Travel Insurance Claim Prediction

Optimizing Travel Insurance Risk and Customer Experience

By Adrian Irshad

Executive Summary

The Challenge :

- Inefficient & costly claims management.
- Lack of predictive insights into potential claims.

Our Solution :

- Developed an ML model (EasyEnsemble + AdaBoost).
- Predicts individual customer claim likelihood.

Key Impact :

- 13% annual loss reduction.
- ~$4.3M in annual savings.

Data Dictionary

Feature	Description
Agency	Name of agency
Agency Type	Type of travel insurance agency
Distribution Channel	Channel used by the insurance agency
Product Name	Name of the travel insurance product
Gender	Gender of the insured individual
Duration	Duration of the travel
Destination	Destination of travel
Net Sales	Sales amount of the insurance policy
Commission (in value)	Commission earned by the agency
Age	Age of the insured individual

Data Description

Feature	Data Type	Missing Values	% Missing	Unique Values	Sample
Agency	object	0	0.00%	16	[C2B, EPX, JZI]
Agency Type	object	0	0.00%	2	[Airlines, Travel Agency]
Distribution Channel	object	0	0.00%	2	[Online, Offline]
Product Name	object	0	0.00%	26	[Annual Silver Plan, Cancellation Plan, Basic Plan]
Gender	object	31,647	71.39%	2	[F, nan, M]
Duration	int64	0	0.00%	437	[365, 4, 19]
Destination	object	0	0.00%	138	[SINGAPORE, MALAYSIA, INDIA]
Net Sales	float64	0	0.00%	1006	[216.0, 10.0, 22.0]
Commission (in value)	float64	0	0.00%	915	[54.0, 0.0, 7.7]
Age	int64	0	0.00%	89	[57, 33, 26]

Why Our Solution Works

1. Data-Driven Foundation

- **Commenced analysis with a robust dataset of 44,328 customer records.**
- **Emphasized data integrity through meticulous processing:**
  - **Eliminated 4,667 duplicate records to enhance data accuracy.**
  - **Disregarded unreliable features (e.g., Gender, exhibiting 71% missing data) to prevent model bias.**
- **Successfully mitigated severe class imbalance:**
  - **Only 1.7% of the dataset comprised actual claims, posing a significant analytical challenge.**
  - **Applied specialized techniques to ensure comprehensive learning from the critical minority class.**

2. Strategic Model Selection & Optimization

- **Conducted a comprehensive benchmark across multiple advanced machine learning models (e.g., Logistic Regression, Random Forest, XGBoost, EasyEnsemble).**
- **Strategically selected EasyEnsemble with AdaBoost for its superior performance, specifically attributed to:**
  - **Exceptional recall for the minority class (claims), which is paramount for loss prevention.**
  - **Proven effectiveness of its ensemble approach in managing imbalanced datasets.**
- **Executed precise hyperparameter tuning:**
  - **Prioritized recall to minimize costly false negatives, a critical business objective.**
  - **Refined model parameters to achieve optimal predictive performance.**

3. Demonstrable Business Value

- **Calculated the financial consequences of prediction errors:**
  - **False positives : Led to unnecessary investigations and increased operational expenditure.**
  - **False negatives : Resulted in missed claims, significant financial payouts, and potential reputational damage.**
- **Targeted model tuning yielded tangible benefits:**
  - **Delivered significant cost savings.**

Enhanced overall operational efficiency.

Supporting Evidence & Detailed Insights

1. Data Preparation

- **Successfully removed duplicates, accounting for 10.5% of the initial dataset.**
- **The 'Gender' column was systematically removed due to 71% missing values, ensuring data reliability.**
- **Rare categories within 'Agency', 'Product Name', and 'Destination' were judiciously grouped into "Others" to reduce noise and prevent overfitting.**

2. Feature Engineering

- **Implemented one-hot encoding for all relevant categorical variables (e.g., Agency, Product Name) to prepare them for model consumption.**
- **Applied robust scaling to numerical features (e.g., Duration, Net Sales, Commission) to normalize their distribution.**
- **Utilized quantile binning for 'Age' to effectively capture non-linear relationships and mitigate the impact of outliers.**

