**PROPENSITY MODELLING AND BANK MARKETING**

Divya Somakumar (06) & Shreya Paul (15)

**ABSTRACT**

Banks don’t have unlimited funds to conduct various campaigns and identify the various factors that lead to the success of the campaign. This paper tries to estimate the various factors that lead to the success of the Portuguese Bank marketing campaign using propensity models. It also attempts to identify the right campaign for the right customer. This could improve the cost efficiency of the campaign.

**Keywords**: Propensity Modelling, Term deposits, Probability.

**INTRODUCTION**

Propensity modelling is a type of statistical technique that attempts to estimate the probability of occurrence of certain kinds of behavior in the subject as a result of a particular external stimuli. It tries to study the impact of different stimuli like campaigns, treatments, advertisements, etc. on different individuals. It is similar to conditional probability. It takes into account all the independent and confounding variables that affect our decision-making behavior. It is mostly carried out on observational Data. The propensity scores vary depending on the different observed characteristics viz age, education, ethnicity, etc. They are mainly carried out because Businesses do not have unlimited Funds to conduct different campaigns and identify which campaign was the most successful. It also tries to identify the various factors that lead to the success of the campaign. In our study, we have used Portuguese Bank Marketing Dataset to identify whether a Telephonic Campaign was successful in getting people on Board to subscribe for Term Deposits. This data was obtained from the paper named “Using Data Mining for Bank Direct Marketing: An application of CRISP DM Methodology” by Sérgio Moro, Raul M. S. Laureano and Paulo Cortez. A term deposit refers to a fixed term investment with a financial institution like a bank, etc.

**LITERATURE REVIEW**

Propensity Modelling was introduced by Paul Rosenbaum and Donald Rubin in their paper Titled “The Central Role of Propensity Scores in Observational Studies for Causal Effect” in the year 1983.Therese A Stutkel, Elliot A Fisher ,et.al have used propensity scores to identify the effects of Cardiac Management on AMI survival .McCaffrey , Daniel F, et. al have tried to estimate the propensity scores using boosted regression to study the causal effects on observational experiments. G Bondi Zoccai, E Romognoli,et.al have tested the superiority of propensity scores over multivariate analysis. Taylor and Francis have studied the statistical and regulatory issues pertaining to propensity modelling. RJ Glynn and S Schneweis have used propensity scores in the field of pharmacoepidemiology. There are several studies based on the Bank Marketing Dataset. Jaino Martinez in his study based on the same bank marketing Dataset found that not more than 3 calls should be made to the potential client to make him buy a Term Deposit. He also found that there is a likelihood of client declining the offer if he/she is called more than 3 times. Shirantha has used a decision tree approach on the same dataset.

**METHODOLOGY**

First and Foremost, we have Encoded our categorical variables like job, marital status, education, credit defaults, housing loan taken, personal loan taken, mode of communication (cellular, Telephonic), outcome of previous campaigns, last contact month of the year, etc. using Label Encoder. A label encoder converts a nonnumerical data into numerical data. Numerical data like age, contact, call duration, Bank balance of the individual, number of times the client was contacted, number of days since the last contact, number of times the client was contacted before this campaign was scaled using a Standard Scaler. Standard Scaler ensures that the variables do not behave unruly.

After Encoding our categorical data and scaling our numerical data we ran the XG Boost Algorithm. XG Boost stands for Extreme Gradient Boosting. It was introduced by Tiaqui Chen and Carlos Guestrin. Gradient Boosting is used for solving regression and classification problems and helps in producing a prediction model. Boosting is an ensemble model wherein new models are added sequentially to reduce the errors. Each model learns from the previous model and updates the residual error. XG Boost is highly efficient as its computations take place very fast. The major difference between Gradient Boosting and XG Boost is that XGBoost can work with any differentiable Loss Function.

After XG Boost we estimated the confusion matrix also known as error matrix in order to gauge the performance of a classification algorithm. It gives us insights regarding the types of errors that may have occurred during classification.

After XG Boost and Confusion matrix we tried to estimate the Propensity Score. Propensity Score matching is a quasi-experimental technique. There are several methods to calculate the Propensity scores like Logistic Regression, Probit Analysis, discriminant analysis, tree based methods, etc.In our study, to estimate the Propensity Score we have used the Pymatch algorithm introduced by Ben Miroglio which was adapted from the Jasjeet Singh Sekhon's Matching package in R. In this algorithm there is a Matcher function which uses Logistic Regression models to estimate the propensity scores. In our case we are using our variables like ‘age’, ‘education, ‘marital status’, ‘campaign’ etc. to estimate the probability of subscription of a term Deposit. The scores thus obtained were matched to the corresponding individuals.

