Write up - Assignment 3

Followed the TA’s (Horace) instructions and critique, and created better (and cleaner) functions for dealing with reading data, preprocessing and cleaning data, and exploring data. Also added proper functions for generating features and predictors by discretization and creating dummies.

This write-up is on the data used in the previous assignment (assignment 2) so credit data has been used, and analyzed.

Comparison of different metrics:

I performed the regression on Random Forest, Logistic Regression, Decision trees, and KNN. I have written the functions to be able to use other models as well, but did not include those in the analysis due to high processing time and low computational capability.

I have used different metrics for evaluation:

Accuracy, auc-roc, auc-pr, (precision, recall, f1 == at 1,2,5,10,20,30,50%).

For accuracy, the following was most accurate:

**accuracy**

|  |  |  |
| --- | --- | --- |
| RF | {'max\_depth': 50, 'max\_features': 'log2', 'min\_samples\_split': 10, 'n\_estimators': 100, 'n\_jobs': -1} | 0.8774522292993631 |

In terms of auc-roc, the following was highest:

LR {‘C’:10, ‘penalty’:’l1’

In terms of auc-pr, the following was highest:

|  |  |
| --- | --- |
| LR | {'C': 1e-05, 'penalty': 'l1'} |

The lowest time taken was by:

|  |  |
| --- | --- |
| DT | {'criterion': 'gini', 'max\_depth': 1, 'min\_samples\_split': 10} |

The results for precision, recall, and F1 vary depending on the percentage, and can be gleaned from the output table.

My recommendation depends on the goals of the organization I am giving the recommendation to. If the goal is to minimize false positives and make sure that a large majority of people who are identified are positive, then I would recommend high precision models such as:

|  |  |
| --- | --- |
| RF | {'max\_depth': 50, 'max\_features': 'sqrt', 'min\_samples\_split': 10, 'n\_estimators': 10, 'n\_jobs': -1} |

However, if the goal is to minimize the number of false negatives and make sure that as few people as possible are oversighted, I would recommend a model with high recall, such as

|  |  |
| --- | --- |
| DT | {'criterion': 'gini', 'max\_depth': 1, 'min\_samples\_split': 10} |

For those proceeding on models with this data, I’d also recommend looking at the distribution of the data. For example, a disproportionately large number of positives belong in the 30-50 age bracket.

It’s also important to notice the long tail in Debt Ratio for SeriousDlqin2yrs positives, as well as a high peak and skewed tail for Monthly Income. There are also some zipcodes where positives happen more than others and that can also be an important factor.

Some limitations of this recommendation involve the data collection and gathering process. We do not know of the degree to which any systematic biases exist in the data (though we see some in our data exploration), and so would recommend a further assessment of the validity of the data for issues involving data entry or other factors. It is also important to note that there might be factors which might be unknown such as family situation, or level of education etc. Having this information could possibly add to the model. We also need to be aware of any systemic biases we might create in this model that might unintentionally discriminate against certain groups based on age, race etc.

Lastly, it is important to remember what the policy goal is that the organization wants to achieve, and then look at the metrics accordingly. All the data exploration graphs and other info is available in the code for further insight.