3. Model Evaluation

- **Rigorous comparison of default and tuned models was conducted using key metrics such as recall, precision, and confusion matrices.**
- **Key Performance Indicators:**
  - **Recall for claims significantly improved from 0.78 to 0.81 following optimization.**
  - **Precision, while impacted by class imbalance, remained a consideration.**
  - **A strategic trade-off was acknowledged: higher recall translates to more claims identified, albeit with an increase in false alarms.**

Cross Validation

Algorithm	All Scores	Mean Score	Std. Dev.
Easy Ensemble	[0.77, 0.75, 0.7, 0.68, 0.72]	0.723	0.032
Logistic Regression	[0.73, 0.75, 0.67, 0.64, 0.70]	0.697	0.040
SVM	[0.73, 0.72, 0.67, 0.66, 0.66]	0.688	0.031
Balanced Random Forest	[0.71, 0.69, 0.66, 0.64, 0.71]	0.680	0.028
LightGBM	[0.60, 0.48, 0.54, 0.53, 0.52]	0.534	0.037
CatBoost	[0.49, 0.46, 0.43, 0.45, 0.39]	0.444	0.032
XGBoost	[0.42, 0.44, 0.41, 0.40, 0.36]	0.405	0.029
Random Forest	[0.07, 0.13, 0.04, 0.06, 0.08]	0.074	0.032

Classification Report

Before :

Class	Precision	Recall	F1-Score	Support
0 (No Claim)	0.99	0.76	0.86	7,730
1 (Claim)	0.05	0.78	0.10	135
Accuracy			0.76	7,865
Macro Avg	0.52	0.77	0.48	7,865
Weighted Avg	0.98	0.76	0.85	7,865

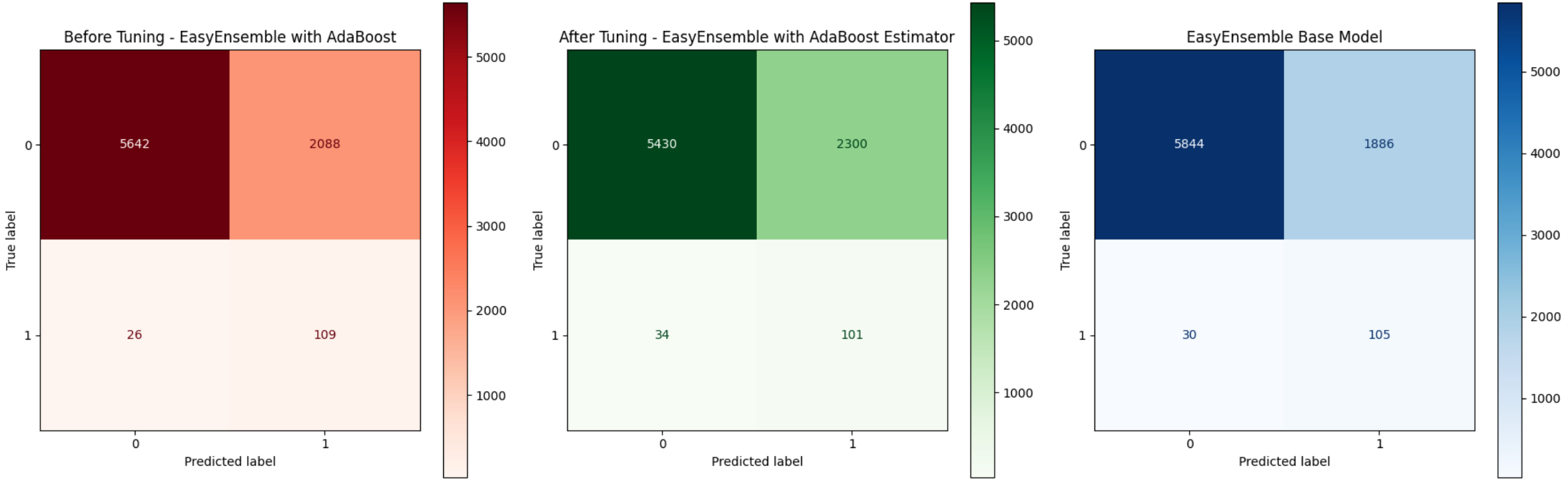
After :

