Guided Capstone Project Report

Problem Identification Overview

- **Data Available -** A generic survey dataset that Stack Overflow conducted with its community of developers.
  - Employment Type
  - Compensation received
  - Development Type
  - Age
  - Job Satisfaction Level
  - Total years as a professional Coder
- This is a Classification Model
- **Data Time frame -** The dataset is survey conducted within Stack Overflow Community in the year 2020

Data Preprocessing Steps of Note

- There is minimal percentage of outliers
- The data has a moderate amount of Null values. Used certain methods such as below to fill the na values based on the column values
  - Fill with specific values
  - Fill with a new but generic value
  - Fill with mean for Numerical values
- Created a couple a new features based on the following methodologies listed below
  - One hot encoding
  - Binning of Column datas
  - Categorising columns

Model Description

- **Input Data Size -** Total **64461** rows/survey details and **62**features
- **Algorithm used -** Tried three different algorithms such as **Logistic Regression**, **Random Forest** and **Gradient Boosting**

Model Performance

Below is the performance metrics of individual models tried using various set of Parameters through Grid Search CV

Logistic regression

Best Score:0.8685690043241499

Best Parameters: {'C': 1, 'penalty': 'l1', 'solver': 'liblinear'}

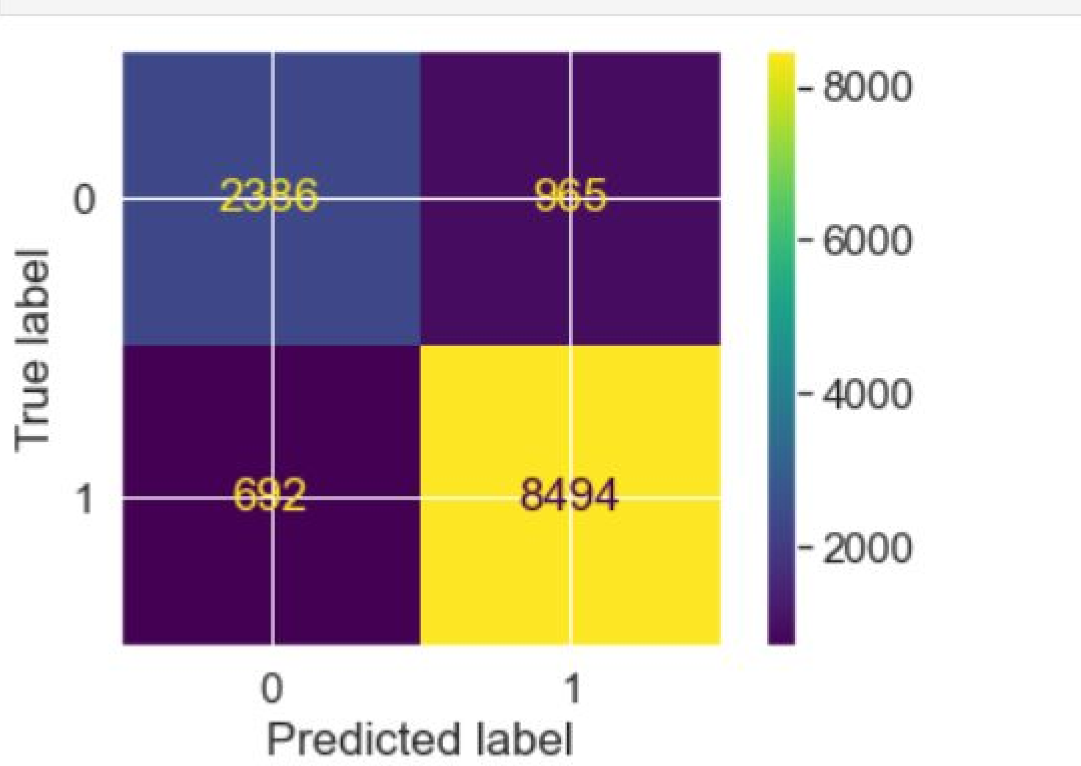
Random Forest Classifier

Best Score:0.8672927393907617

Best Parameters: {'criterion': 'gini', 'max_depth': 5, 'n_estimators': 10}

Model Findings

Below is the Confusion Matrix indicating the Positive and negative class identifications performed by our best model (Logistic Regression)


True label
Predicted label