Effectiveness of Social Media Ads.

GAM - Classification | Assignment - 5 | Pattern Recognition

RAMIT NANDI || MD2211

2023-11-15

Contents

DATA **2**

GAM in R using mgcv **5**

Full Model 5

Simplified Model after variable selection 7

Comparison with GLM (Logistic Regression) **7**

Accuracy & Confusion Matrix **8**

Conclusion **8**

DATA

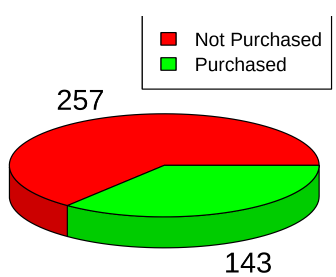
The dataset contains details of the purchase of a product based on social network advertisements. The data has 400 observations, looks as follows . . .

User.ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0

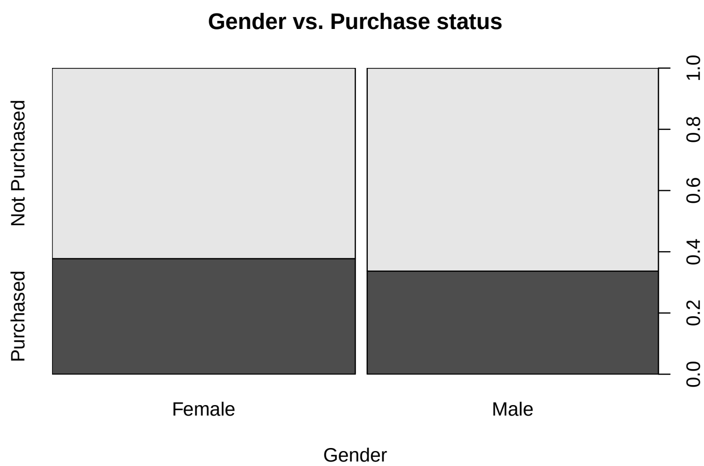
GOAL: Predicting whether a person will buy a product displayed on a social network advertisement based on his/her gender , age and approximate Salary.

EDA

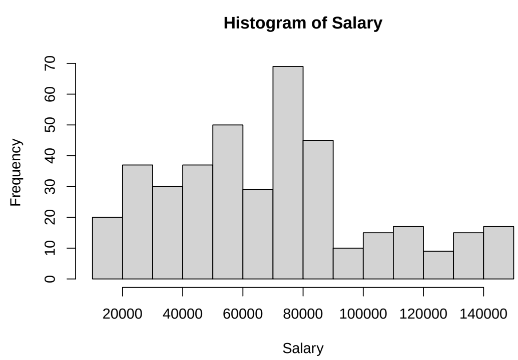
- Target class belongs to two discrete categories of purchased and not purchased. [Throughout the report , Red colour will denote 'Not Purchased' , Green colour will denote 'Purcased']

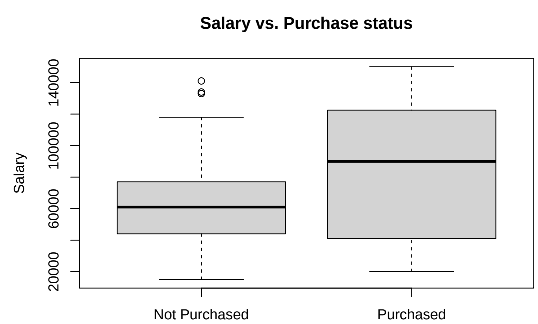


- *Gender:* Gender does not affect the purchase status much here. Given a person is male, the chance that he will buy the product is very similar to the chance of purchase ,given the person is female .

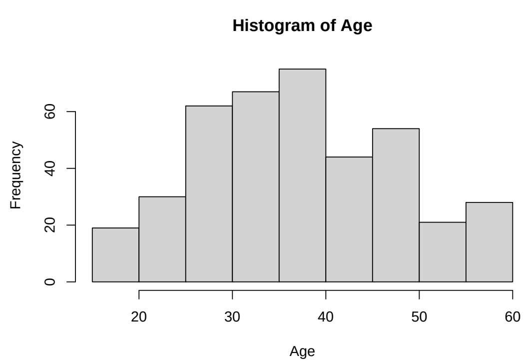


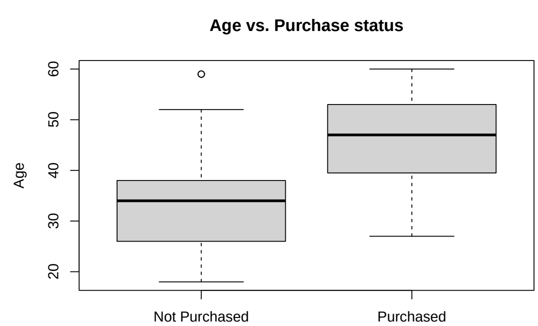
- *Salary:* Those who purchase the product , have on average higher salary than those who do not.





