Capstone II Final Report

Introduction

This project models customer churn in a large dataset from a French telecom provider. The goal of this project is to develop a predictive model that identifies customers who are likely to terminate service; which would help the business direct tactics towards these customers aimed at retaining them.

Problem Statement

Customer churn represents lost revenue to the business. The capacity to proactively identify customers at risk of terminating service allows the business to attempt to retain these customers; potentially retaining revenue that would otherwise be lost.

Data Overview and Preprocessing

The dataset, originally part of a machine learning competition, contained a large number of unlabeled features and required several preprocessing steps. Missing values were handled, categorical variables were encoded and continuous variables were standardized.

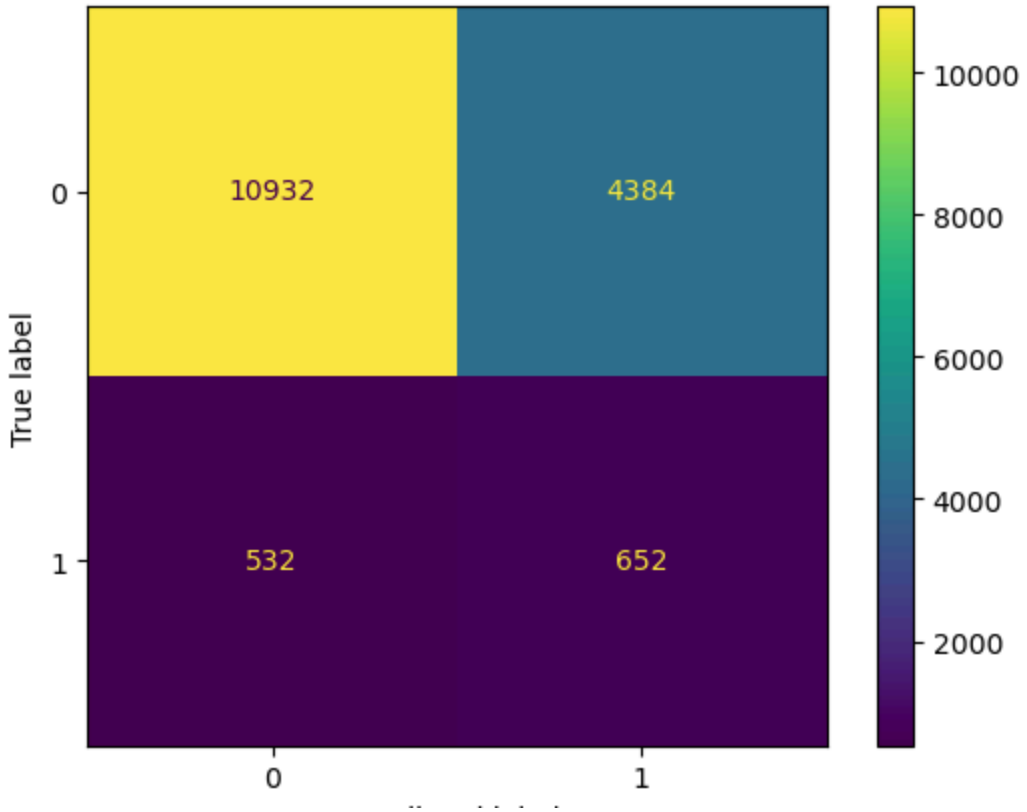
Visualizations and statistical analysis was conducted to understand feature distributions and pairwise relationships. Variables that did not show a relationship to the outcome variable were dropped and variables that were highly correlated to each other were also removed from the data.

Modeling

Three models were evaluated - Logistic Regression, Random Forest and Gradient Boosting. It was found that logistic regression, with steps taken to handle class imbalance, was the most effective model. This model achieved an F1 score of 0.82, a precision score of 0.13 and a recall score of 0.55.

Model Metrics

	precision	recall	f1-score	support
0	0.95	0.71	0.82	15316
1	0.13	0.55	0.21	1184
accuracy			0.70	16500
macro avg	0.54	0.63	0.51	16500
weighted avg	0.89	0.70	0.77	16500



Conclusion and Recommendations

The model developed in this project achieved performance metrics which indicate that it could potentially be used to identify customers which are more likely than others to terminate service; adding business value by enabling the business to target these customers with tactics aimed at retaining their service.