**GRoup - 01**

**Laxmikanta Gauda**

**Ambili Jayachandran**

**Maya Vasudevan K**

**Tarun Kamal Khiani**

**M Sarath**

**M Sarath tarun kamal khiani**

Data Mining

Assignment – DEv & Hold out sample

Contents

[Goal 2](#_Toc485355778)

[Exploratory data analysis 2](#_Toc485355779)

[Conclusion 3](#_Toc485355780)

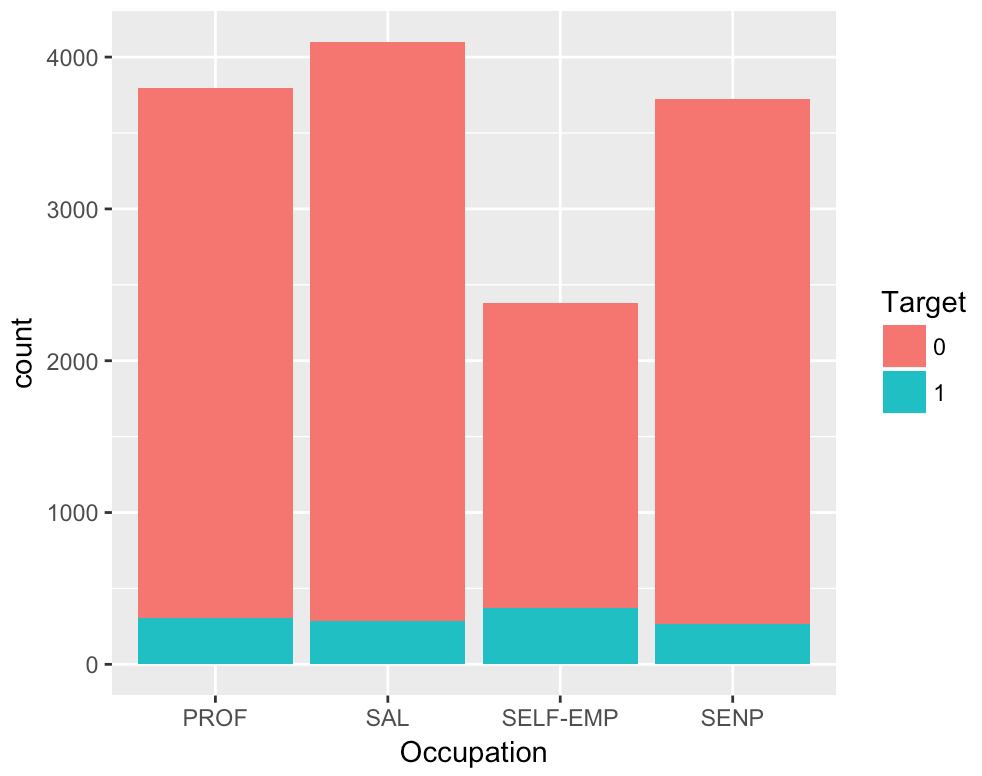
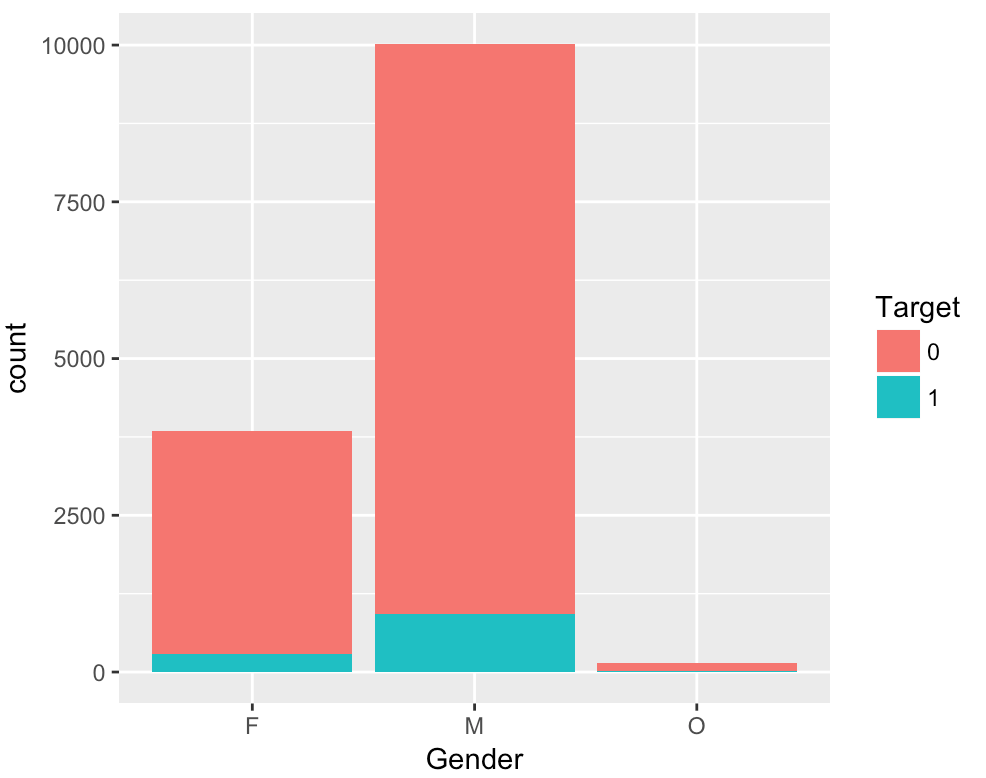
# Goal

