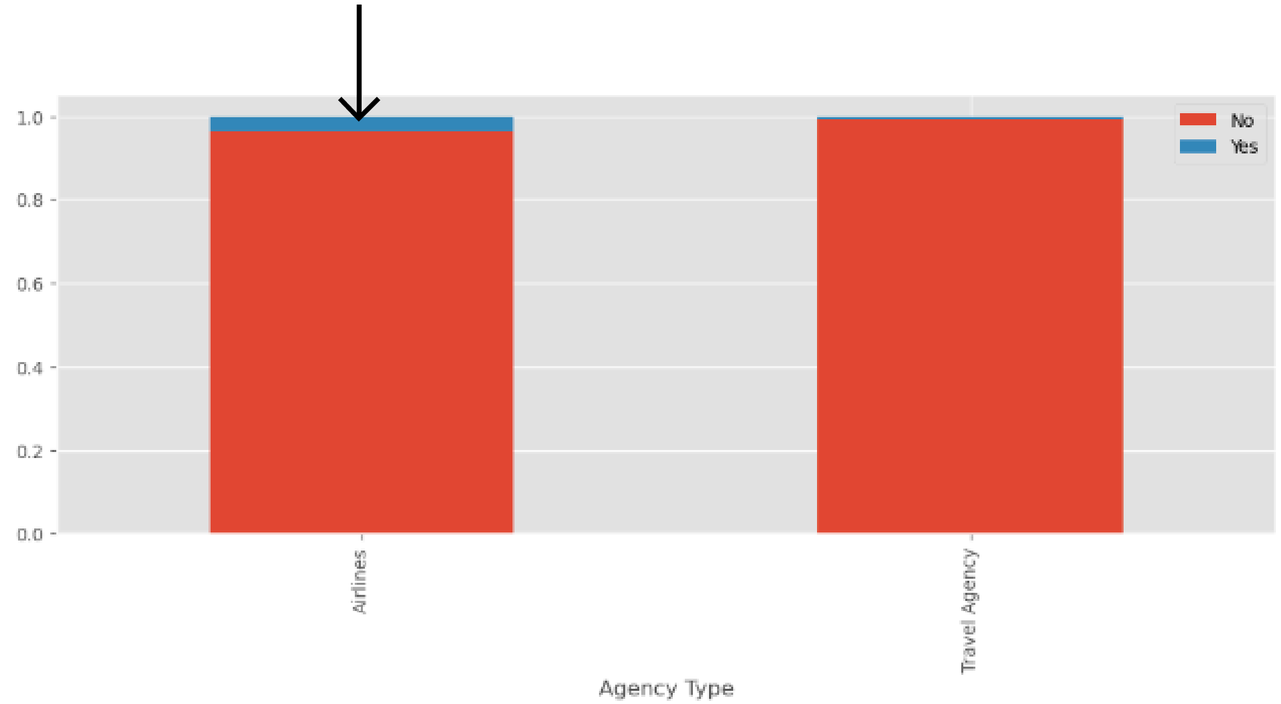




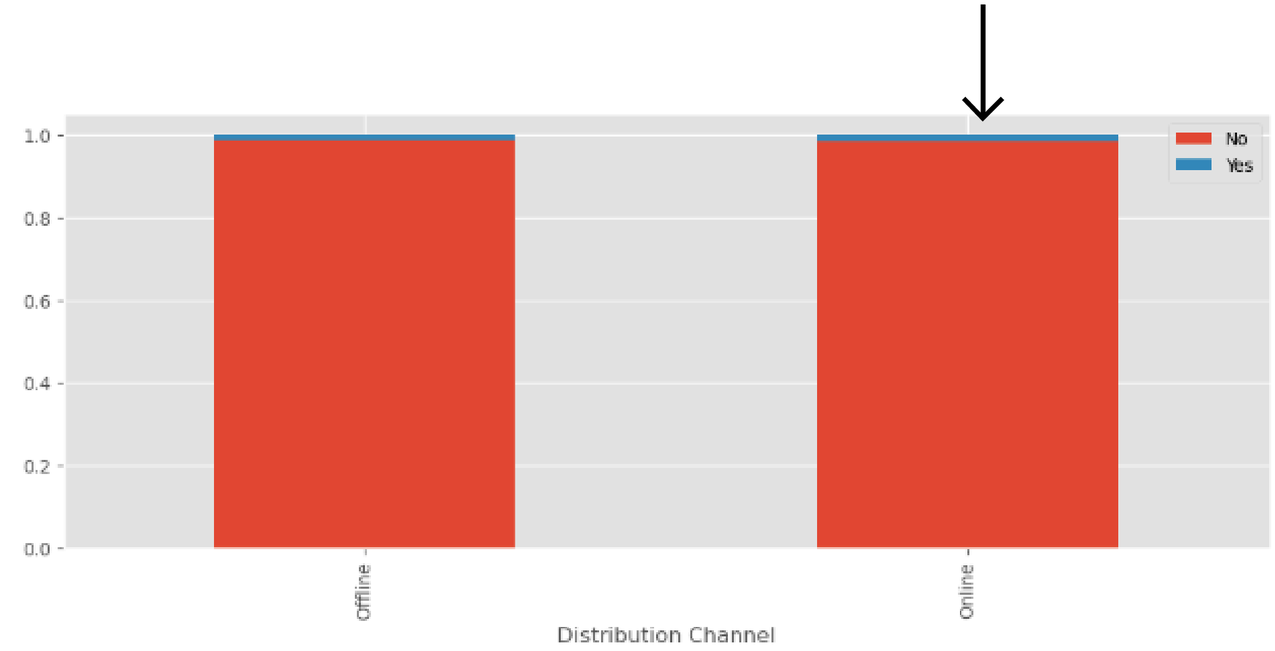


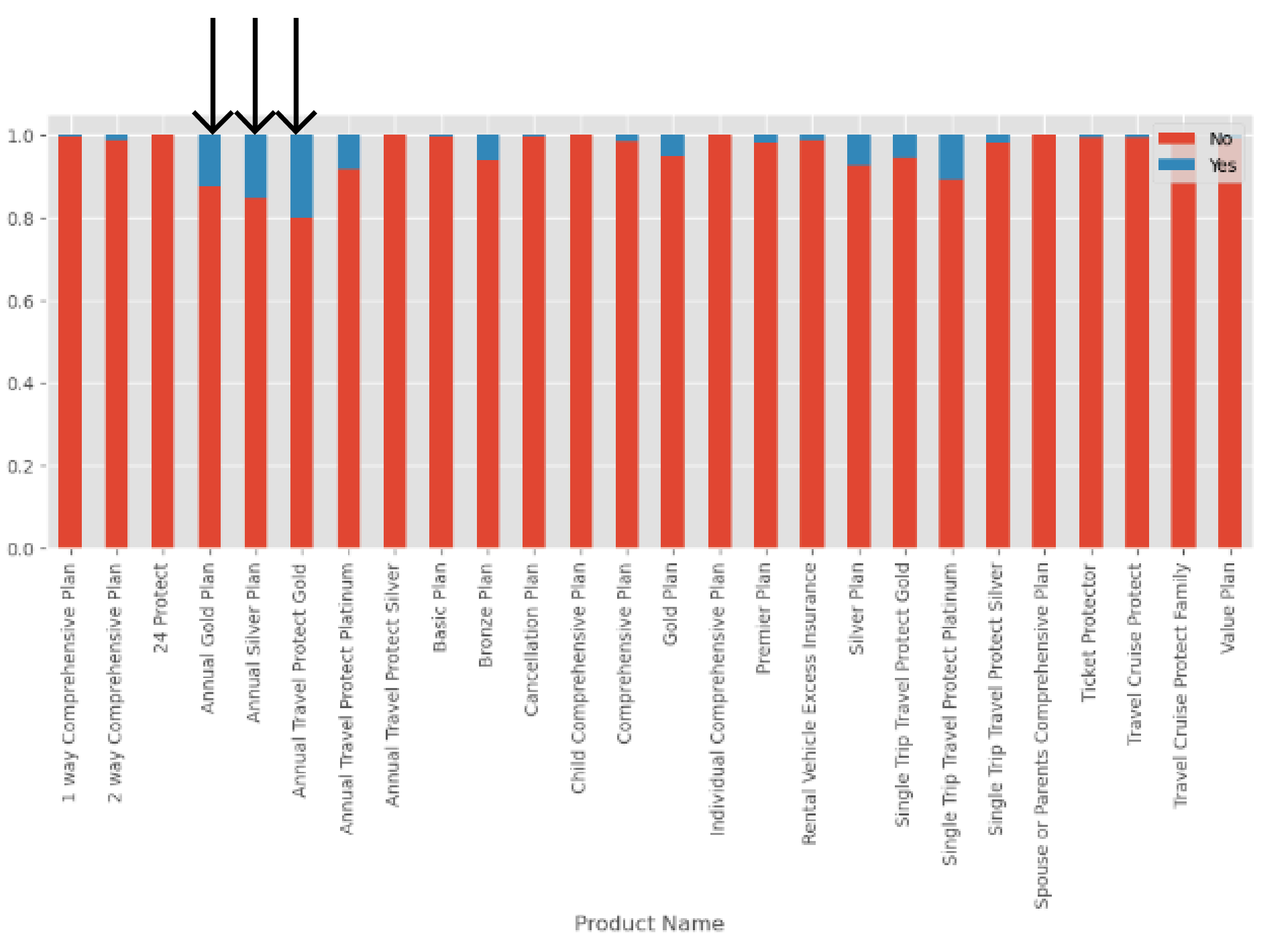
Categorical vs Target



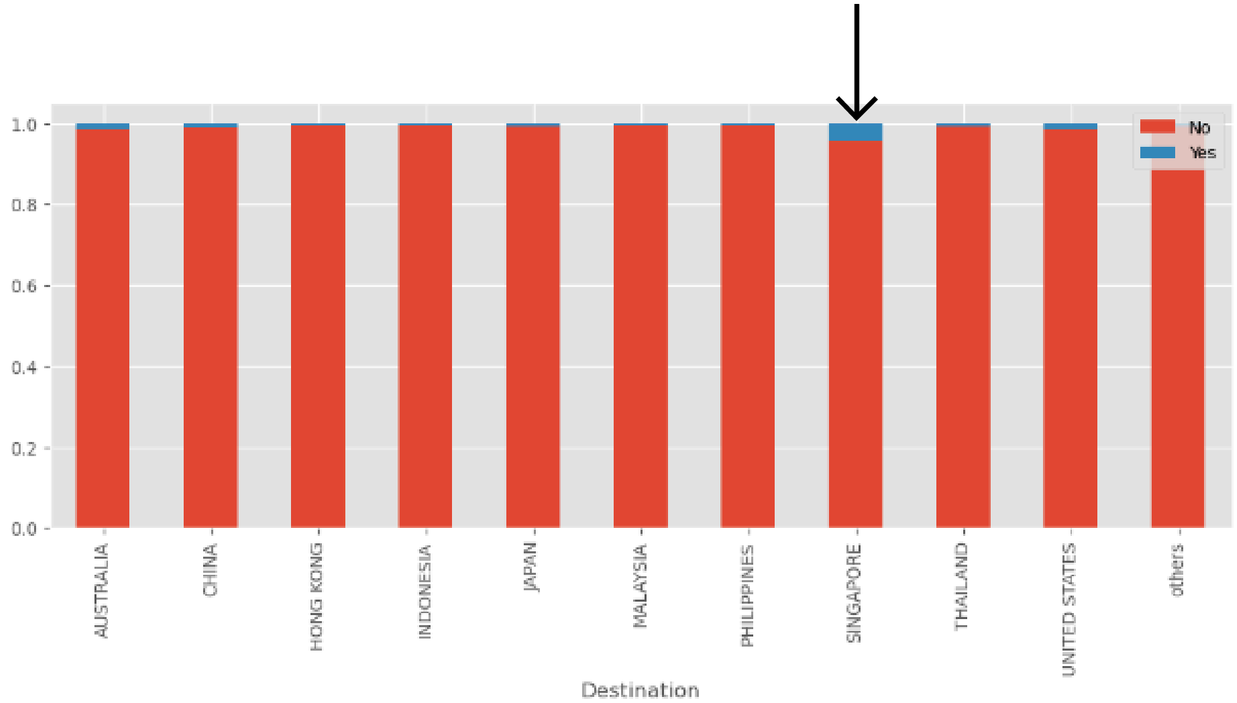


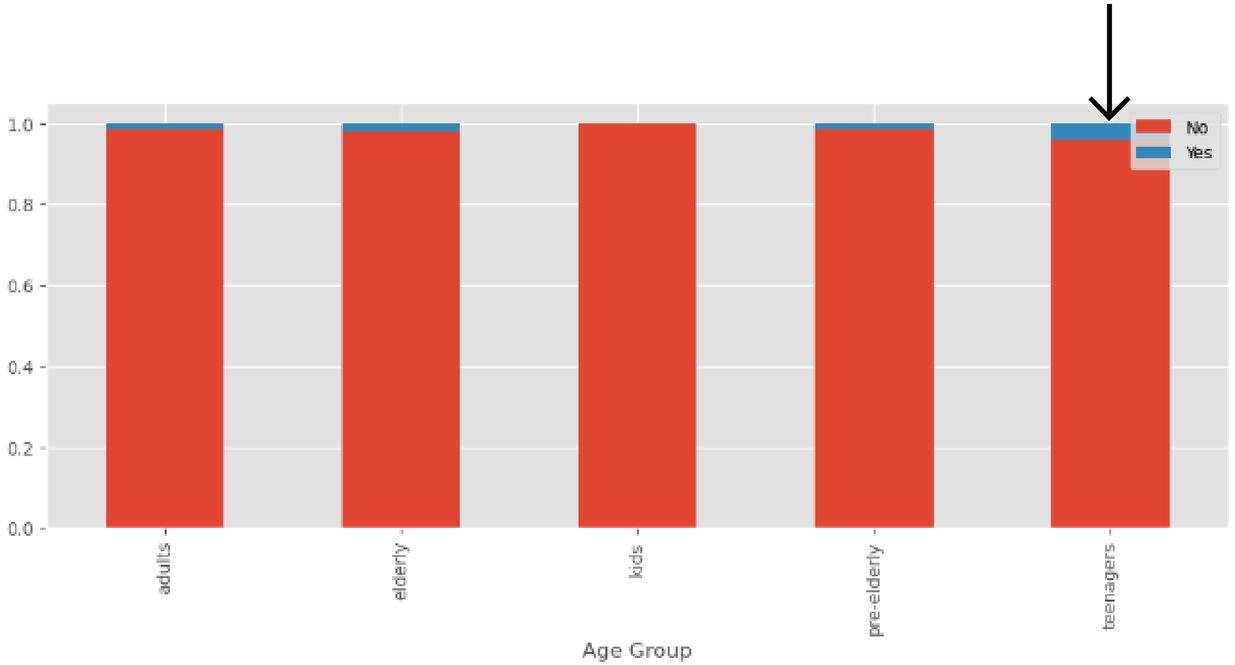
Categorical vs Target



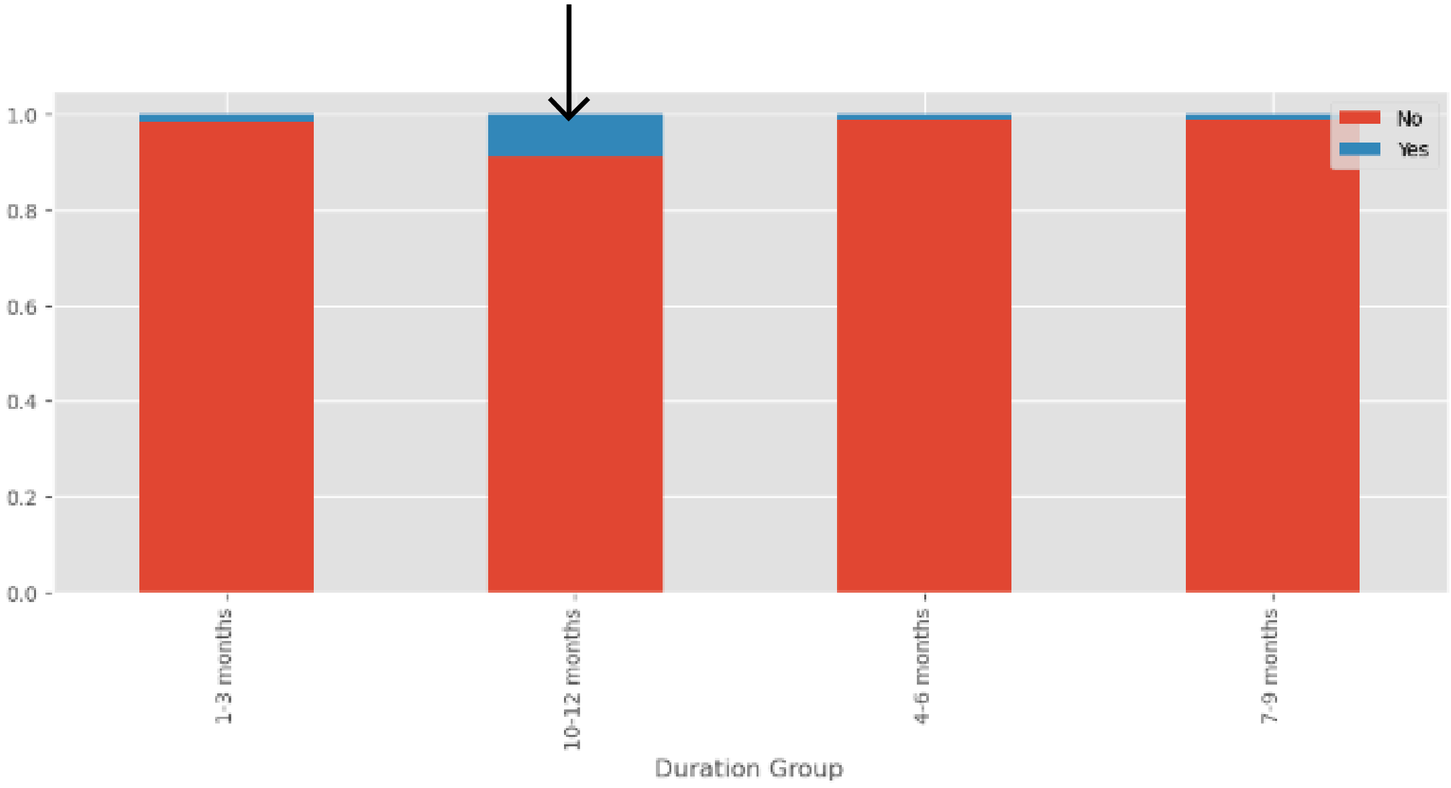


Categorical vs Target





Categorical vs Target



VIF, Cardinality, & Imbalance Data

Variance inflation factor (VIF)

