Predictive modelling

George, John & William

11/17/2021

Data Preparation and cleaning

```
##   Number_Missing Missing_Rate        Variable
## 1           7087     56.53318      enrollment
## 2           7087     56.53318      employment
## 3           9448     75.36694 employment_type
## 4           9448     75.36694 weekly_work_hrs
## 5           7361     58.71889       ethnicity
## 6           7383     58.89438          gender
```

19 columns with 5095 observations

missing value treatment

```
## [1] 3415   18
```

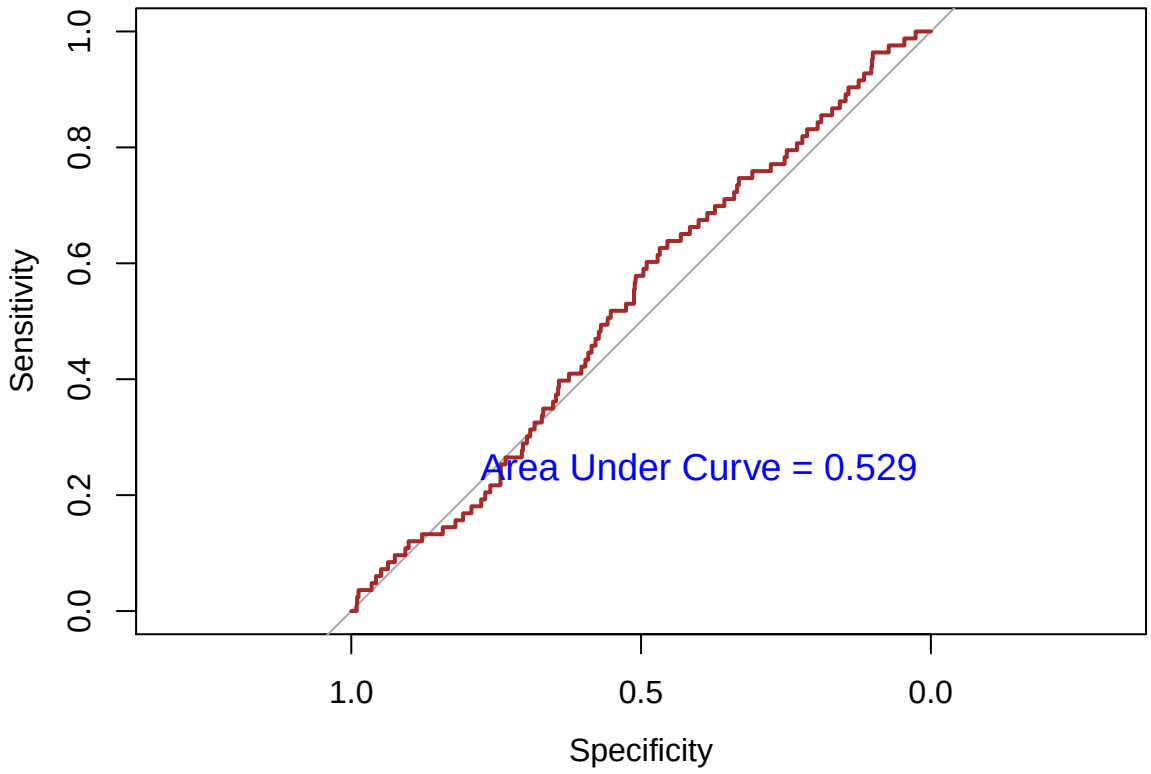
```
## [1] 1680   18
```

The training data has 76 observations with 1887 now (old =1057 when compared) variables. The testing data has 32 observation with 1887 now (old= 1057 when compared) variables.