The goal of this assignment is to build a model to predict if a customer will opt for a scheme or not.

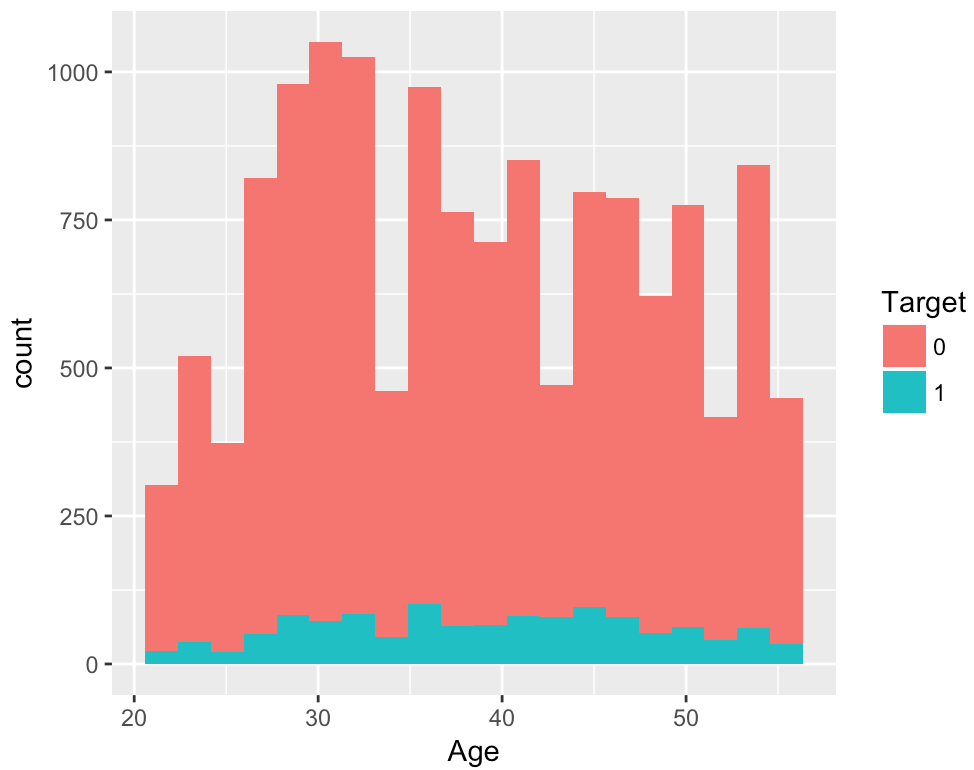
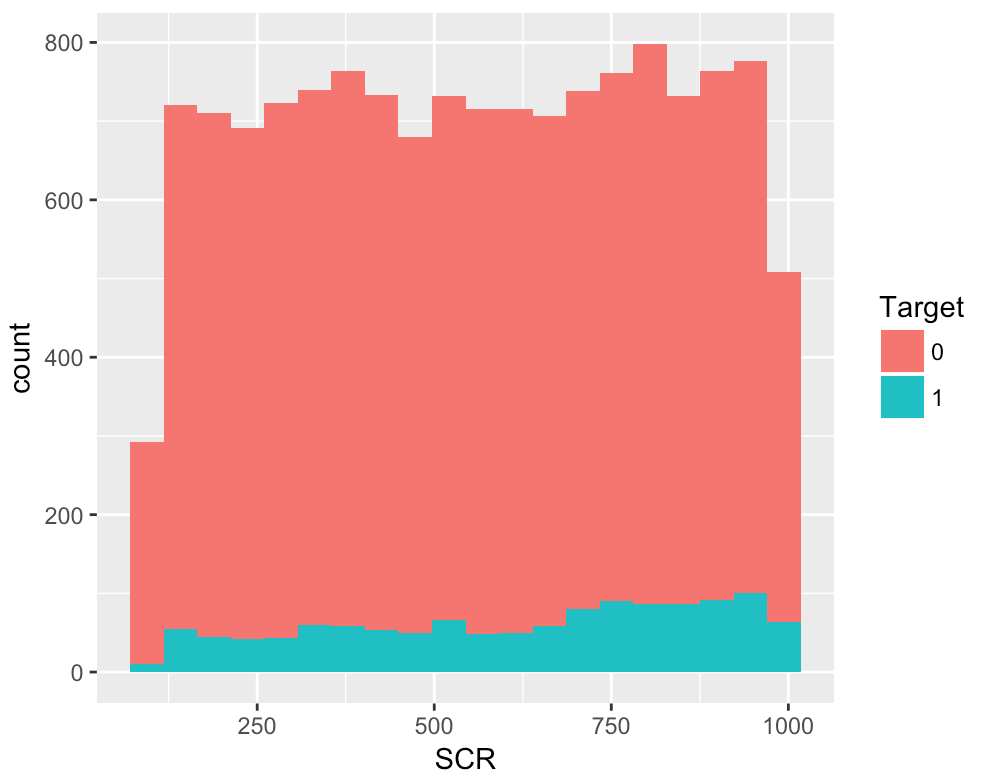
# Exploratory data analysis

