Claim fraud detection report

Introduction

The report will include the overall approach of the assignment, covering the problem statement, methodology, techniques used and key insights.

Overall Approach

The overall approach was to do analysis, clean-up, modelling and predictions. Follow various steps learnt so far in ML journey.

The objective is to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features such as claim amounts, customer profiles, claim types and approval times, the company aims to predict the claims that are likely to be fraudulent before they are approved.

Methodology

Steps to solve the problem:

Logistics regression

- Raw Data
- Data cleaning : handle missing values, outliers
- Train-Test Split
- Encoding categorical features : get_dummies
- Scale numeric features
- Features selection using RFECV
- Resampling
- Train final model
- Predict
- Evaluate performance : accuracy, F1, AUC, confusion matrix

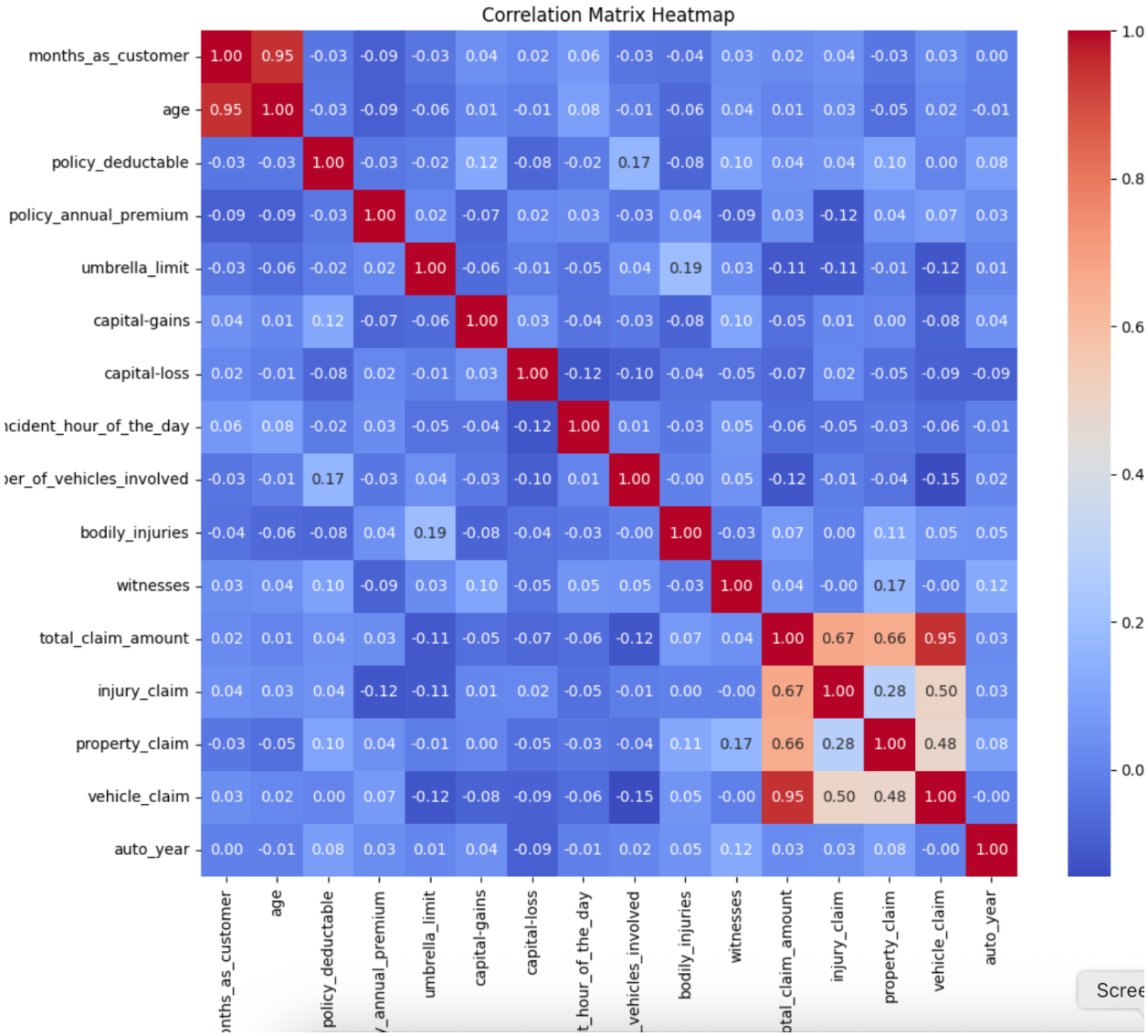
Random Forrest

- Raw Data
- Data cleaning : handle missing values, outliers
- Train-Test Split
- Encoding categorical features : get_dummies
- Features selection using feature importance
- Resampling
- Train final model
- Predict
- Evaluate performance : accuracy, F1, AUC, confusion matrix

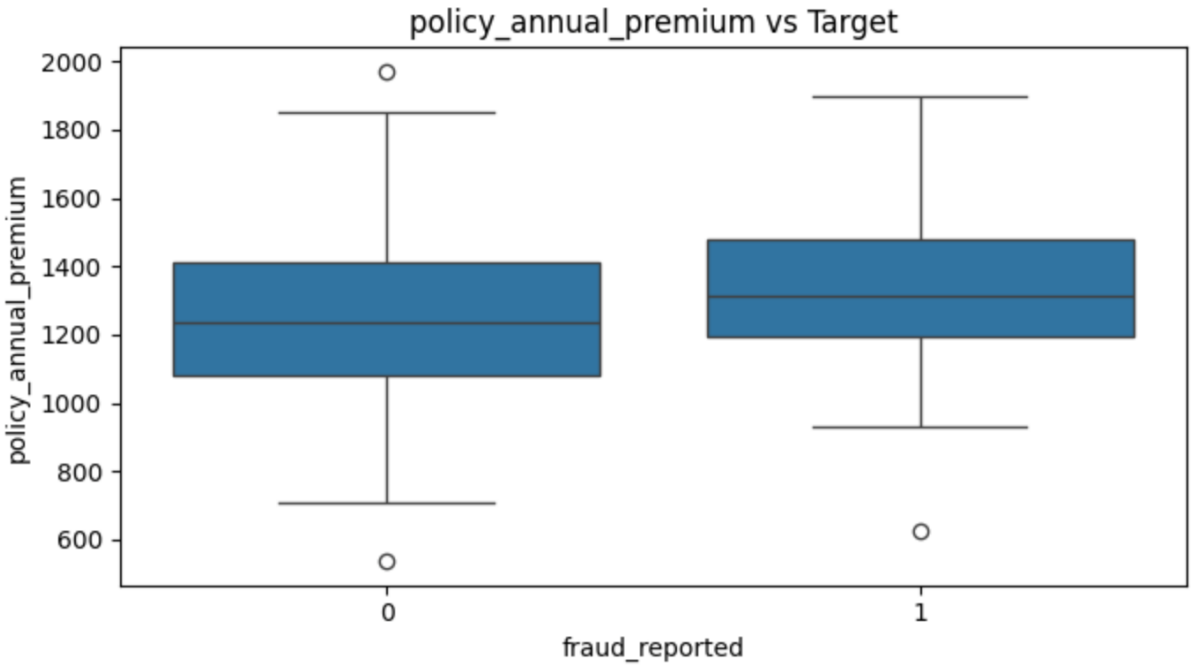
Key Insights

It was lengthy & tough assignment. It has to be done carefully considering so many steps. Mixing the steps can create much different results. Metrics used to find different aspects of data, model & prediction:

