DA2_Homework_Classification_2

Employee turnover

We want to determine based whether or not an employee will leave *(i.e., turn)*. We'll build 2 models: a glm binomial regression, and we'll also build an logistic equation and compare results to glm. The dimensions we want to consider in the model are:

- Satisfaction *(latest emp survey)*
- Last_Eval *(lastest evaluation)*
- Number_Projects *(average number of project per month)*
- Avg_Mo_Hrs *(average hours per month)*
- Tenure *(years with company)*
- Promotion *(promotion recieved in last year)*

First check the balance between 0 *(did not leave)* and 1 *(left)*

```
##
##         0         1
## 0.7619175 0.2380825
```

It's impbalanced enough to consider runing SMOTE *(remember, the response variable needs to be a factor to run SMOTE, and needs to be numeric (0,1) to apply logistic regression analysis, so prepare to transform between operations)*

After applying SMOTE, the balance should be near the following *(play with the perc.over and under to get what you want)*:

```
##
##         0         1
## 0.4936547 0.5063453
```

Now run glm to estimate your coefficients *(you want to use the smote data to train the model, but retain the orginal data for pulling testsets - keep the datatypes in sync!)*.

```
##     (Intercept)    Satisfaction       Last_Eval Number_Projects      Avg_Mo_Hrs
##     0.477977877    -4.519804184     1.343789948    -0.478094011     0.005153499
##          Tenure       Promotion
##     0.518688334    -2.389893336
```

Now, that you have coefficients, create a test file with 100 records *(just use sample_n, 100 on the original data)*. Using the glm coefficients, build a logistic regression equation, and calcuate probabilities *(write these to the test dataframe)*. Just for confidence, also run the test data through the glm fitted model and compare to your equation results to make sure all agree.

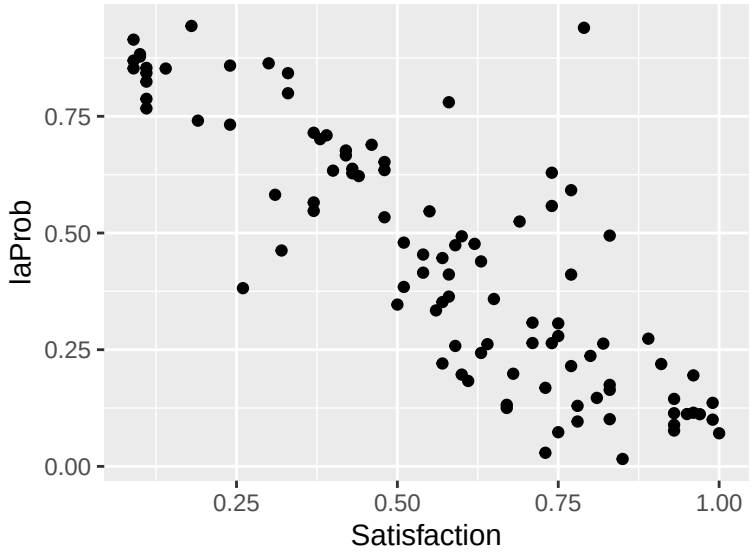
Now set all the records with a probabiliy over 50% to 1 *(Left)*, and use a confusion Matrix to score.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 56  3
##          1 17 24
##
```

```
##                Accuracy : 0.8
##                  95% CI : (0.7082, 0.8733)
##     No Information Rate : 0.73
##     P-Value [Acc > NIR] : 0.06843
##
##                   Kappa : 0.5639
##
##  Mcnemar's Test P-Value : 0.00365
##
##             Sensitivity : 0.7671
##             Specificity : 0.8889
##          Pos Pred Value : 0.9492
##          Neg Pred Value : 0.5854
##              Prevalence : 0.7300
##          Detection Rate : 0.5600
##    Detection Prevalence : 0.5900
##       Balanced Accuracy : 0.8280
##
##        'Positive' Class : 0
##
```

An accuracy score $> 75\%$ is fine *(there are ways we can impove this which we'll study later)*.

Finally, show the relationship between Employee Satisfaction and whether they left or not.



Classification with Random Forest

Set up your