	Feature	VIF
1	Net Sales	3.424214
3	Age	2.061769
2	Commision (in value)	1.959678
0	Duration	1.889018

Cardinality

	Feature	Cardinality	Warning
0	Agency	15	High
1	Agency Type	2	Low
2	Distribution Channel	2	Low
3	Product Name	26	High
4	Destination	136	High
5	Claim	2	Low

Imbalance Data

Claim	proportion
No	0.984711
Yes	0.015289

Preprocessing

Handling Outliers

Winsorization

Encoding

OneHot Encoder

‘Agency Type’, ‘Distribution Channel’, ‘Duration Group’

Binary Encoder

‘Agency’, ‘Product Name’, ‘Destination’

Ordinal Encoder

‘Age Group’

Scaling

Robust Scaler

Resampling

SMOTE

Benchmark Model

01 Logistic Regression

02 KNN Classifier

03 Decision Tree Classifier

04 Random Forest Classifier

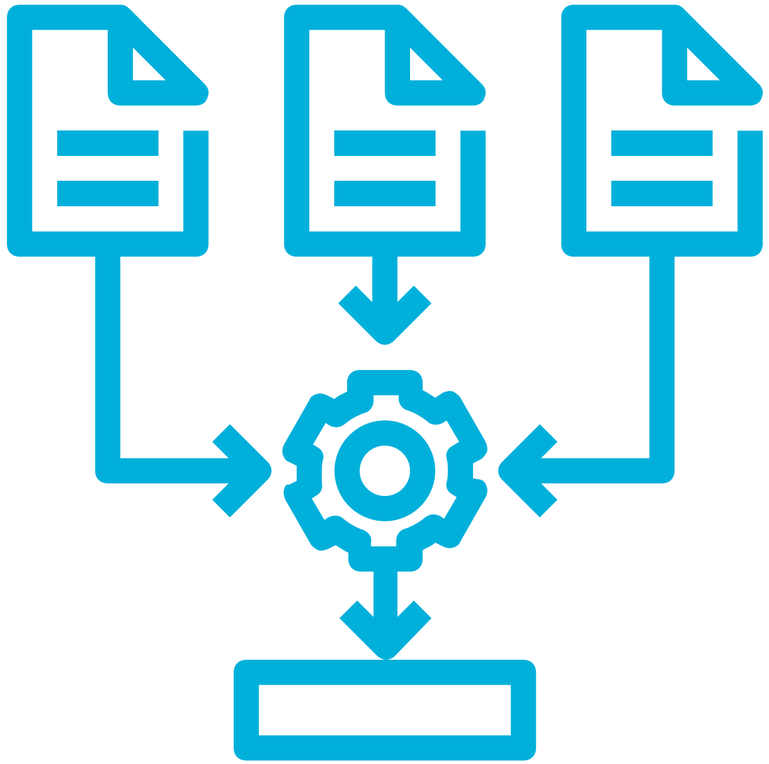
05 Adaptive Booster Classifier

06 Gradient Booster Classifier

07 Categorical Booster Classifier

08 XGBoost Classifier

09 LGBM Classifier



Model Benchmarking

Comparison performance train and test

model	roc_auc train	roc_auc test
Logistic Regression	0.812410	0.711222
AdaBoost	0.786380	0.707284
GradienBoost	0.782484	0.705383
LightGBM	0.768805	0.669062
XGBoost	0.740831	0.668101
CatBoost	0.751149	0.631497
Random Forest	0.685882	0.585911
Decision Tree	0.623156	0.579780
KNN	0.599462	0.533732