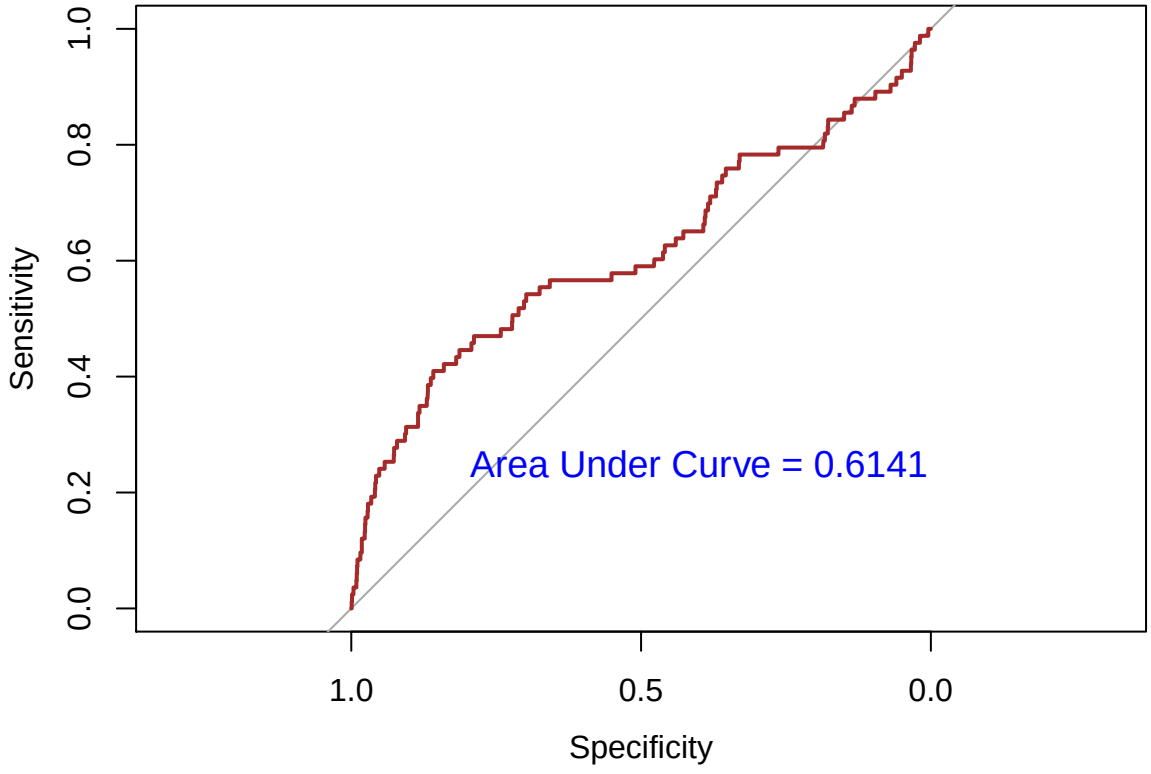
SVM with kernlab package with tuned parameters