Important Observations:

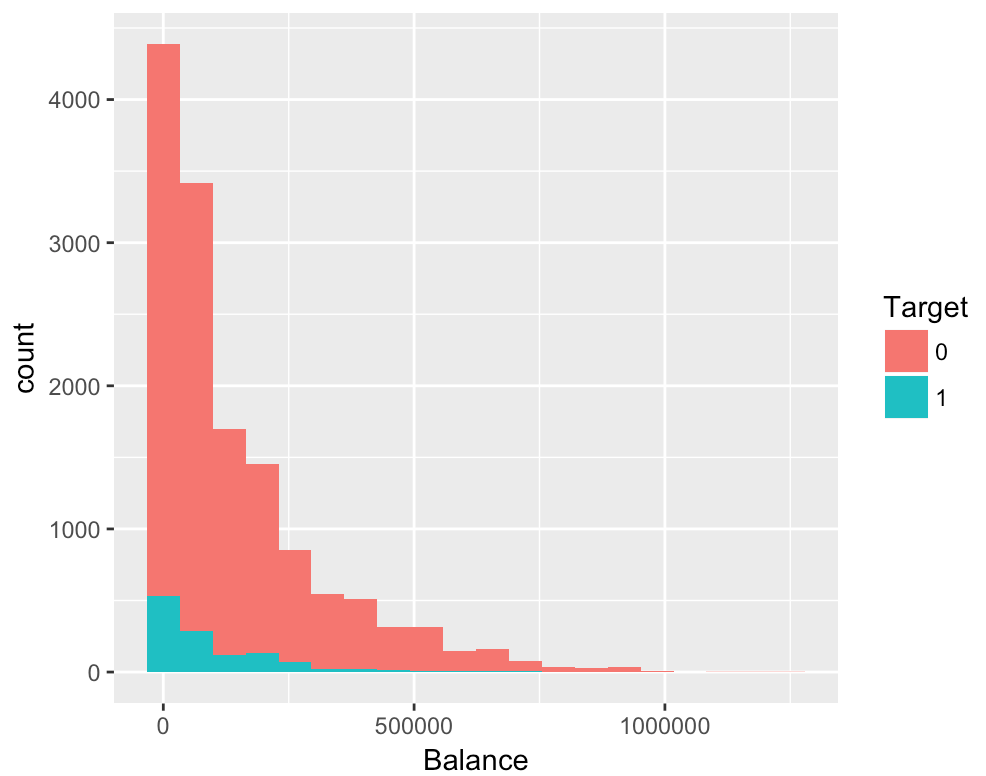
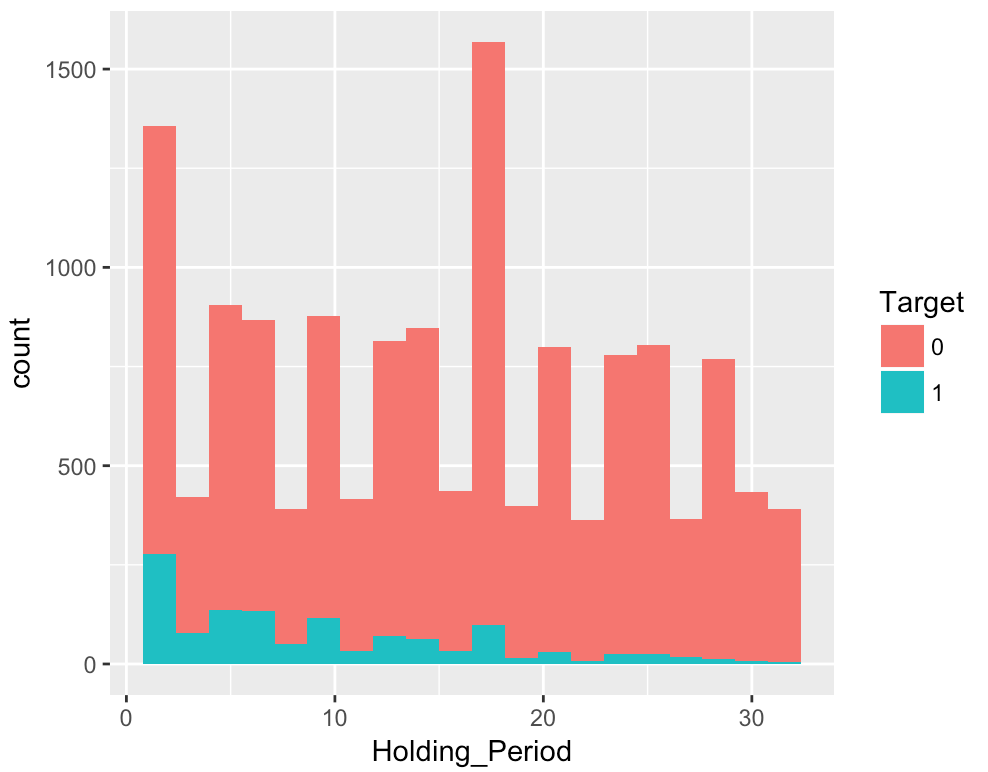
* The data quality is good as there are no null or empty fields.
* AGE\_BKT is similar to Age hence not needed for classification.
* Customer ID is unique for every customer and hence not needed for classification.
* Self Employed have a higher success rate and more males who have a success rate of 1 but proportion wise gender does not make a difference



* Success Rate is higher where the SCR is between 750 – 1000 & Age is between 35 – 45



* Success Rate is higher if the holding period is between 0 – 10 whereas success rate does not seem to be dependent on Balance



* Success Rate is 8.82% which is not good for classification hence there is need to oversample or under sample. We have decided to under sample the customers who do not opt for the scheme.
* Variances of the indices are widely different. Hence scaling of variables is necessary for further analysis.

# Conclusion

**Classification Accuracy on the development sample**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| CART | 84.95% | 15.14% | 98.46% | 69.15% |
| Random Forest | 85.66% | 22.75% | 97.83% | 78.13% |
| Neural Network | 66.23% | 16.84% | 75.79% | 53.90% |

**Classification Accuracy on the hold out sample**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification Model** | **Accuracy** | **Sensitivity** | **Specificity** | **AUC** |
| CART | 90.95% | 13.25% | 97.98% | 68.54% |
| Random Forest | 90.86% | 19.67% | 97.34% | 75.77% |
| Neural Network | 90.03% | 27.10% | 95.72% | 75.01% |
| Ensemble (Combination of the above models) | 88.46% | 32.12% | 93.56% | 75.54% |

The random forest has the best classification performance in the development sample

All the three models perform well on the hold out sample. Surprisingly the neural network classification performance is better than the CART but the Random Forest performs better.

We would recommend the Random Forest Model over the neural network model or the CART model.

We can also use an ensemble model as its performance is far superior on all aspects.