Logistic Regression

AdaBoost

GradientBoost

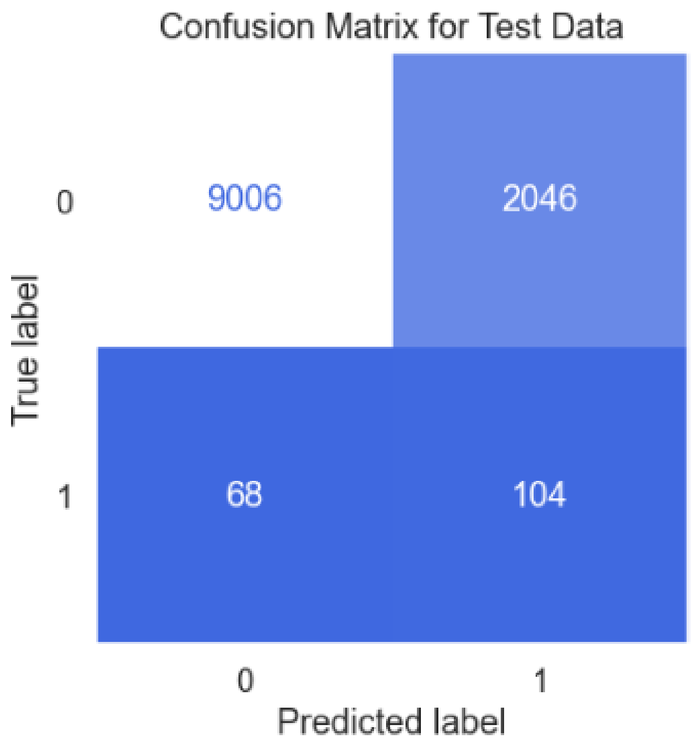
Hyperparameter Tuning

Model Performance Before & After Tuning

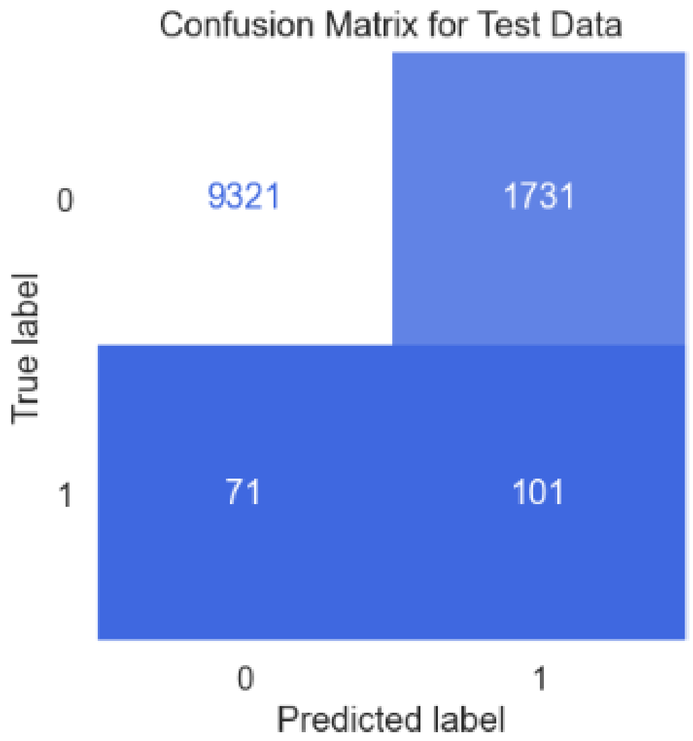
Model	Conditions	Train score	Test score
Logistic Regression	Before Tuning	0.812	0.711
Logistic Regression	After Tuning	0.812	0.709
AdaBoost Classifier	Before Tuning	0.786	0.707
AdaBoost Classifier	After Tuning	0.807	0.715
GradBoost Classifier	Before Tuning	0.780	0.705
GradBoost Classifier	After Tuning	0.806	0.691

