Project 1

Instructions

1. Create a Data Summary section describing everything (showing your subject matter expertise as well). You may need to look up diabetes information relevant to the data.
2. Do a comprehensive EDA (make sure to reference the 7 stories one tells with data)
3. Make sure to do any necessary feature engineering and/or imputation throughout using at least three techniques to test it such as the chi-square and p-values, variance inflation factors, etc. For the imputation, ensure that the distribution is the same and check multiple methods.
4. Make sure to follow the different options you've created
5. Create a logistic regression (Logit) model and return a classification "vector"
6. Create your own evaluation/performance metrics (write your own code for Precision, Accuracy, Recall, Specificity, and F1 score) and test it against the Python library for said metrics. Use Python's crosstab as well.
7. Test hyperparameters throughout.
8. Show the data is imbalanced and use SMOTE and at least one other approach to fix the issue.
9. Retest your logistic regression model with the newly SMOTE data
10. Retest your evaluation/performance metrics and also create a confusion matrix, as well as a ROC-AUC curve
11. You may want to reevaluate any feature engineering
12. Create models using SVM, KNN, and some type of Tree model and compare it to your logistic regression model. Be mindful of hyperparameters. Make sure that your models take into account train/test/split with k-fold cross validation or the like.
13. Create an ensemble model
14. Summarize your approach and discuss the full data science pipeline as well as what you went back and fixed based on your findings and then what your final model includes.