- *Age:* Those who purchase the product , are on average older than those who do not.





NOTE: Based on some earlier analysis on this data , we know -

- The decision boundary is not linear in terms of the predictors.
- The predictor 'Gender' is not that much important.

Let's check what GAM concludes on this data.

```r
# preprocessing .............
Data$Gender = factor(Data$Gender)
Scaled_ = scale(Data[c("Age","EstimatedSalary")])
Data[c("Age","EstimatedSalary")] = Scaled_
# 80-20 Train-Test split
split_ix = caTools::sample.split(Data$Purchased,0.8)
TRAIN = subset(Data,split_ix==T)
TEST = subset(Data,split_ix==F)
```

GAM in R using mgcv

Full Model

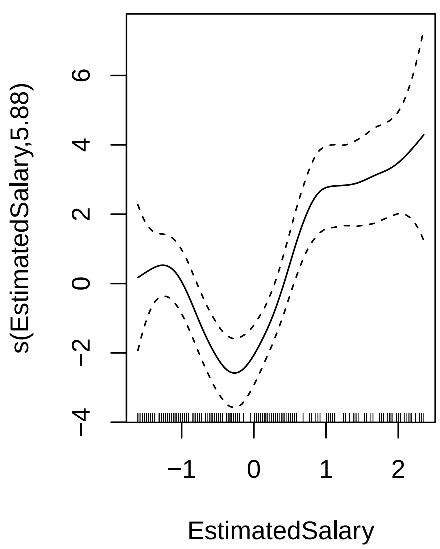
We start our model with cubic splines on 'Age' and 'Salary' , the categorical predictor 'Gender' and its interaction with 'Age' and 'Salary'.

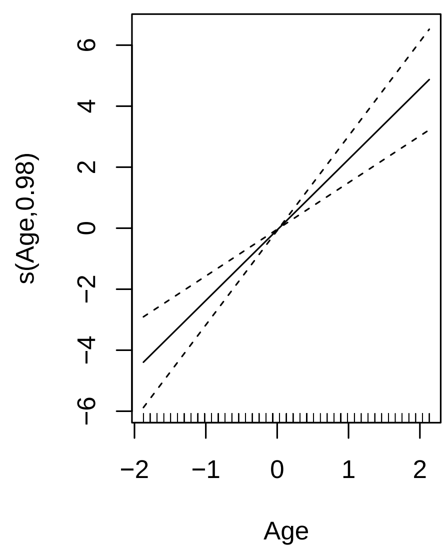
```
full_GAM = gam(Purchased ~
                 s(EstimatedSalary,bs='cr',k=20) +
                 s(Age,bs='cr',k=20) +
                 Gender +
                 s(EstimatedSalary,by=Gender) + s(Age,by=Gender),
               data = TRAIN, family = binomial,
               method = 'REML', select = TRUE)


summary(full_GAM)
```

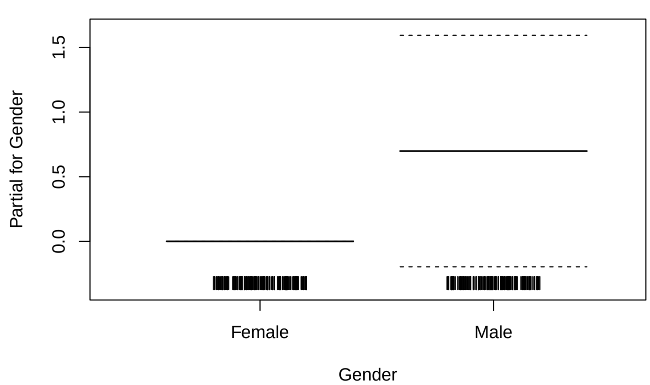
```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Purchased ~ s(EstimatedSalary, bs = "cr", k = 20) + s(Age, bs = "cr",
##     k = 20) + Gender + s(EstimatedSalary, by = Gender) + s(Age,
##     by = Gender)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6016     0.3437  -4.660 3.16e-06 ***
## GenderMale    0.6986     0.4477   1.561    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                    edf Ref.df Chi.sq p-value
## s(EstimatedSalary)           5.884e+00     19 62.909  <2e-16 ***
## s(Age)                       9.761e-01     11 34.930  <2e-16 ***
## s(EstimatedSalary):GenderFemale 4.127e-05   9  0.000   0.458
## s(EstimatedSalary):GenderMale   2.126e-01   9  0.237   0.282
## s(Age):GenderFemale          2.816e-05      9  0.000   0.340
## s(Age):GenderMale            6.148e-01      9  1.534   0.105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.736   Deviance explained = 66.4%
## -REML = 85.181  Scale est. = 1         n = 320
```