Class	Precision	Recall	F1-Score	Support
0 (No Claim)	1.00	0.73	0.84	7,730
1 (Claim)	0.05	0.81	0.09	135
Accuracy			0.73	7,865
Macro Avg	0.52	0.77	0.47	7,865
Weighted Avg	0.98	0.73	0.83	7,865

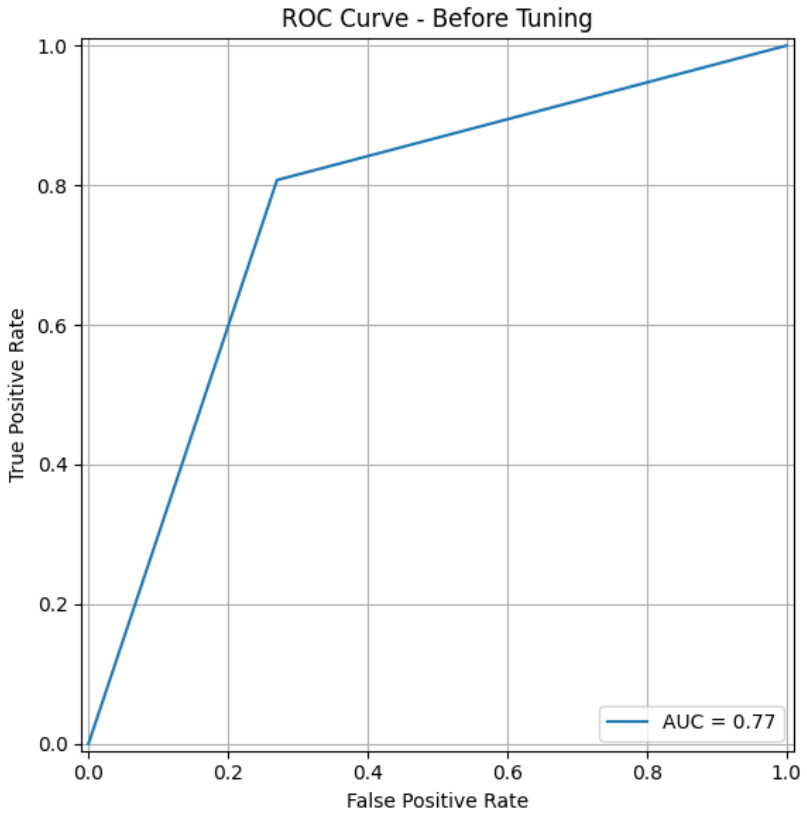
Confusion Matrix

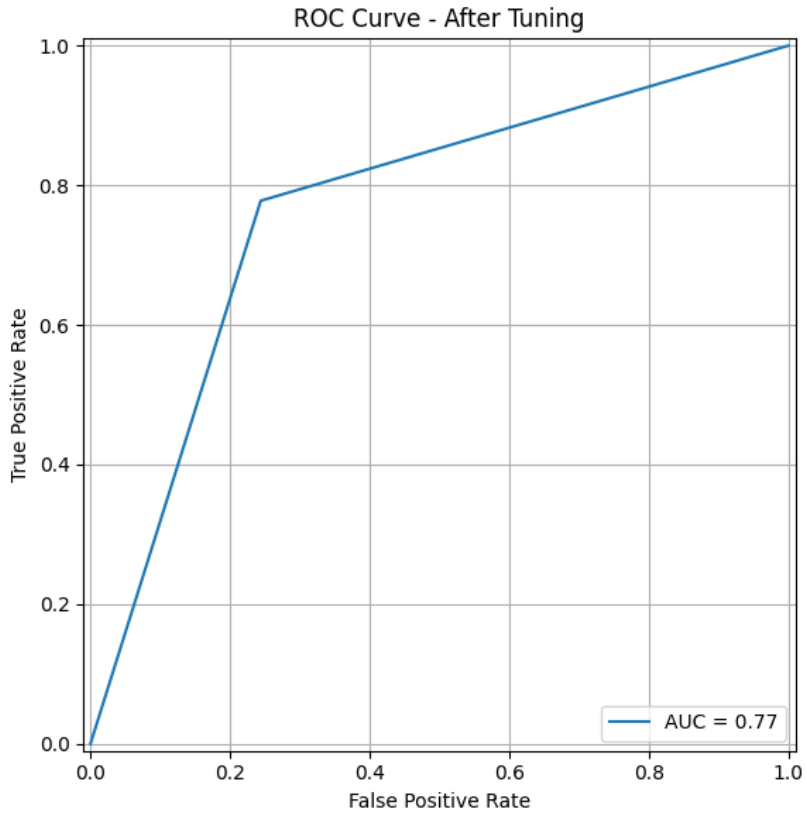
		Predicted	
		Not Claim (0)	**Claim (1)**
Actual	**Not Claim (0)**	**TRUE NEGATIVE (TN)** Model predicts no claim, actual is no claim	**FALSE POSITIVE (FP)** Model predicts claim, actual is no claim
	Claim (1)	**FALSE NEGATIVE (FN)** Model predicts no claim, actual is claim	**TRUE POSITIVE (TP)** Model predicts claim, actual is claim

