Data Science Project Report

Introduction

This project aims to build a predictive model for Olusola Insurance Company to determine if a building will have an insurance claim during a certain period or not. The model is based on the building characteristics and the target variable, `Claim`, is binary:

- 1 if the building has at least a claim over the insured period.
- 0 if the building doesn't have a claim over the insured period.

Data Preprocessing

The dataset provided was divided into a training set and a test set. Both datasets required cleaning as they contained missing values and irrelevant data. The numerical features, `Building Dimension` and `Date_of_Occupancy`, were examined for their distribution. Based on their skewness, missing values were filled with either mode or median.

Feature Selection

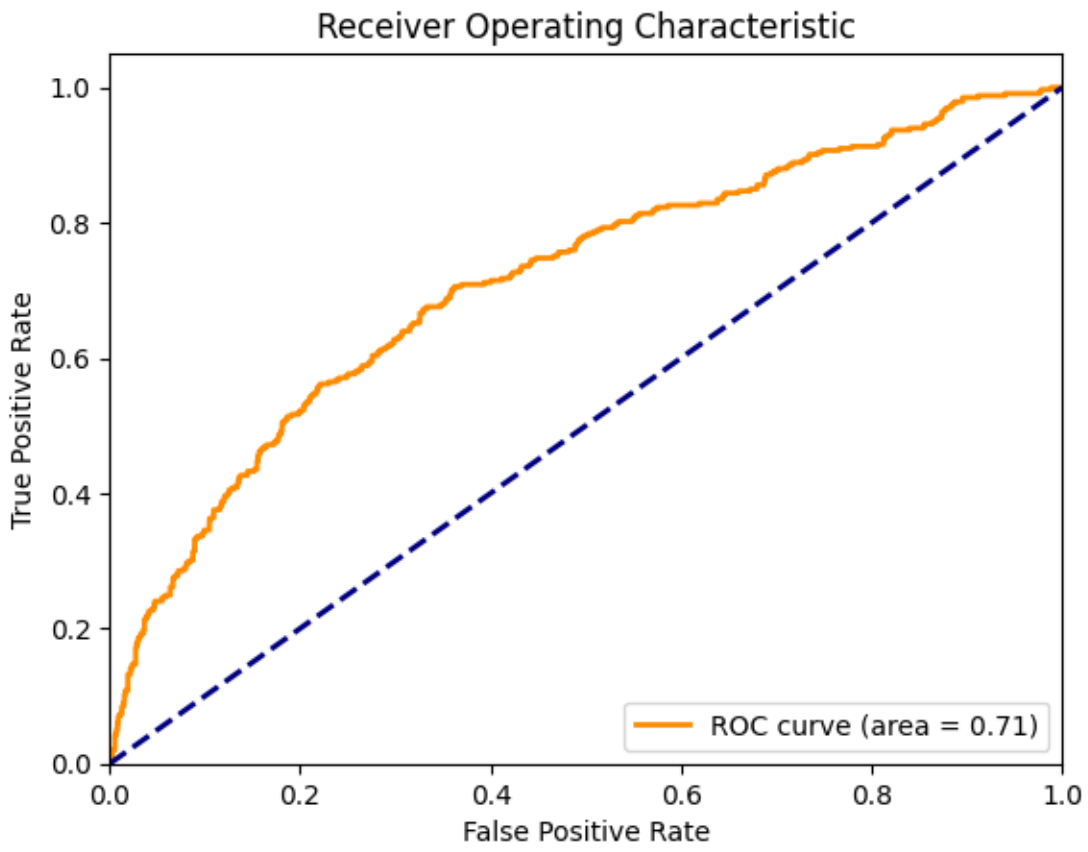
To select the features, a correlation matrix was computed. The `Claim` has a positive correlation with `Building Dimension` (0.295558) and `Building_Type` (0.112168), suggesting that these features might be good predictors for `Claim`. `Residential` and `Building_Type` are positively correlated (0.334039), indicating that these two variables move in the same direction. `Date_of_Occupancy` and `Building_Type` are negatively correlated (-0.137001), suggesting that as one increases, the other decreases.

In addition to correlation, feature importance was also checked. The selected features based on both correlation and feature importance are `Building Dimension`, `Date_of_Occupancy`, `YearOfObservation`, `Insured_Period`, `Building_Type`, and `Residential`.

Model Training and Evaluation

The main data was split into a training set (80%) and a test set (20%). Before this split, the numerical and categorical columns were transformed to make them appropriate for training.

Different models were evaluated for their performance. The Random Forest Classifier achieved an accuracy of 75%, while the Logistic Regression model performed slightly better with an accuracy of 78%. Unhappy with the model performance, model tuning was performed using GridSearchCV. The best parameters were `{'C': 1, 'penalty': 'l2'}`. Defining the model with the best parameters did not improve the accuracy, which remained at 78%. Therefore, this model version was selected for further steps.



The performance of the model was evaluated using various metrics. One of the key metrics used was the Receiver Operating Characteristic (ROC) curve, which is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity.

The area under the ROC curve (AUC) is a single number summary of the overall performance of the model. The AUC for our model was found to be 0.71. An AUC of 0.5 suggests no discrimination (i.e., ability to distinguish the classes), 1 indicates perfect discrimination, while 0.71 suggests a good level of discrimination.

This indicates that our model has a good predictive performance, with a 71% chance that the model will be able to distinguish between positive class and negative class.

Test Data Preprocessing and Prediction

The test data was preprocessed in the same way as the main data. After preprocessing, the cleaned test data was imported into the model to generate predictions. The prediction only contains data with `Customer Id` and `Claim`, and was saved as a CSV file.

Future Work

In the future, the focus will be on deploying the model for practical use.

Prepared by Endework Abera