Interpretation: Based on the partial effects plots (i.e. effect of a variable on the response , given the other variables are fixed) we see -

- Salary has quite non-linear effect.
- The effect of Age is almost linear.
- We do not see significant effect of Gender, as the confidence interval for estimated effect of Gender-Male contains the baseline 0.







Simplified Model after variable selection

Based on the partial effects plots as well as the p-values in model summary above, we get -

- 'Age' & 'Salary' as significant predictors.
- The 'Gender' is not significant, so it is discarded along with its interaction terms.
- Also we replace spline of 'Age' with just a linear term.

```
Model = gam(Purchased ~
                 s(EstimatedSalary,bs='cr',k=20) +
                 Age ,
               data = TRAIN, family = binomial,
               method = 'REML', select = TRUE)
```

Comparing AIC and BIC with full-model, we see the simple one is actually better one here (rule of thumb: >2unit difference in AIC or BIC is considered as statistically significant, lower the AIC or BIC better the model).

	df	AIC
full_GAM	11.570960	163.0930
Model	8.606467	164.4732

	df	BIC
full_GAM	11.570960	206.6961
Model	8.606467	196.9052

Hence, our simplified model is accepted.

Comparison with GLM (Logistic Regression)

```
Model_GLM = glm(Purchased ~ EstimatedSalary + Age + Gender,
               data = TRAIN, family = binomial)
```

	df	AIC
Model	8.606467	164.4732
Model_GLM	4.000000	216.8761

	df	BIC
Model	8.606467	196.9052
Model_GLM	4.000000	231.9494

Clearly, GAM performs much better than GLM, though it uses more df, so more complex model and more computation involved.

Accuracy & Confusion Matrix

On Train Data

```
##
##                          true Not Purchased true Purchased
##   pred. Not Purchased                  194             13
##   pred. Purchased                       12            101
```

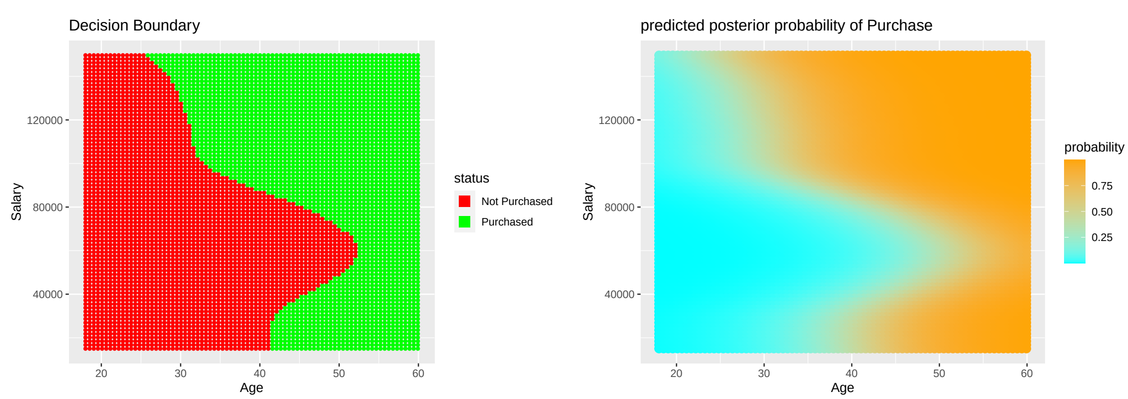
```
## Accuracy:  0.921875
```

On Test Data

```
##
##                          true Not Purchased true Purchased
##   pred. Not Purchased                   47              8
##   pred. Purchased                        4             21
```

```
## Accuracy:  0.85
```

Conclusion



GAM classification works well on this dataset.

Based on the decision boundary we can say - Those with high salary are likely to puchase the product. Even with low salary, Young peoples are likely to purchase the product based on social-media ads.

For time constrain here we used default hyperparameters for penalty coefficients and used only one 80-20 split to check model performance instead of k-folds. With those adjustments may me the model could be improved or generalized better.