**RESULTS**

Firstly, we ran our XGBoost algorithm. On applying the XG Boost algorithm we got an accuracy score of 100% which implies that there is no difference between the actual values and predicted values.

After running the XG Boost algorithm , we built our Confusion Matrix also known as the Error Matrix. We obtained the following confusion matrix;

|  |  |  |
| --- | --- | --- |
|  | Predicted  No | Predicted  Yes |
| Actual  NO | 778 | 0 |
| Actual  Yes | 0 | 101 |

There are 778 True Positive implying that the observation was positive and was predicted to be positive is positive and 101 True negative which implies that the observations which were negative were predicted to be negative. We don't have any False positives or False Negative. Thus, our classification model has not committed any classification errors.

Mostly propensity modelling is applied in the Medical field to identify the impact of the various treatment. In our model we are trying to apply propensity scores to determine the course of action that has bank has to undertake in order to get their customers to subscribe for a term deposit. Propensity score is the conditional probability that a customer would subscribe for a Term Deposit given the confounding variables like age, education, marital status, campaign, bank balance, loan, mode of communication, etc.

**P(x)= P(Y=0|X=x)**

Where ‘Y’ corresponds to whether a customer would subscribe to a term deposit or no.

Y=0 or Y=1

Where x = age, marital, education, loan default, bank balance, housing loan, personal loan, mode of contact, number of days before previous contact, outcome of previous campaigns.

On applying the Matcher function of the pymatch algorithm the following results were obtained;

**Y~age + job + marital + education + default + balance + housing + loan + contact + day + month + duration + campaign + pdays + previous + poutcome**

**N majority:4000**

**N minority :521**

This implies that out of 4521 customers 4000 customers had subscribed for a term deposit while 521 had not subscribed for a Term deposit. We have used the 4000 customers who have subscribed to the term deposit to fit our model. We have taken 100 samples from it and have got an Accuracy score of 79.27%. This means that 79.27% of time our predictions have come true. In our case 79.2% of the time a Telephonic campaign was successful in making a customer subscribe to a term deposit. The propensity scores were matched to the individual customers. This gave us the individual conditional probabilities of the customers of the banks.

We have also tried to plot several graphs based on the various confounding variables like age, education, bank balance, etc. and the propensity scores. These gave us interesting insights about the effectiveness of the campaign.

* Firstly, there is a positive relation between the ages of the customers and the probability that the customer would subscribe for a term deposit i.e. People belonging to the age group of 30-60 are more likely to subscribe for the term deposit. As any model cannot be perfect, there are certain exceptions in this case also. (Please Refer to the Graph:3 in the Appendix)
* There is a positive relation between bank balance and Term deposits. The people who have low bank i.e. having less than 1000 Euros balance are less likely to subscribe to a term deposit. (Refer to Graph:4 in the Appendix)
* The graph which plots the campaign and the propensity scores have interesting insights. It is less likely that a customer would subscribe to a term deposit if contacted for less than 3 times. The probability that the customer will subscribe for a term deposit is very high when the customer is contacted more than 3 times (Refer to Graph:5 in the Appendix).
* The duration of the call also has a high impact on the subscription of Term Deposit. As the duration of the call increases the probability of the customer subscribing to a term deposit also rises. There are certain exceptional cases wherein although the call duration was not long the customer subscribed to the term deposit. (Refer to Graph in the Appendix)

**CONCLUSION**

On the basis of the propensity scores obtained in this paper, it can be concluded that banks should focus their campaigns more on their customers who belong to the age group 30-60 and have either secondary or Tertiary education with a bank balance of more than 1000 Euros. To get the customers subscribe to the term deposit the bank has to contact the customer more than thrice and the call duration should also be long.

# **References**

(n.d.). Retrieved from Infolitik: http://www.infolitik.com/analytics/customer-analytics/

Katiyal, A. (2017). *In-class Kaggle Classification Challenge for Bank's Marketing Campaign.* Kaggle.

Mahata, H. Q. (2018). *Bank Marketing Classification, ROC,F1,Recall.* Kaggle.

Mandot, P. (2019). *How XG Boost Exactky works.* Medium.

Miroglio, B. (2017). *Introducing Pymatch Python Package.* Medium.

Morde, V. (2019). *XG Boost Algorithm: Long May She Reign!* TowardsDataScience.com.

Mulin, S. (2019). *Propensity Modeling: Using Data (and Expertise) to Predict Behavior.* CXL.

Sérgio Moro , Raul M. S. Laureano,Paulo Cortez. (n.d.). *USING DATA MINING FOR BANK DIRECT MARKETING:AN APPLICATION OF THE CRISP-DM METHODOLOGY.* Portugal.

Shirantha. (2017). *Bank Marketing Data: A Decision Tree approach.* Kaggle.

