Data Mining - Kaggle Competition
Banking Dataset

Stat 642 – 674 : Yamini Rege

Agenda

- Data understanding
- Data preparation
- Modeling
- Evaluation methodology
- Managerial implications

Data Understanding

Introduction

- Banks allocates huge amount of money on campaigning to promote their products.
- Increasing the number of campaigns leads to increase in cost, it also reduces effects on the general public. Therefore, the campaigns should be effective..

Objective

- Increasing efficiency of the campaign by effectively using the data
- Identifying the main factors which will impact the success of a campaign.
- Predict accuracy of customers who will open savings account.
- Recommend strategies to increase campaign efficiency based on predictions.

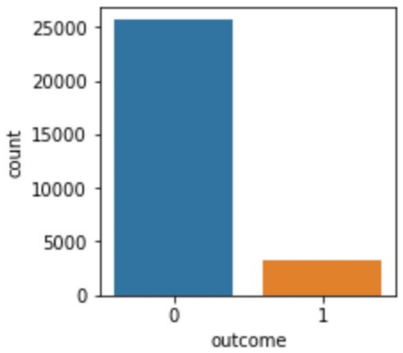
Dataset

28831

Customers

21

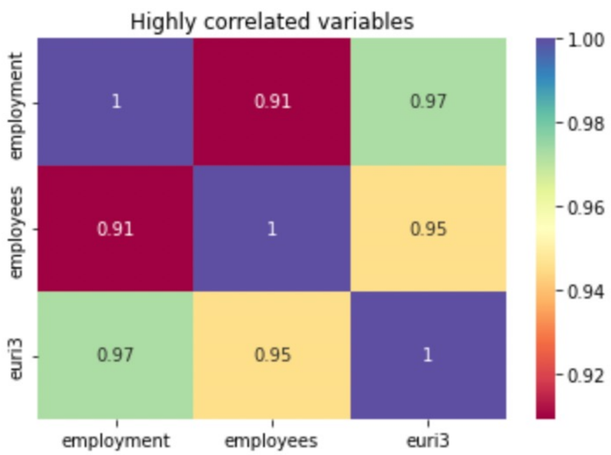
Variables



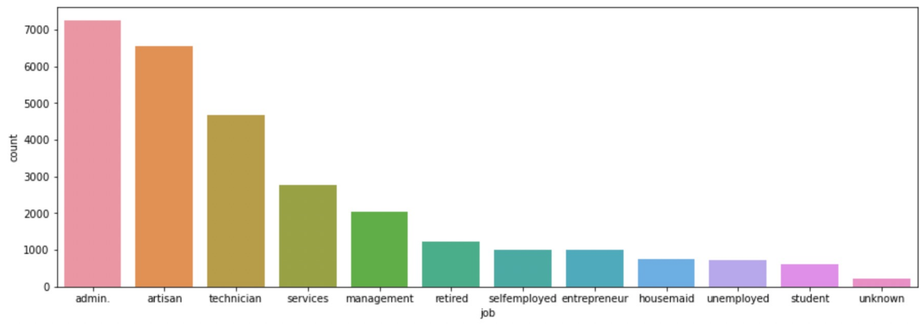
The graphs shows that the data is imbalanced

Correlation of Variables

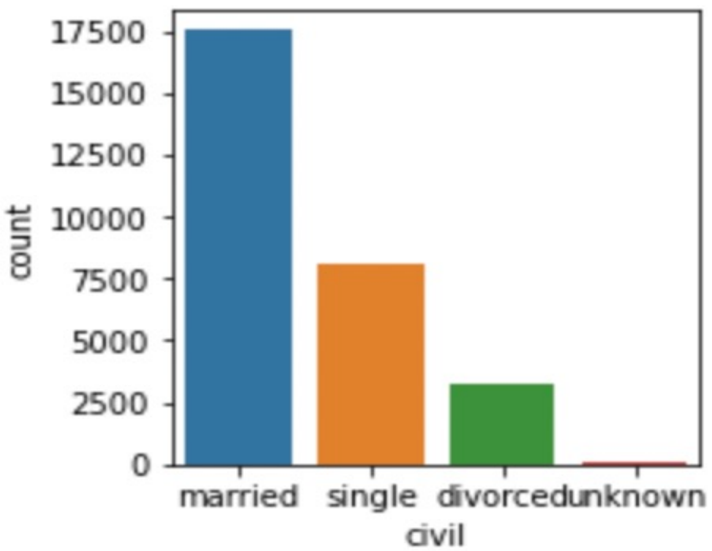
- Employees, Employment and Euri3 are highly correlated variables.



