Machine Learning In Depth Analysis

In this project, I am going to compare different classification models with each other to build a best performing model to predict the credit card default. Firstly, I am going to handle class imbalance present in the dataset using different resampling methods. Then, I will find the best method of handling class imbalance for the dataset we have. Then, I will use that method to handle the class imbalance. The new data set will be used to train different models.

First, I created dummies features for all the categorical data from the dataset. All the categorical data must be converted into machine readable language which is numerical. Then, our classification model will be able to understand and learn from that dummy feature. We have 3 categorical features i.e. gender, education, and marital_status. When we convert the categorical features to dummy variable, the features would be subdivided into its category and given 1 or 0 values for yes and no. The status column for different months had values less than 0 like -1,-2,-3 and those numbers were changed to 0. -1 represented that the client paid in time. -2 represented that the client paid 1 months early. -3 represented that the client paid 2 months in advance.

Then, I imported all the necessary packages that i would be using for the machine learning analysis. Then, I divided the dataset into input 'X' and label 'y'. By the help or using input 'X', we will find out, predict, or classify the label 'y'. Then I scaled the input features of X using robust scaler because or dataset contains some outlier values in paid and balance column. Then, I divided the X and y into training and testing data set. The training set is 70% of our total data and training set is 30% of the data. Our dataset has class imbalance which needs to be taken cared to build a better model which would not over/underfit when learning for classification. We need to handle class imbalance before we start building different model. I used different methods of handling class imbalance techniques to find the best method for our data set. I tried balancing the class weight, undersampling and oversampling, and ensemble method. Balancing the class weight will add weight to the lower class by copying the data from the lower class to balance the weight of the both class. Undersampling will ignore or not include some data from the upper class in our dataset to balance with the lower class. Oversampling will copy the data from the lower class and add them to our dataset to balance with upper class. Ensemble method will combine two techniques together to handle the class imbalance.

I build a logistic regression model as our base classification model. Then we used different methods of handling class imbalance techniques on our dataset. I calculated different scoring metrics to measure the performance of our model. I stored the score of different class imbalance handling methods to compare which method was the best. I used F1 score to measure the performance of our model because we have class imbalance in our dataset. Higher the f1 score, better would be our model.

Approach 1

The ensemble method SMOTEEEN which is a combination of SMOTE and Edited Nearest Neighbour method had the best f1 score. According to the F1 Score of the classification model, the SMOTE+ENN method to handle the class imbalance performed the best. So we will use SMOTE+ENN to handle class imbalance, then compare different models with each other.

Now, I am using SMOTEEEN as the resampling method and building different models with the new resampled dataset. I build Logistic Regression, Decision Tree, GaussianNB, Random Forest, and Support Vector Machine classifier. Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default. Let's try to tune all our model to see if we can improve the performance and then decide which classification model would be the best.

Hyperparameter is a user provided argument or input which effects the performance of the model. Hyperparameter tuning is done by providing a set of hyperparameter values for the machine learning model, and the model will choose its best parameter which achieves the best performance score. I used GridSearchCV and RandomizedSearchCV to perform the Hyperparameter Tuning. I used 5 fold cross validation to find the best f1 score for the all the model. After tuning the hyperparameter of all different model, we can see that the f1 score for Decision Tree had decreased after hyperparameter tuning, and the f1 score for Random Forest has improved and increased. Logistic Regression, GaussianNB, and SVM has not changed even after tuning. Logistic Regression has the best f1 score compared to all other classification model, so I will choose Logistic Regression for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Logistic Regression model to predict credit card default.

Approach 2

For approach 2, the Edited Nearest Neighbour method to handle the class imbalance performed the best according to the F1 score, so I used ENN to handle class imbalance, then compare different models with each other. Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

Approach 3

According to the F1 Score of the classification model, the Edited Nearest Neighbour method to handle the class imbalance performed the best, so I used ENN to handle class imbalance for approach 3, then compared different models with each other. Random Forest has the best f1 score compared to all other classification model, so I will choose Random Forest for this project. Since the F1-score is significantly better than a random classifier, I would recommend this Random Forest model to predict credit card default.

The highest f1 score is 56% for approach 2 Random Forest Classifier, followed by 54% for approach 1 Logistic Regression Classifier, and 53% for approach 3 Random Forest Classifier.