Model performance comparison

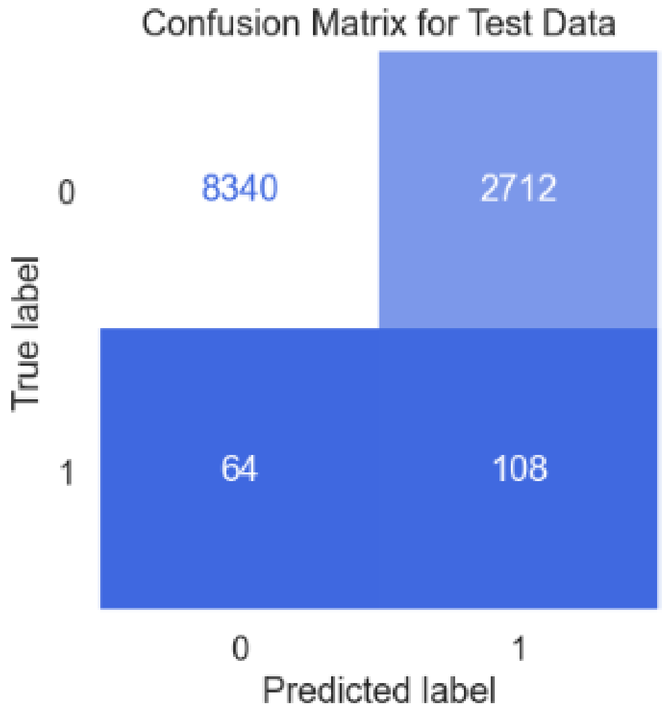
Logistic Regression



AdaBoost Classifier

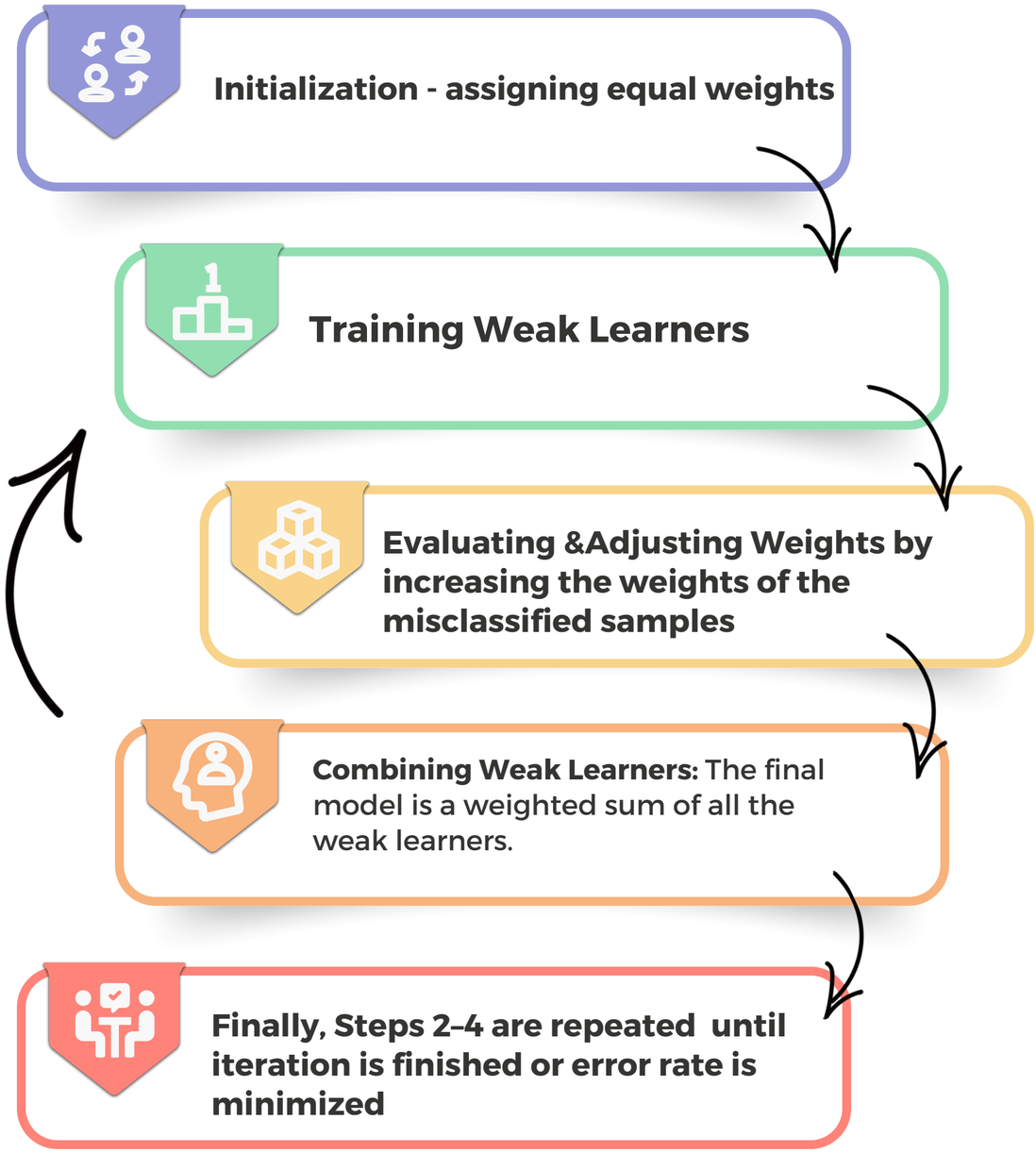


