Understanding Patient Persistency Model selection

Name: Data Science :: Healthcare - Persistency of a drug:: Group Project

Internship Batch: LISUM30:30 Jan24 - 30 Apr 24

Data intake by: David Paul Nalumenya

Specialization: Data science

Email: nalumenyad@gmail.com

Standardization and normalization

- Prior to model building, the standard scaler was used to scale the data variables within the 0,1 scale.
- This was a classification problem, therefore classification models were more appropriated;
  - ➢ Logistic regression model
  - ➢ Support vector machine
  - ➢ Random forest
  - ➢ Decision trees
  - ➢ Xgboosting

SMOTE() was used to handle outcome variable class imbalance, this would affect model quality and fit, this method is more preferable if the dataset is smaller, and fewer categories in the feature are present, in this case, they were two

- The Persistence_flag was the outcome variable, with rest of 67 variables as predictor variables
- Model accuracy, F1, precision and recall, AUC/ROC and feature importance for each model were reported, models were compared and best model was chosen basing on the criteria above.

Model selection and comparison before correcting class imbalance

- Xgboost model performed well on data where the class imbalance was not corrected with SMOTE, the reported accuracy was **0.8145985401459854, better than the rest of the models.**
- Logistic Regression Accuracy: 0.7941605839416058
- Random Forest Accuracy: 0.7985401459854015
- Decision Tree Accuracy: 0.7357664233576642
- SVM Accuracy: 0.7795620437956204

- Random forest and xgboost model had the highest accuracy compared to the other models.

Model selection and comparison after correcting class imbalance

After correcting class imbalance;

- Logistic Regression Accuracy: 0.7775175644028103
- Random Forest Accuracy: 0.8290398126463701
- Decision Tree Accuracy: 0.7353629976580797
- SVM Accuracy: 0.7681498829039812
- XGBoost Accuracy: 0.8138173302107728

- **Two models, random forest, and Xgboost performed well, they gave the highest accuracy.**
- **However its important to note that, after correcting class imbalance, the model accuracy for all the models reduced significantly, suggesting that previously the models were overfit**

Correcting class imbalance and hyperparameter tuning for random forest model

- There after, random forest and xgboost model were taken further for analysis until we choose the best model.
- SMOTE was maintained on the dependent variable, hyperparameter tuning done for both models and results were as below
- For random forest, hyperparameters gotten, max_depth=8, min_samples_leaf=1, min_samples_split=5, n_estimators=200, gave an accuracy of 0.775175644028103, ROC AUC area was 0.86.
- The predicted important features were;

Persistency_Flag,Age_Bucket,Region,Dexa_Freq_During_Rx,Dexa_During_Rx,Ntm_Speciality,Comorb_Long_Term_Current_Drug_Therapy,Comorb_Encounter_For_Screening_For_Malignant_Neoplasms,Comorb_Encounter_For_Immunization,Comorb_Encntr_For_General_Exam_W_O_Complaint,_Susp_Or_Reprtd_Dx.

Correcting class imbalance and hyperparameter tuning for random forest model

- For xgboost, SMOTE was used on the dependent variable, hyperparameter tuning done and results were as below;

- For xgboost, hyperparameters gotten were; learning_rate= 0.1, max_depth = 8, min_child_weight= 1, n_estimators=200.

- These consequently were used to build the model that gave an accuracy of 0.8255269320843092 and a ROC AUC area of 0.89

Model choice

- After exploratory data analysis, hyperparameter tuning and feature engineering, xgboost performed better, it fit the data well compared to the other classification models.

Below is the classification report and model accuracy;

- XGBoost Accuracy: 0.8255269320843092
- AUC = 0.89

XGBoost Classification Report

	Precision	Recall	F1
Non persistent	0.82	0.82	0.82
Persisten	0.83	0.83	0.83