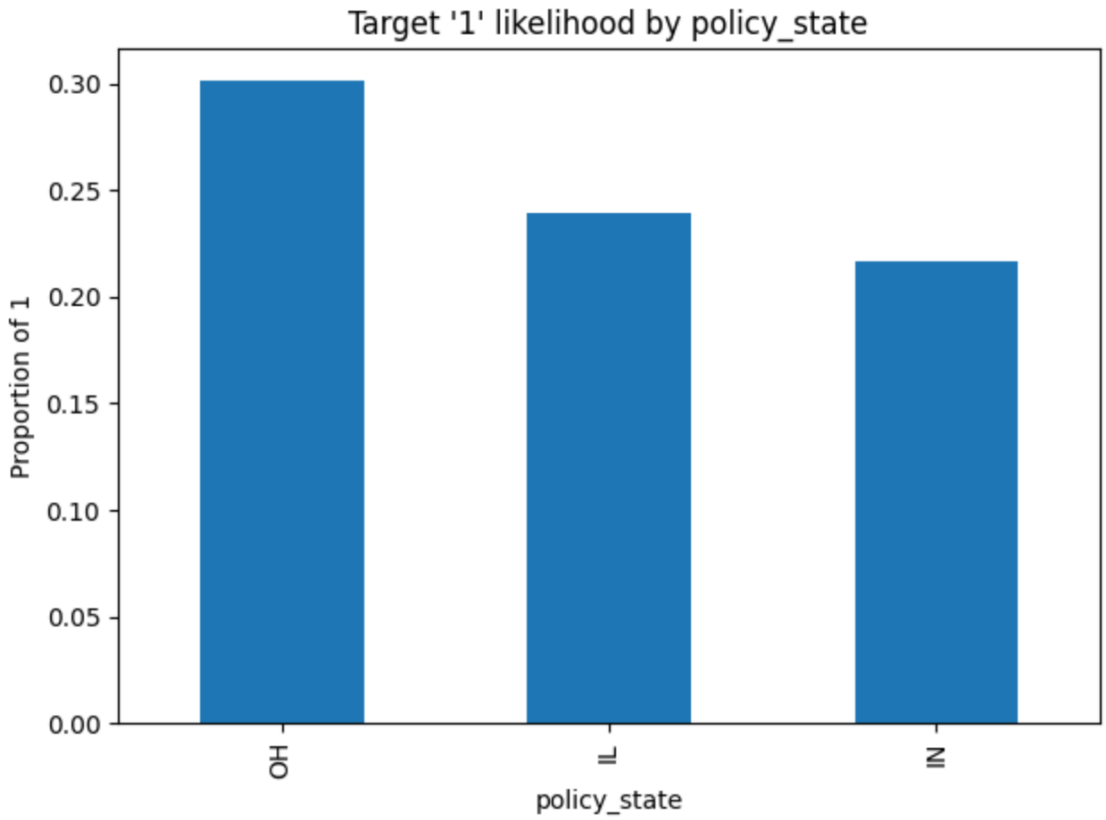
Heatmap


Correlation Matrix Heatmap
months_as_customer 1.00 0.95 -0.03 -0.09 -0.03 0.04 0.02 0.06 -0.03 -0.04 0.03 0.02 0.04 -0.03 0.03 0.00
age 0.95 1.00 -0.03 -0.09 -0.06 0.01 -0.01 0.08 -0.01 -0.06 0.04 0.01 0.03 -0.05 0.02 -0.01
policy_deductable -0.03 -0.03 1.00 -0.03 -0.02 0.12 -0.08 -0.02 0.17 -0.08 0.10 0.04 0.04 0.10 0.00 0.08
policy_annual_premium -0.09 -0.09 -0.03 1.00 0.02 -0.07 0.02 0.03 -0.03 0.04 -0.09 0.03 -0.12 0.04 0.07 0.03
umbrella_limit -0.03 -0.06 -0.02 0.02 1.00 -0.06 -0.01 -0.05 0.04 0.19 0.03 -0.11 -0.11 -0.01 -0.12 0.01
capital-gains 0.04 0.01 0.12 -0.07 -0.06 1.00 0.03 -0.04 -0.03 -0.08 0.10 -0.05 0.01 0.00 -0.08 0.04
capital-loss 0.02 -0.01 -0.08 0.02 -0.01 0.03 1.00 -0.12 -0.10 -0.04 -0.05 -0.07 0.02 -0.05 -0.09 -0.09
ıcident_hour_of_the_day 0.06 0.08 -0.02 0.03 -0.05 -0.04 -0.12 1.00 0.01 -0.03 0.05 -0.06 -0.05 -0.03 -0.06 -0.01
ɔer_of_vehicles_involved -0.03 -0.01 0.17 -0.03 0.04 -0.03 -0.10 0.01 1.00 -0.00 0.05 -0.12 -0.01 -0.04 -0.15 0.02
bodily_injuries -0.04 -0.06 -0.08 0.04 0.19 -0.08 -0.04 -0.03 -0.00 1.00 -0.03 0.07 0.00 0.11 0.05 0.05
witnesses 0.03 0.04 0.10 -0.09 0.03 0.10 -0.05 0.05 0.05 -0.03 1.00 0.04 -0.00 0.17 -0.00 0.12
total_claim_amount 0.02 0.01 0.04 0.03 -0.11 -0.05 -0.07 -0.06 -0.12 0.07 0.04 1.00 0.67 0.66 0.95 0.03
injury_claim 0.04 0.03 0.04 -0.12 -0.11 0.01 0.02 -0.05 -0.01 0.00 -0.00 0.67 1.00 0.28 0.50 0.03
property_claim -0.03 -0.05 0.10 0.04 -0.01 0.00 -0.05 -0.03 -0.04 0.11 0.17 0.66 0.28 1.00 0.48 0.08
vehicle_claim 0.03 0.02 0.00 0.07 -0.12 -0.08 -0.09 -0.06 -0.15 0.05 -0.00 0.95 0.50 0.48 1.00 -0.00
auto_year 0.00 -0.01 0.08 0.03 0.01 0.04 -0.09 -0.01 0.02 0.05 0.12 0.03 0.03 0.08 -0.00 1.00
onths_as_customer
age
policy_deductable
/_annual_premium
umbrella_limit
capital-gains
capital-loss
t_hour_of_the_day
_vehicles_involved
bodily_injuries
witnesses
otal_claim_amount
injury_claim
property_claim
vehicle_claim
auto_year
1.0
0.8
0.6
0.4
0.2
0.0
Scree