Linear SVM through kernlab

```
##  Setting default kernel parameters
```

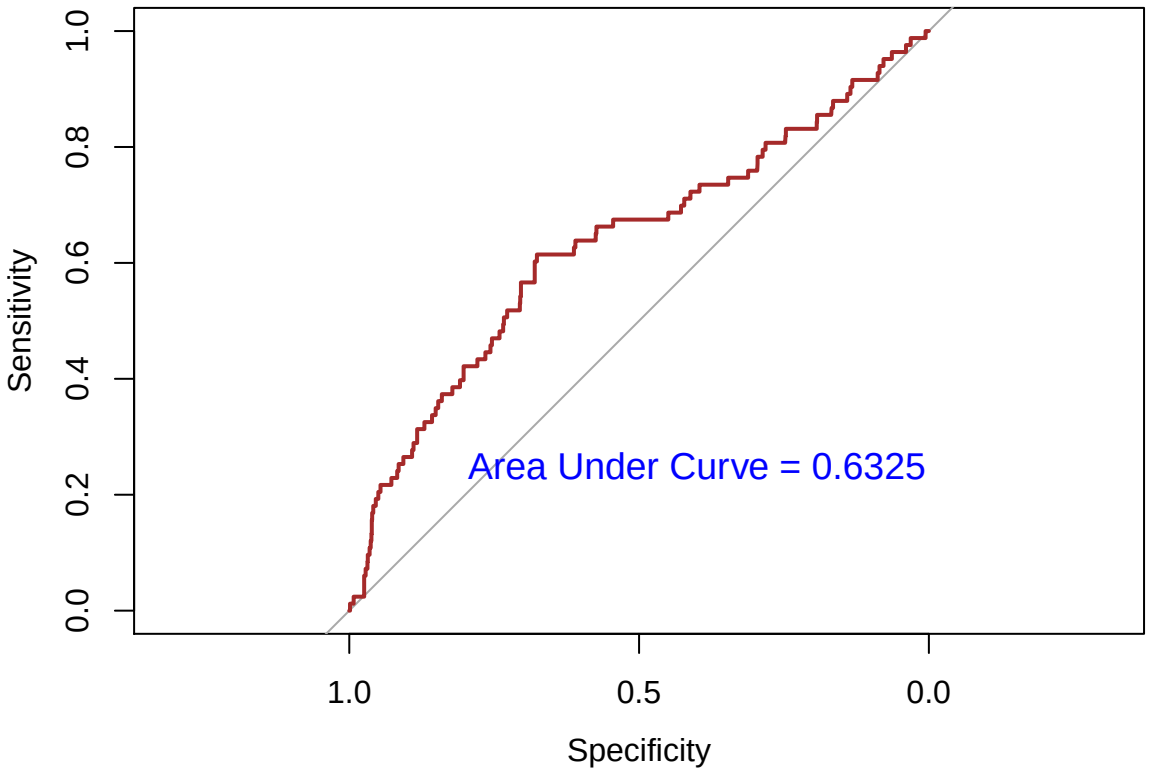


Computing SVM using radial basis kernel (rbfdot)