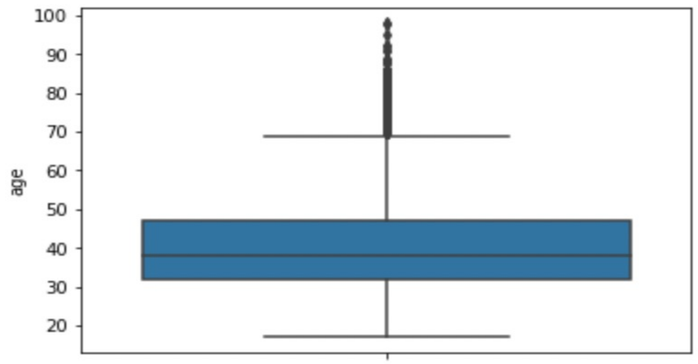
Data Visualisation



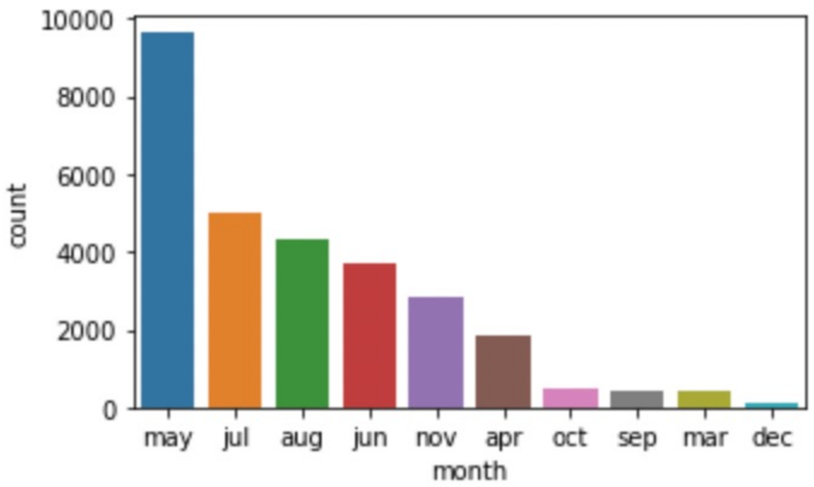
Top 2 types of job in the dataset is admin and artisan.



Out of total customers 61% are married and 28% are single. 10% of married and 13% of single customers have opened savings accounts.

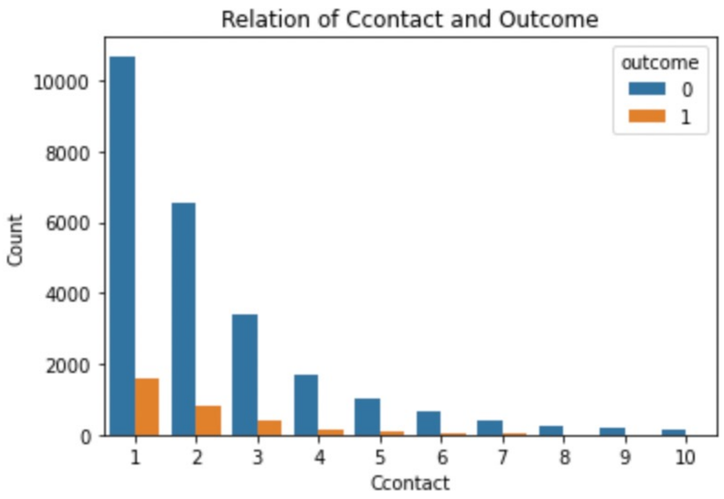


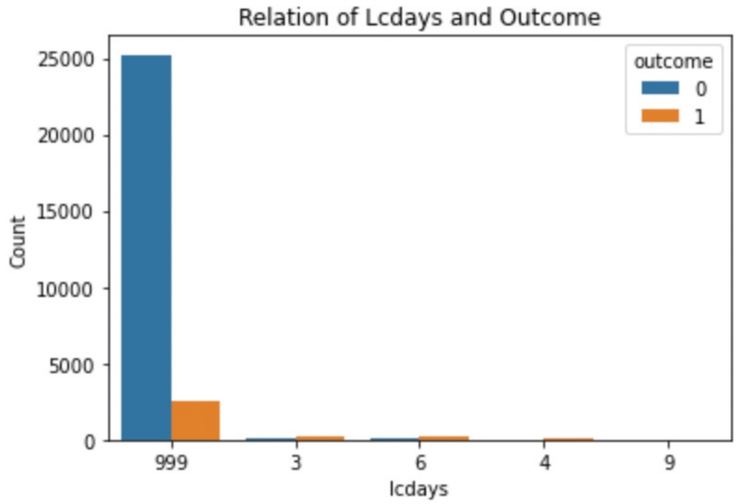
Min and Max age of customers is 17 and 98, respectively. Median age is at 38. Out of total customers who opened savings account, age group 25 to 40 contribute to 53%