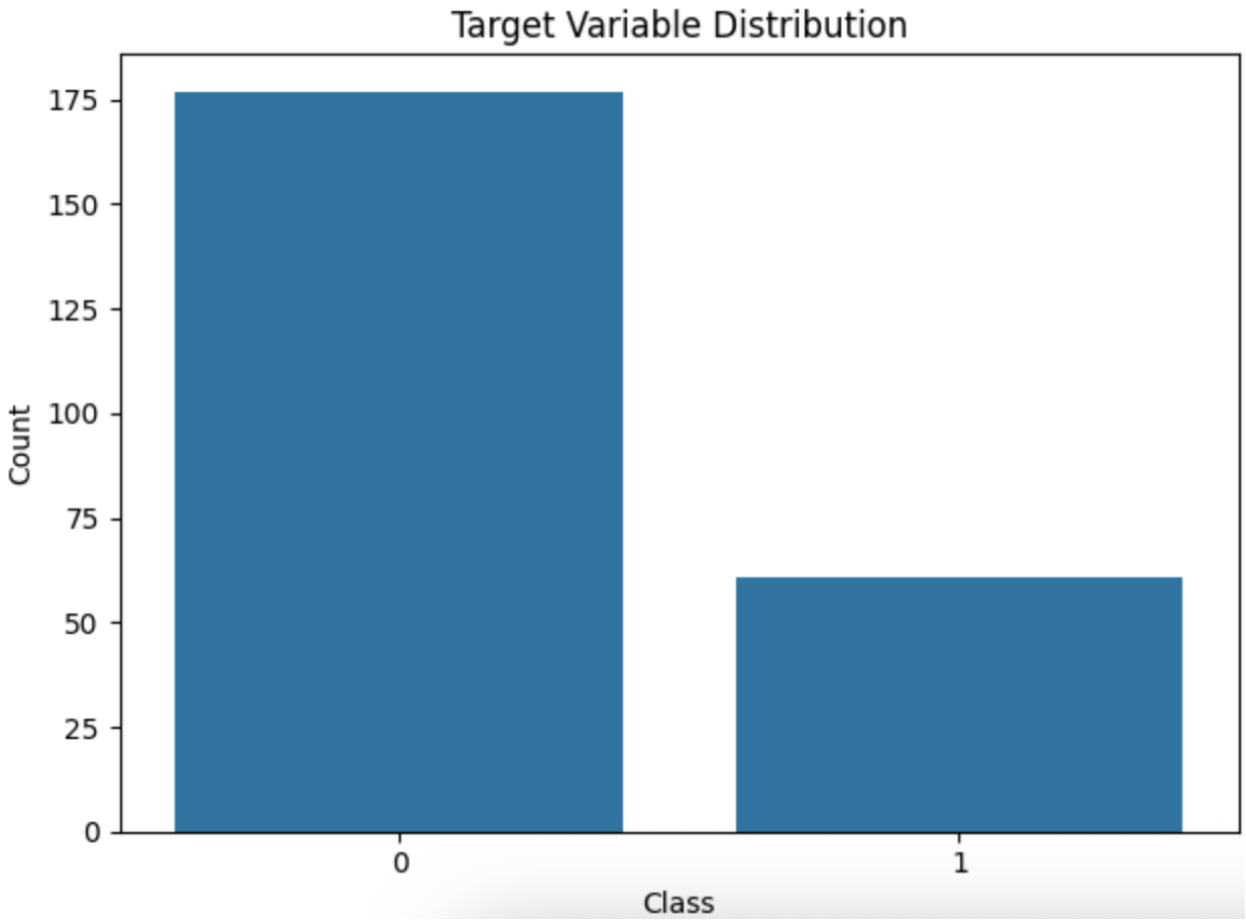
Relationship between numerical & target variable