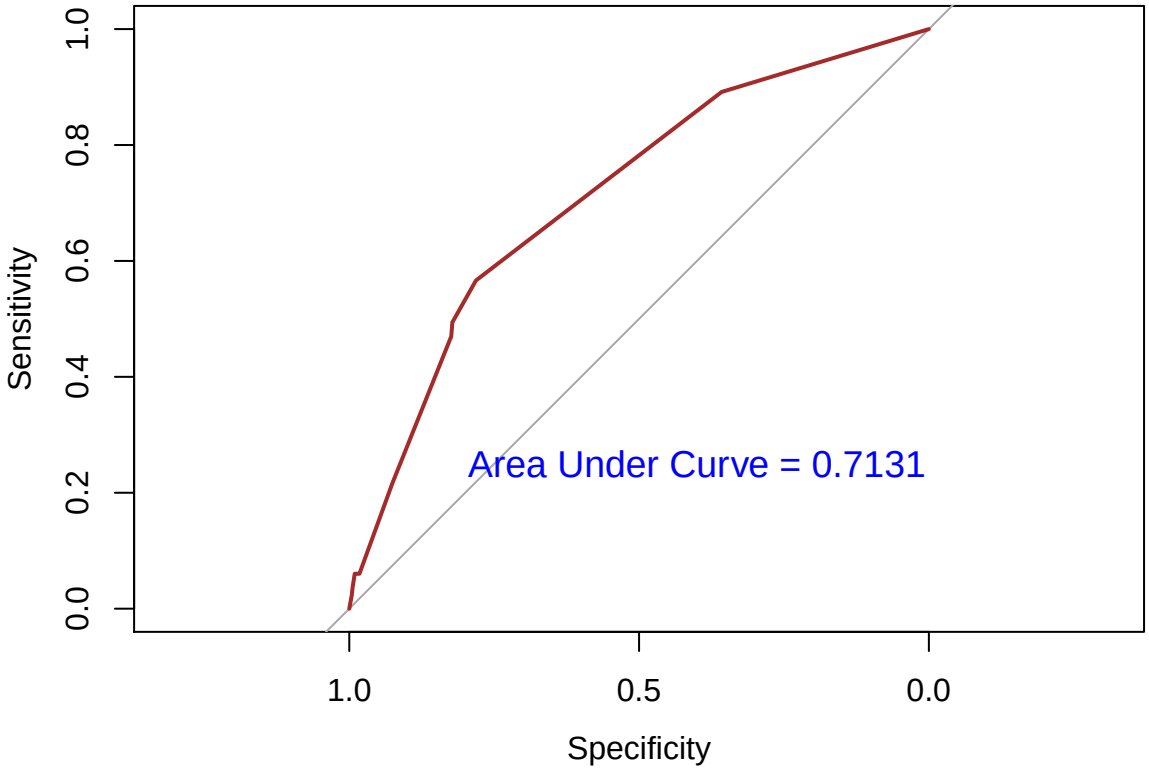


Computing SVM using polynomial basis kernel

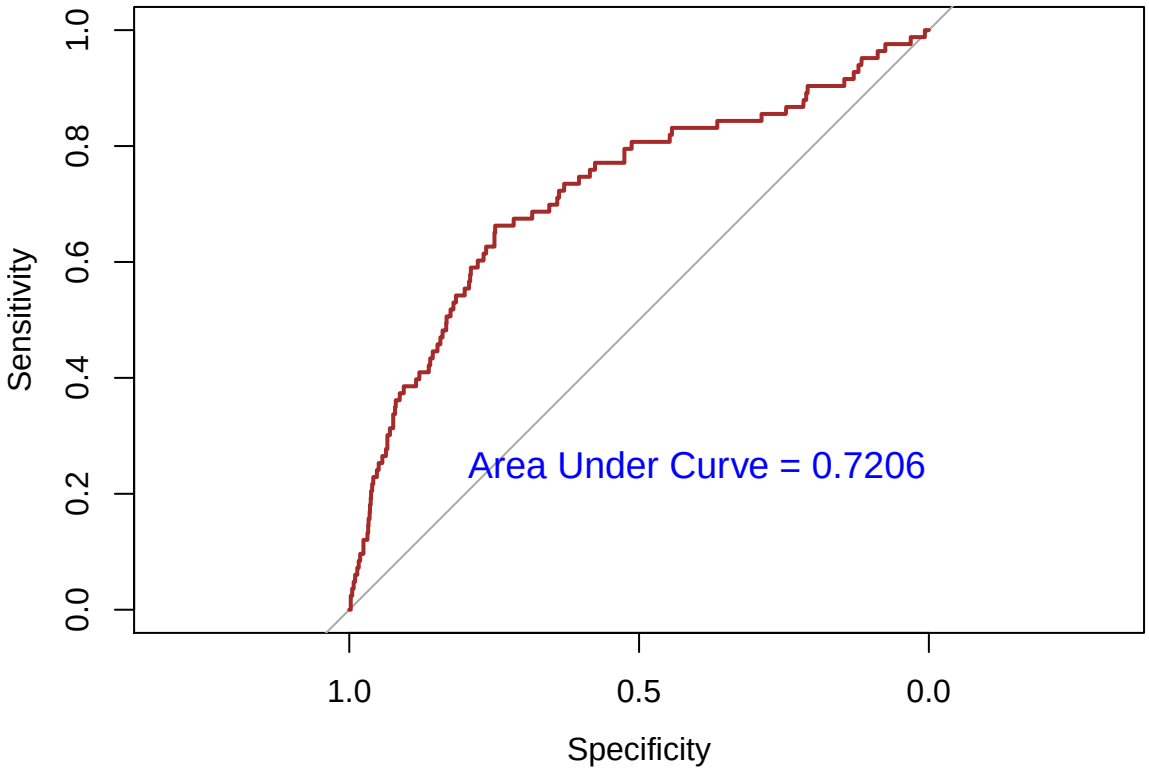
```
##  Setting default kernel parameters
## maximum number of iterations reached 0.002164177 0.002088773
```



Classification through Decision Trees



Classification through Naive Bayes



Comparison between the kernels base on their AUC and Misclassification

Methods	AUC	Missclassification
SVM_Linear	0.529	4.94
SVM_Radial	0.6141	4.94
SVM_Polynomial	0.6325	4.94
Decision Trees	0.7131	5.3
Naive Bayes	0.7206	7.56

Modelling for Spend night elsewhere

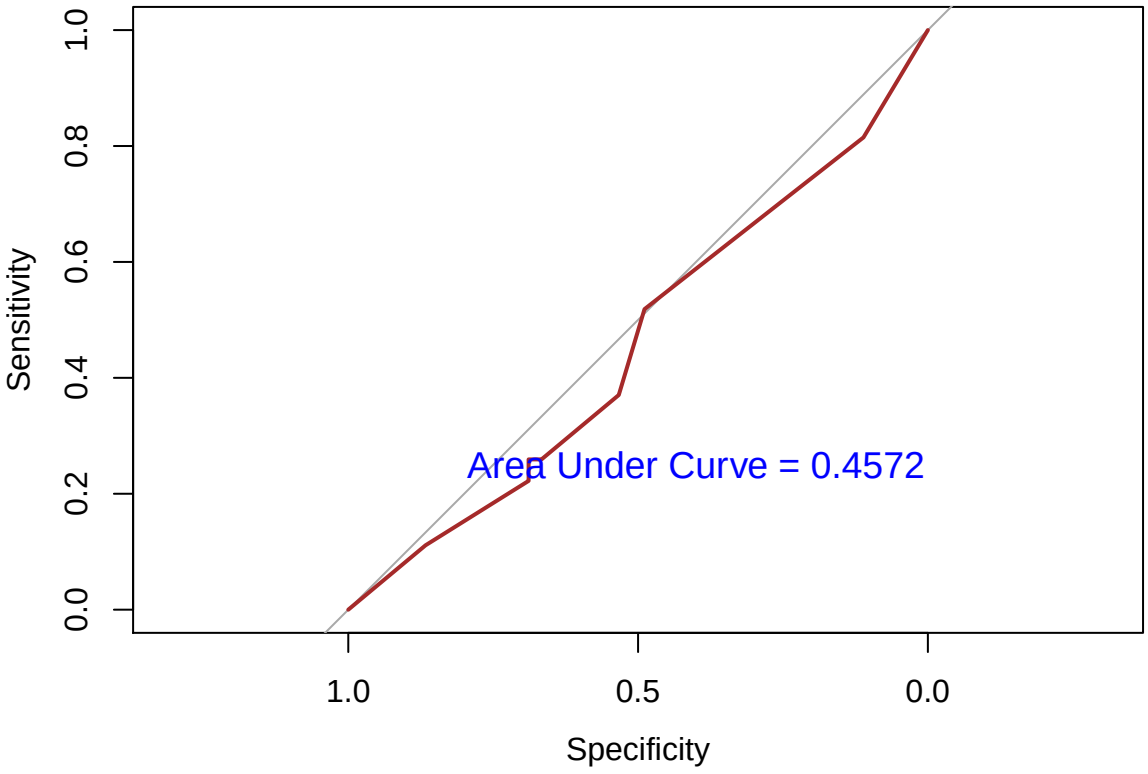
missing value treatment

```
## [1] 172  18
```