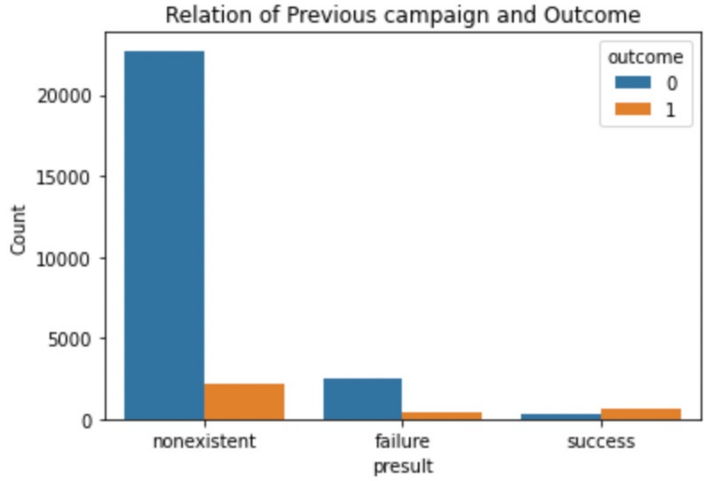


Customers contacted in month of May were highest followed by July and August and in terms of day have similar trend.

Data Preparation







Variable treatment

- Customers contacted upto 4 days were having impact on the outcome. Customers <=4 were converted to 1 and others as 0

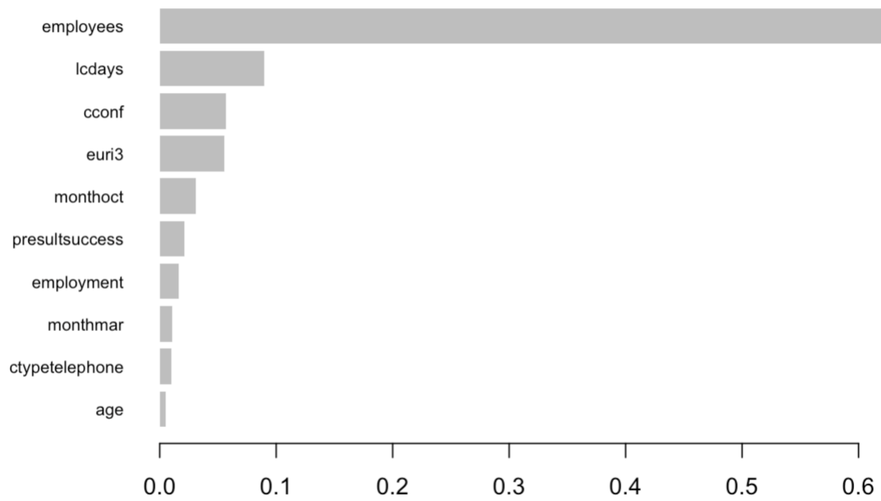
- Customers with 999 are new customers and thus they were made 0 and others who were contacted as 1.

- Presult_success was one of the important variable and thus it was treated separately

- Customers with >2 had no great impact on the outcome thus clubbed Customer with <=2 as 1 and >2 as 0

- Other :
  - Rows with unknowns (~2k) were not impacting the outcome thus same was imputed.
  - One hot encoding was performed on categorical variables.
  - All numerical variables were scaled.

Top 10 important variables as per Model



Following variables which were not important were dropped from the dataset.

- Credit
- Ploan
- Job_entrepreneur/Job_housemaid
- Edu_6k/Edu_illiterate

Models

Steps performed before Machine Learning Models

- Basic pre-processing of data as explained in data preparation slide
- Principal component Analysis (PCA) : Used PCA for dimensionality reduction.
- Reduce rows of minority class (0) and used oversampling

Additional Steps

- Cross Validations
- Stratified K-Folds
- Hyper parameter tuning and Grid search