Confusion Matrix

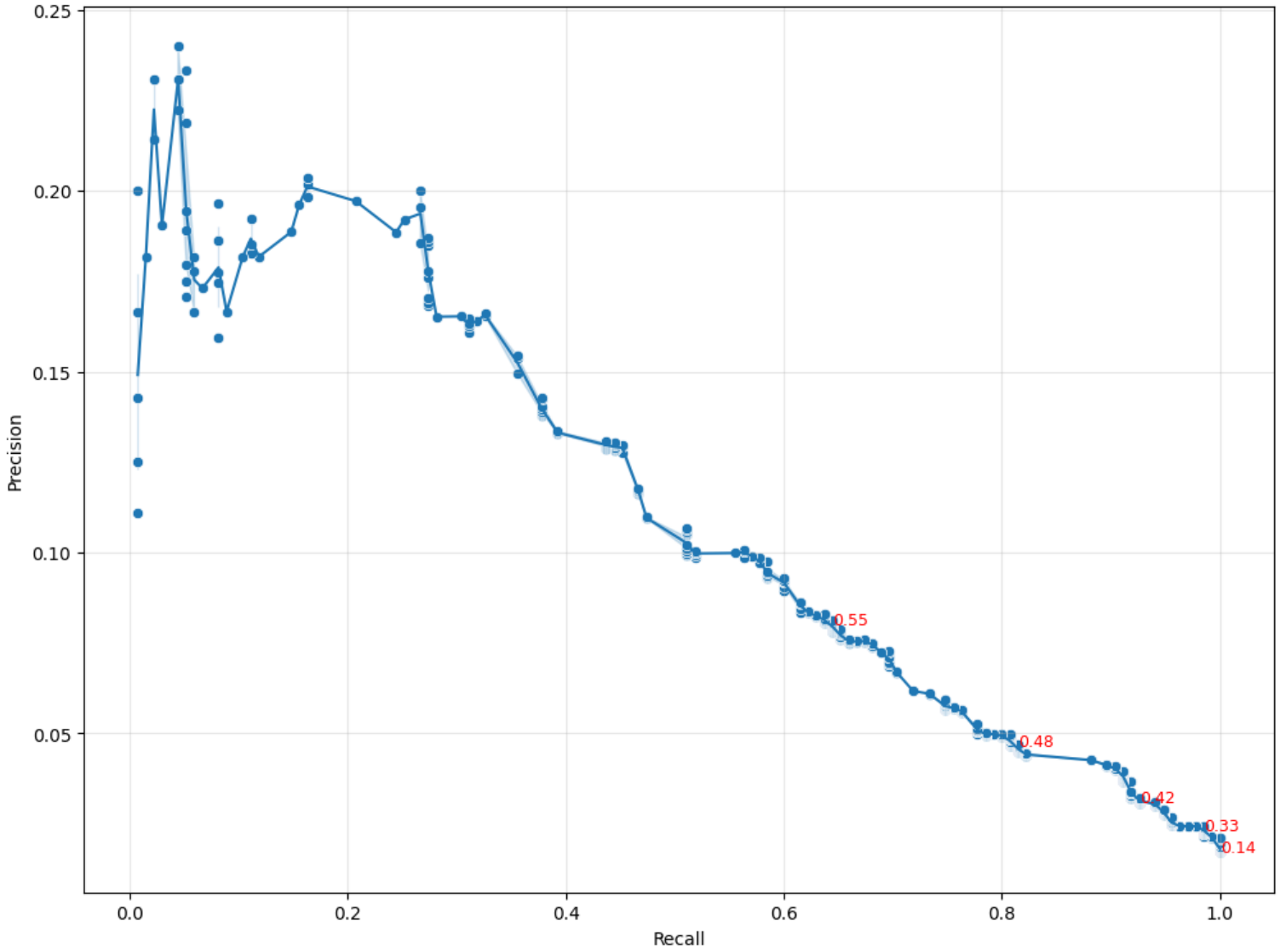


Receiver Operating Curve (ROC)





Precision-Recall (PR) Curve



4. Financial Analysis

PERFORMED A COMPREHENSIVE CALCULATION OF ANNUAL LOSSES ATTRIBUTABLE TO BOTH FALSE POSITIVES AND FALSE NEGATIVES.

DEFAULT MODEL : EXHIBITED AN ESTIMATED TOTAL ANNUAL LOSS OF $32.2 MILLION.

TUNED MODEL : SIGNIFICANTLY REDUCED THE TOTAL ANNUAL LOSS TO $27.9 MILLION.

THIS OPTIMIZATION RESULTED IN SUBSTANTIAL ANNUAL SAVINGS OF $4.3 MILLION, REPRESENTING A 13% REDUCTION IN TOTAL LOSSES.

Comparative Financial Evaluation

Model	False Positives	False Negatives	Cost of False Positives	Cost of False Negatives	Total Estimated Loss
Default Model	1,855	30	$84,403	$32,100,000	$32,184,403
Tuned Model	2,087	26	$94,939	$27,820,000	$27,914,939

- **Annual Savings from Model Tuning**
  - **Absolute Savings:** $4,269,464
  - **Percentage Reduction in Losses:** 13.27%

Conclusion

A comprehensive approach to data cleaning and model development led to more reliable predictions of travel insurance claims.

Despite a highly imbalanced dataset, the final solution significantly improved our ability to identify valid claims with minimal risk to overall accuracy.

The optimized model helped reduce the potential annual financial loss by **over $4.2 million**, representing a **13.27% improvement** compared to the baseline.

The business now has a data-driven tool that supports **smarter risk management decisions** and enhances claim prediction strategies.

This project demonstrates how strategic use of machine learning can deliver **tangible financial value** and drive better operational outcomes.

Recommendations

Monitor Model Performance
Review metrics quarterly; focus on reducing costly false negatives.

Enhance Data Quality
Increase valid claims via cross-team efforts; run regular data audits.

Update Key Features
Reassess inputs like age, agency, and destination; add behavioural trends.

Maintain & Benchmark Model
Keep using EasyEnsemble; compare with new models as data evolves.

Optimize Risk Threshold
Adjust threshold (e.g., 0.48) to balance detection and review cost.

Embed into Operations
Use predictions to flag risks or adjust premiums; ensure workflow integration.

Improve Communication
Build dashboards; clearly report model scope and limitations.

Thank You