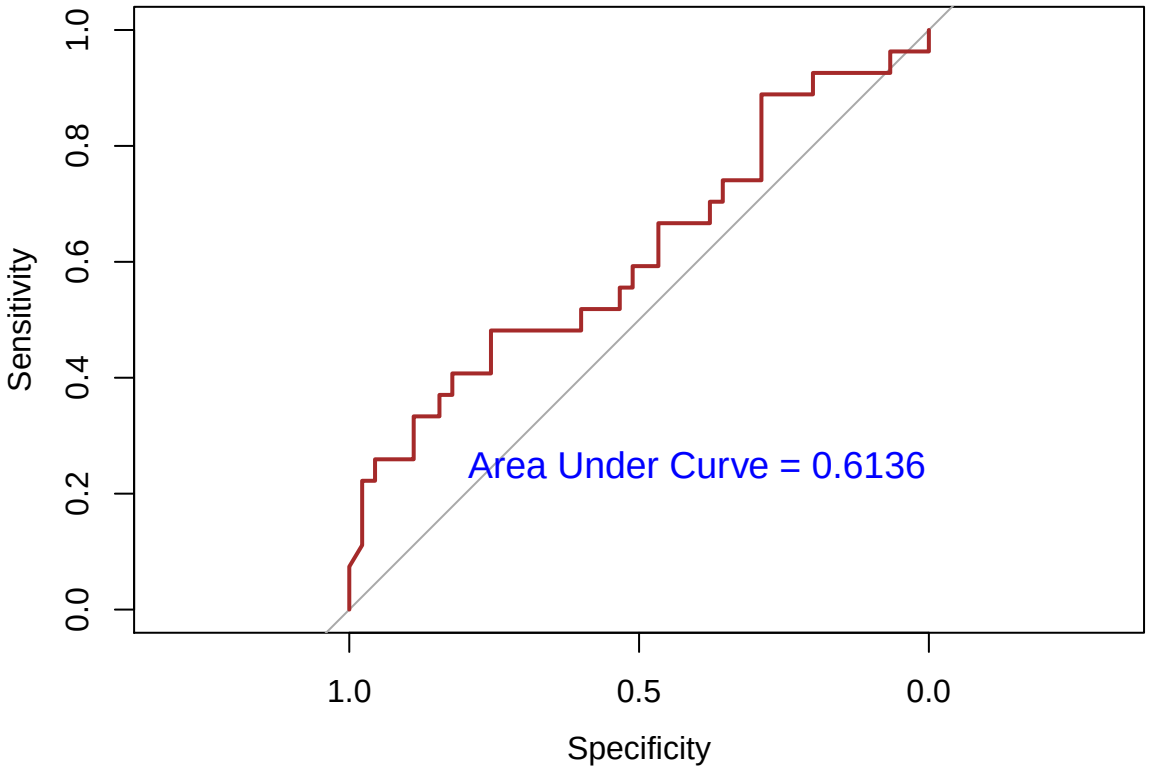
```
## [1] 82 18
```

Due to the three levels SVM is not appropriate, hence

Classification through Decision Trees



Classification through Naive Bayes



Comparison between the kernels base on their AUC and Misclassification

Methods	AUC	Missclassification
Decision Trees	0.4572	54.88
Naive Bayes	0.6136	54.88