Machine Learning Models

Applied algorithms on the models

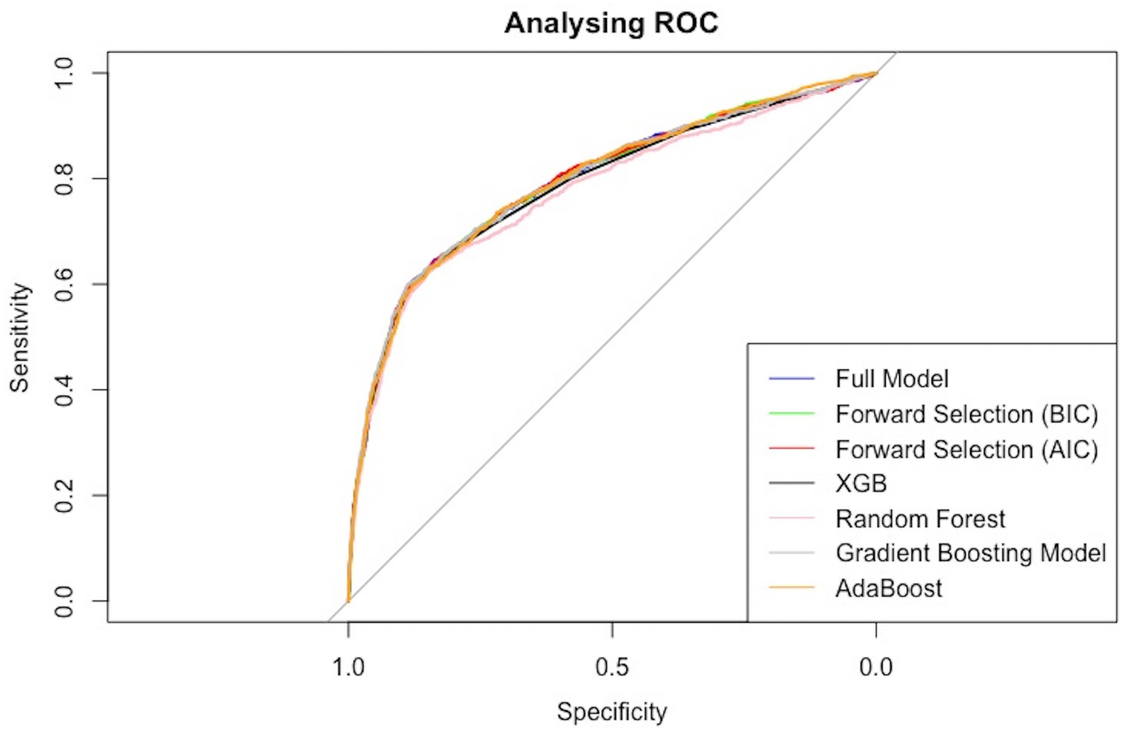
1. **Logistic Regression**
2. **Random Forest**
3. **Support Vector Machine**
4. **Forward and Backward selection**
5. **Lasso & Ridge**

Boosting Algorithms

1. **Ada Boost**
2. **XG Boost**
3. **Gradient Boosting**

Evaluation Methodology

Analyzing ROC Curve



Models	AUC
Logistic Regression	0.7926
Forward Selection - BIC	0.7931
Forward Selection – AIC	0.7939
XG – Boost	0.7872
Random Forest	0.7766
Gradient Boosting Model	0.7950
Ada Boost	0.7944

- Models like Lasso, ridge, SVM did not have good accuracy on the dataset and thus excluded from above graph.
- PCA and sampling technique was excluded from final script, the auc was not impacted by the same.
- AUC for all the models performed were in the range of ~0.78-0.79.
- Gradient boosting model has the maximum AUC in both validation test data and unseen test data.
- Gradient boosting can handle both numerical and categorical data in the same model and thus worked better on the banking data.

Managerial Implications

Learnings from dataset and Actionable

- Around 60% of the dataset comprised of married individuals and 22% of which opened savings account.
- Age feature demonstrates that the majority savings account are opened by customers between age group 25 to 58 years.
- Employment variation rate has negative influence. A stable employment rate denotes a stable economic environment in which people are more confident to make their investment.
- From Presult variable it is observed that customers who have positively reacted to previous campaign have higher chance of opening the bank account. (i.e. 63% conversion)
- ccontact variable depicts that there is reduction in impact of contact after 4+ days

1. Banks can use the said learning to align and further improve their **marketing campaign**.
2. Additionally, Banks can **hire more people and/or train existing resources** to improve the quality their campaigns.
3. Some of the below variables can be used to improve their **lead generation strategy.**
   - Married
   - Age group between 25 to 58
   - Jobs - Artisans or Retired

Limitations

Bank can add following information to the dataset for better prediction of the outcome :

- Customer characteristic and their preferences.
- What are the type of marketing campaigns and are most successful.
- Information regarding competitors.

If such information is available in the future, we could compare and analyze the customer behaviour to further tune our model.

Thank You