GradBoost Classifier

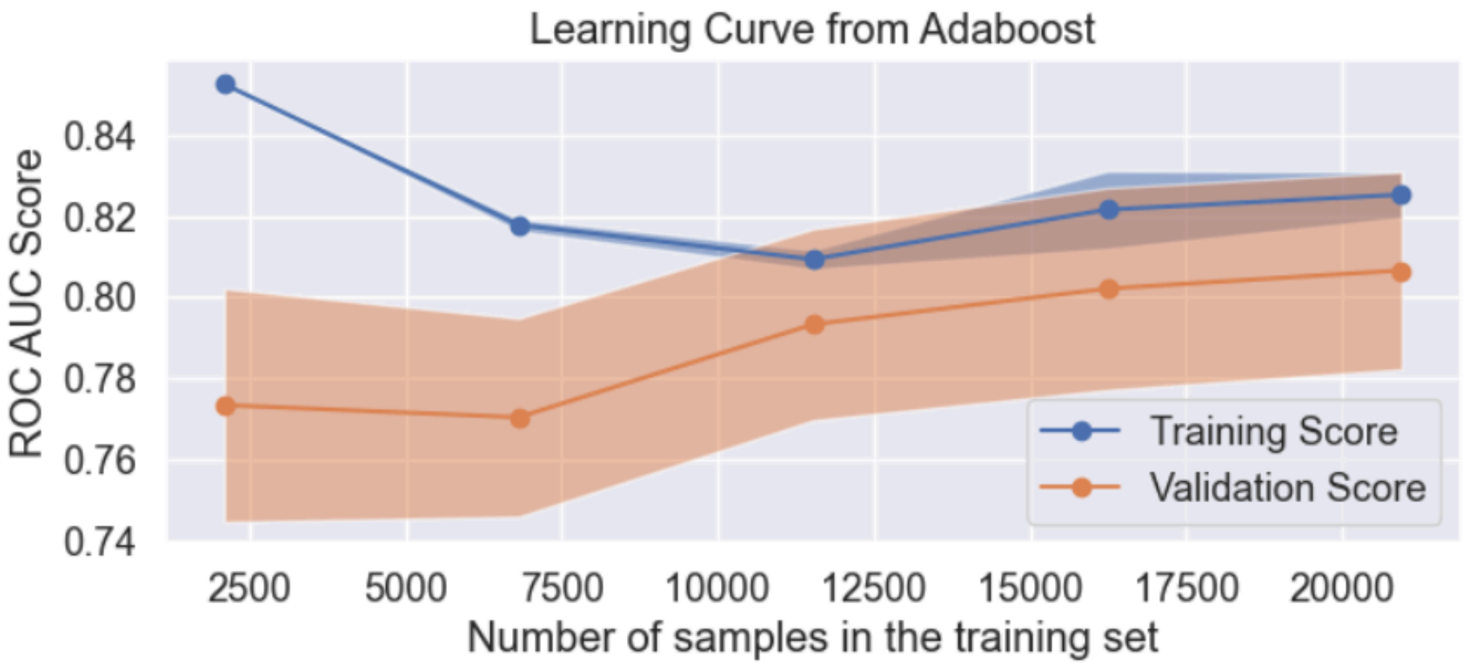


FINAL MODEL: ADABOOST

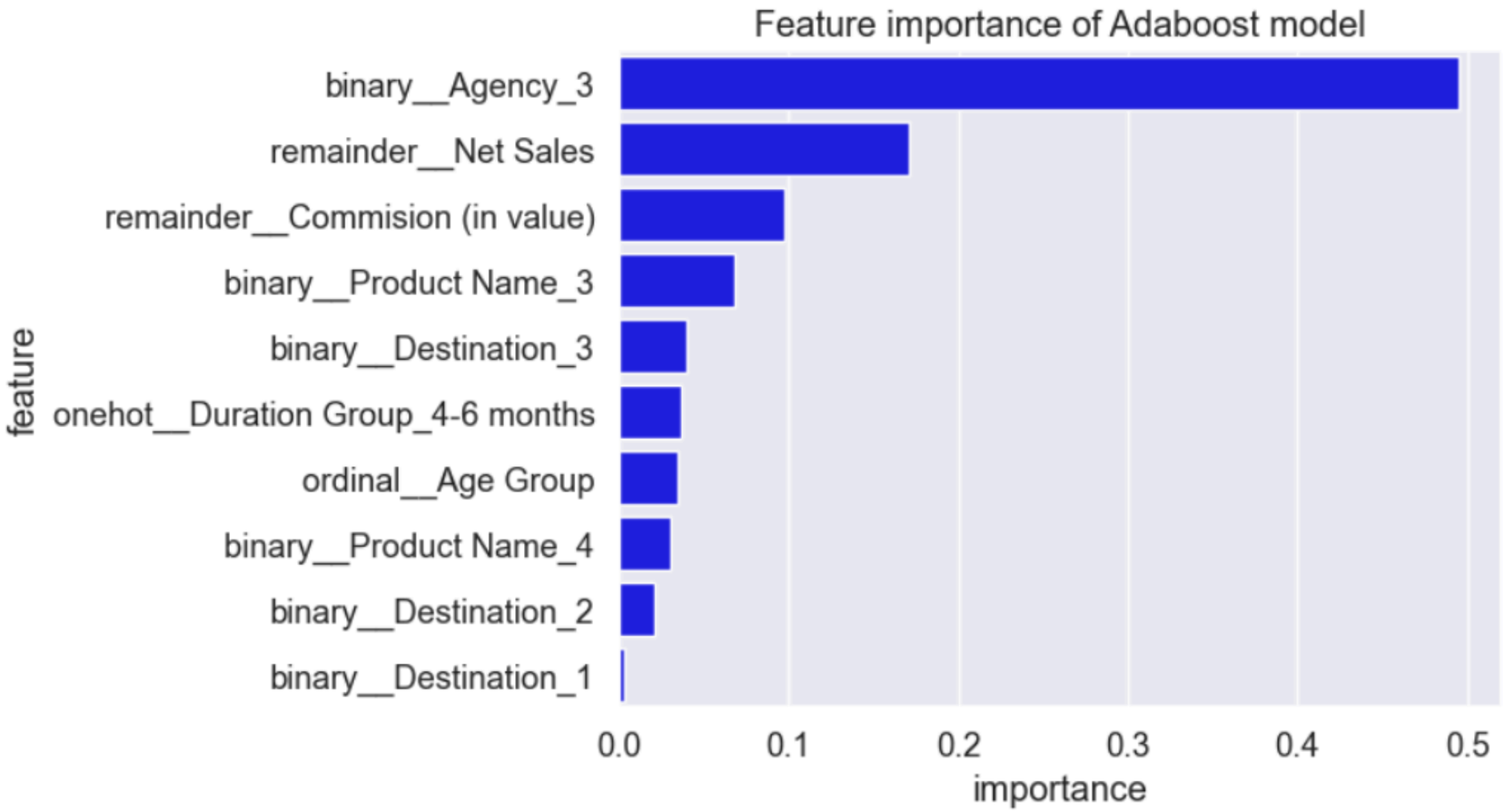
How does the best model work?