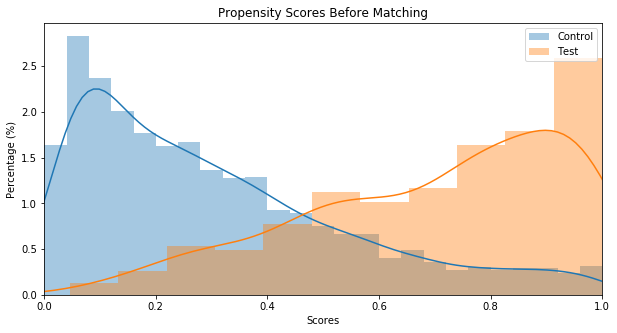
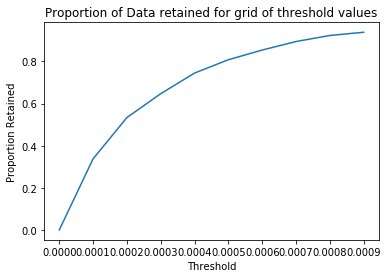
Webb, T. R. (2018). *Propensity Modelling for Business.* Data Science Foundation.

White, T. (2019). *Hadoop The Definitive Guide.* Sebastpool: O'Reilly.

**APPENDIX**

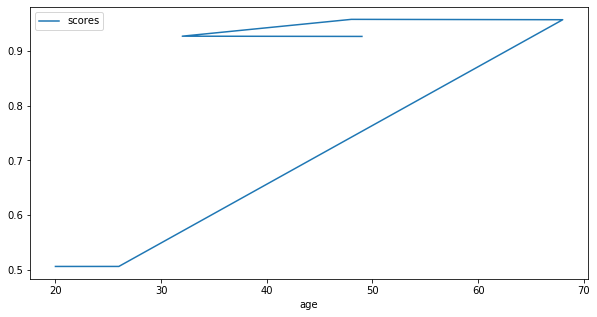
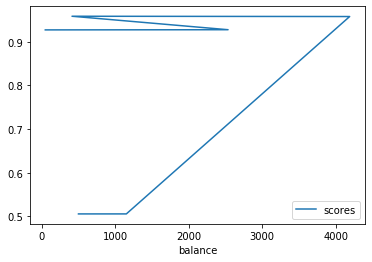
Graph:1

Graph:2

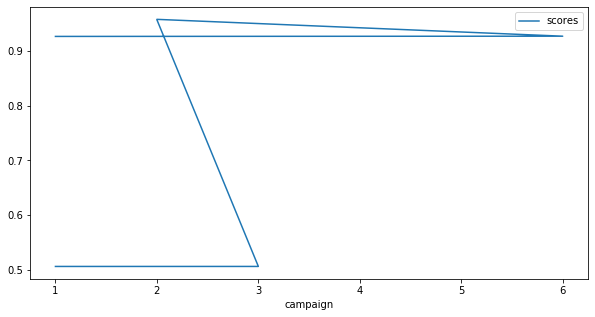
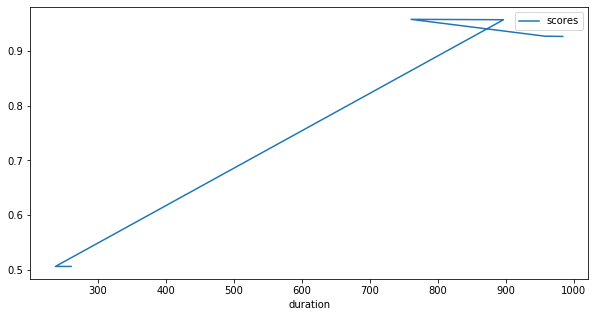
Graph:4

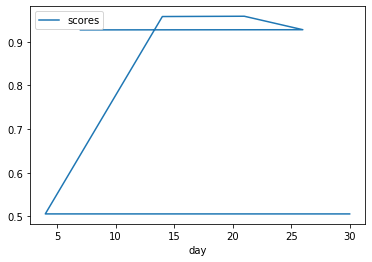
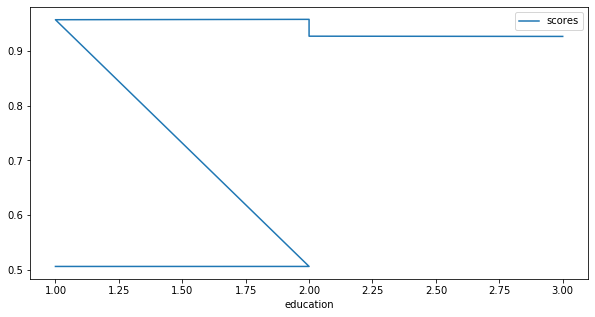
Graph:3

Graph:5

Graph:6



Graph:8

Graph:7