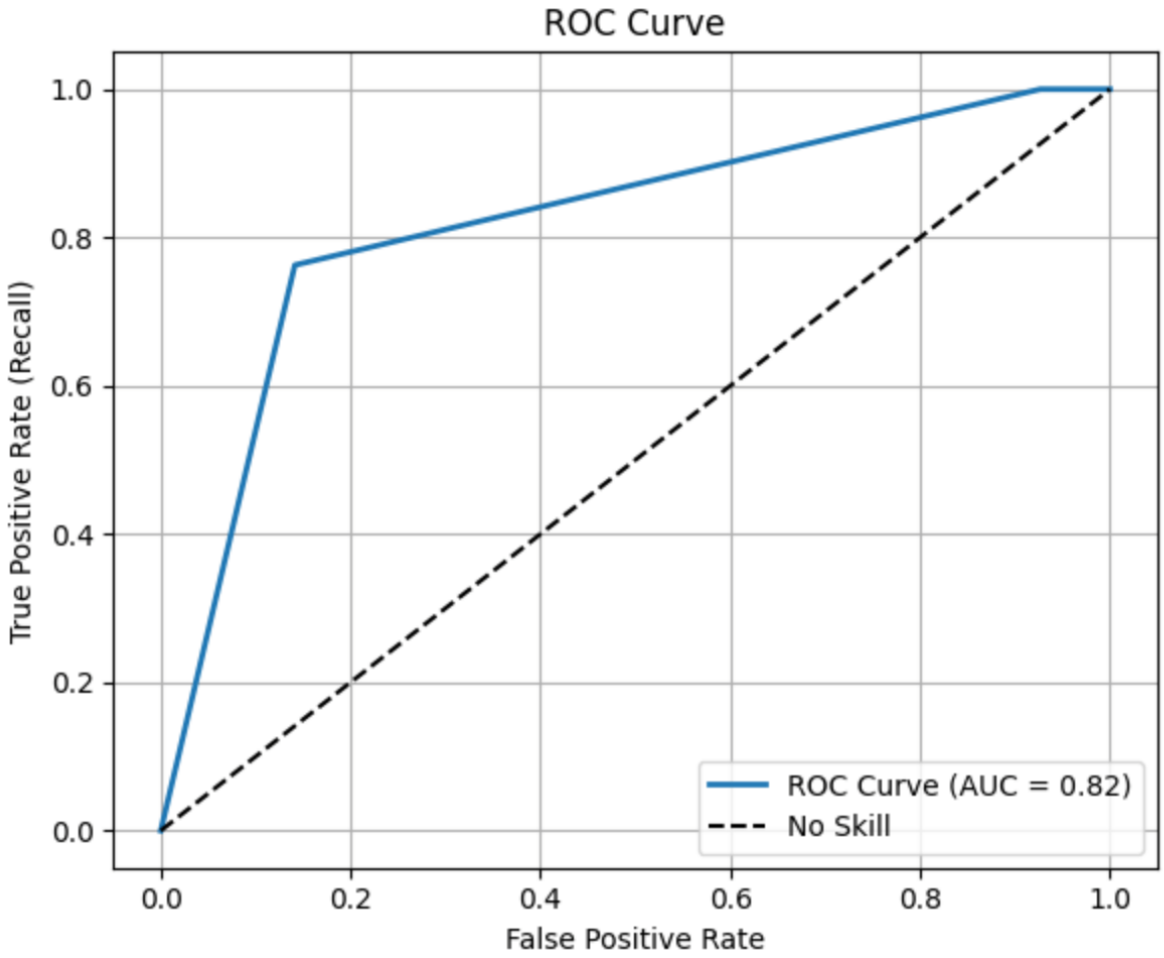
Categorical feature likelihood



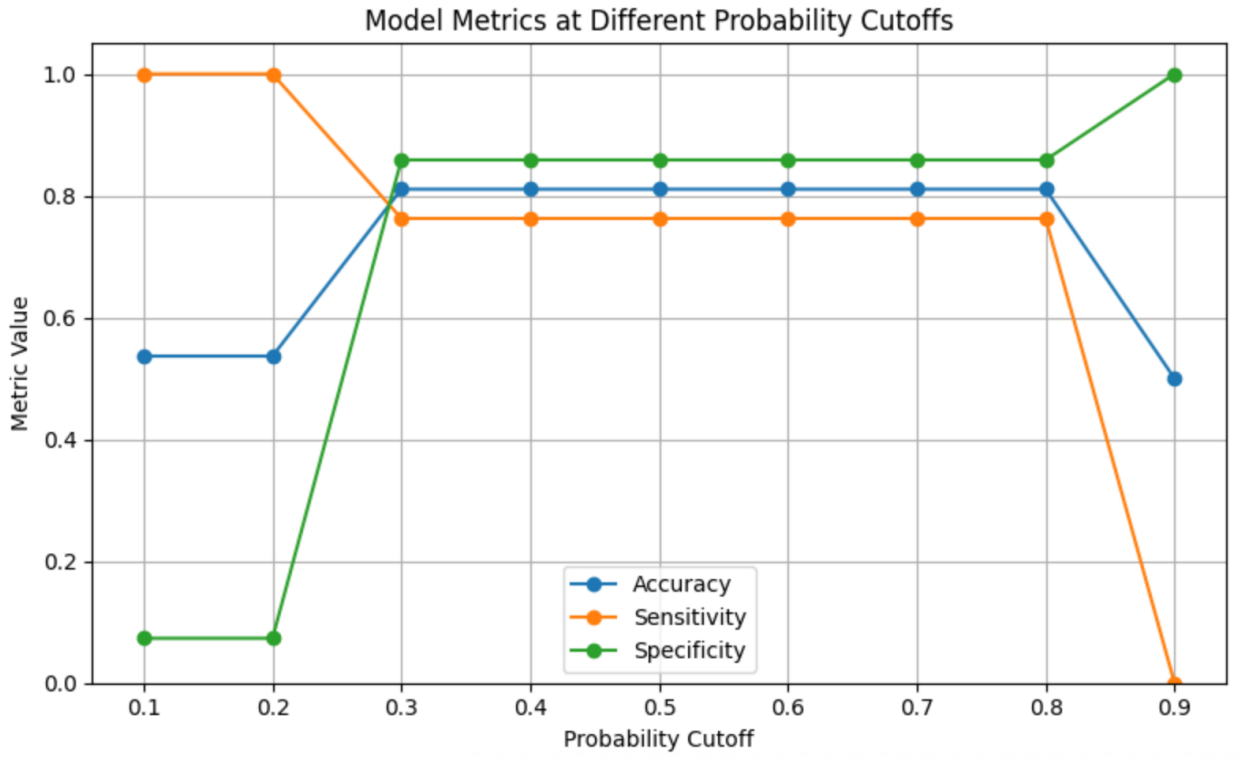
Class balance


Target Variable Distribution
175
150
125
100
75
50
25
0
Count
0
1
Class

ROC Curve


ROC Curve
True Positive Rate (Recall)
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC Curve (AUC = 0.82)
No Skill

Probability Cutoff



Precision-Recall Tradeoff