Learning curve



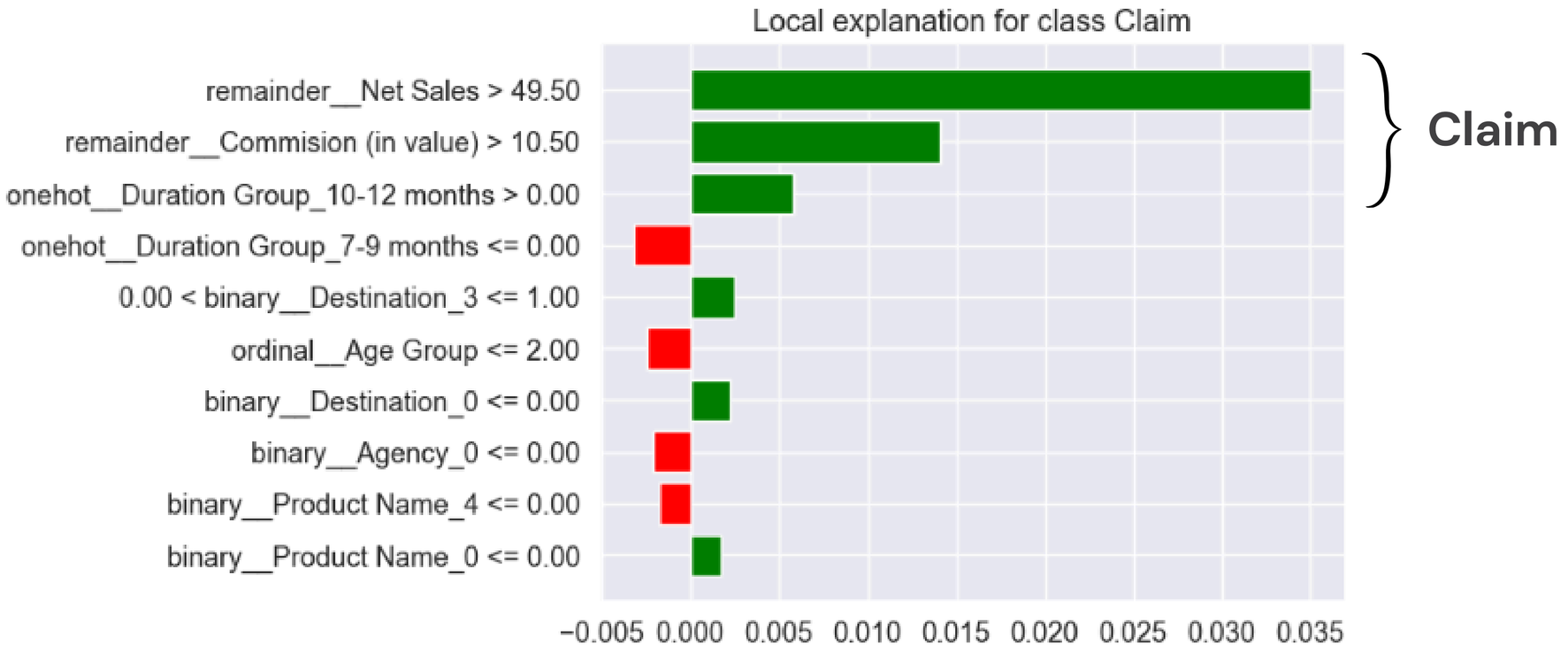
Feature Importance



Explainable Model
LIME (Local Interpretable Model-agnostic)

Customer in the row - 4282 (Predicted as claim)

	Agency	Agency Type	Distribution Channel	Product Name	Destination	Net Sales	Commision (in value)	Age Group	Duration Group
4282	C2B	Airlines	Online	Annual Gold Plan	SINGAPORE	333.0	83.25	adults	10-12 months



Explainable Model
LIME (Local Interpretable Model-agnostic)

Customer in the row - 570 (Predicted as not claim)

	Agency	Agency Type	Distribution Channel	Product Name	Destination	Net Sales	Commision (in value)	Age Group	Duration Group
570	EPX	Travel Agency	Online	Cancellation Plan	SINGAPORE	10.0	0.0	adults	1-3 months

Conlusion

Best Model

Tuned - AdaBoost Classifier

Performance

ROC AUC score 0.715

Feature importance

- Agency
- Net Sales
- Commission (in value)

Lime Explanation

- Net Sales
- Commission (in value)
- Duration
- Destination

RECOMMENDATION FOR ML MODEL

USE OTHER RESAMPLING & ALGORITHMS

Use ADASYN, SMOTE ENN

ADD MORE FEATURES

Policy price

USE OTHER ML METRICS

That improve precision and reduce False Positive

RECOMMENDATION FOR BUSINESS

Implementing ML to automate initiate screening

Provide discounts to low risk customer profile

Establishing higher premium to high-risk customer profile


THANK YOU