

Data Analysis Project for Marketing Campaign

Kushal Ahuja

(Master of Business Analytics)

University of Technology Sydney

Table of Contents

[Introduction 4](#_Toc150369290)

[Method 6](#_Toc150369291)

[Results 8](#_Toc150369292)

[Discussion 12](#_Toc150369293)

[Limitation of the work 13](#_Toc150369294)

[Conclusion 13](#_Toc150369295)

[Reference 14](#_Toc150369296)



Introduction

# **Introduction**

The report focuses on the insights generated from statistical models about the marketing campaign; these methods will help the company to target the desired customers, resulting in increasing the subscription rate among the consumers.

This project discusses customer segmentation, which is examining the large amount of data that exists on customers (and possible customers) to accurately separate customers into distinct groups based on certain social, behavioural, and other factors. (*Customer Segmentation Meaning & Analysis Models | Optimove*, 2023)

The aim of this project is to determine customer segmentation using statistical methods, following the results, and generating business insights that can help the company to target customers effectively and increase the subscription rate among the current customer base.

Research questions-

* How do statistical learning models perform in predicting the marketing campaign's success?
* Which features are important to consider for targeting the customers?
* How do these features affect the likelihood of a subscription?



**Method**

# **Method**

The methods implemented are Logistic Regression with MLE as model estimation as a parametric model and Random Forest as a non-parametric model.

The reason behind implementing the logistic regression is due to the following –

* It acts as a baseline model, and it is reliable with a larger dataset.
* It provides easier interpretation using coefficients, which will be helpful in presenting to non-technical stakeholders.

The reason for implementing Random Forest is due to the following-

* Its robustness to fit the dataset and its effectiveness for the complex and bigger dataset where there is a chance of overfitting the model.
* It handles well the mixed data of categorical and numerical variables like in this dataset; this is because this model doesn’t need scaling of the dataset.
* Due to the feature importance of this model, business insights will be generated.

**Summary Table**

|  |  |  |
| --- | --- | --- |
| **Features** | **Details** | **Observations** |
| **Duration** | Last contact duration with the customer | Duration attribute has more significant importance than any other feature. |
| **Job** | The type of job customers do. | The job column has shown feature importance from the random forest model, which implies that some jobs can have better chances of subscription. |
| **Contact** | The kind of contact channel used | The contact showed no correlation in the Logistic Regression model, meaning this attribute does not have much importance. |
| **Education** | The level of education customer has or pursue | This attribute demonstrates the importance of the Marketing campaign. |
| **Default** | The customer has ever credit in default | This attribute doesn’t show much importance towards the campaign. |
| **Age** | The age group of customers | This attribute has shown significant importance. |



**Results**

# **Results**

Both models have been evaluated based on evaluation metrics, and then insights are generated which will be useful for the company to make better decisions about the marketing campaign.

**Subheadings of research questions-**

How do statistical learning models perform in predicting the marketing campaign's success?

The below results show the model performance based on certain metrics-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **F1- Score** | **ROC-AUC Score** |
| **Logistic Regression** | 91.45% | 68.8% | 52.4% | 93.8% |
| **Random Forest** | 91.51% | 64.6% | 58% | 94.5% |

It is evident that the Random Forest Model demonstrate higher model performance compared to Logistic Regression.

For better interpretability of the model comparison, the following bar chart is generated

A graph of a graph of a graph

Description automatically generated with medium confidence

1. **Which features are important to consider for targeting the customers?**

To acknowledge the customers most responsive to the marketing campaign, the model predictions are used for insights; these insights can prove to be beneficial when implemented,

1. The insights generated can be visualised from the Coefficients of the Logistic Regression Model.

A graph with blue squares

Description automatically generated

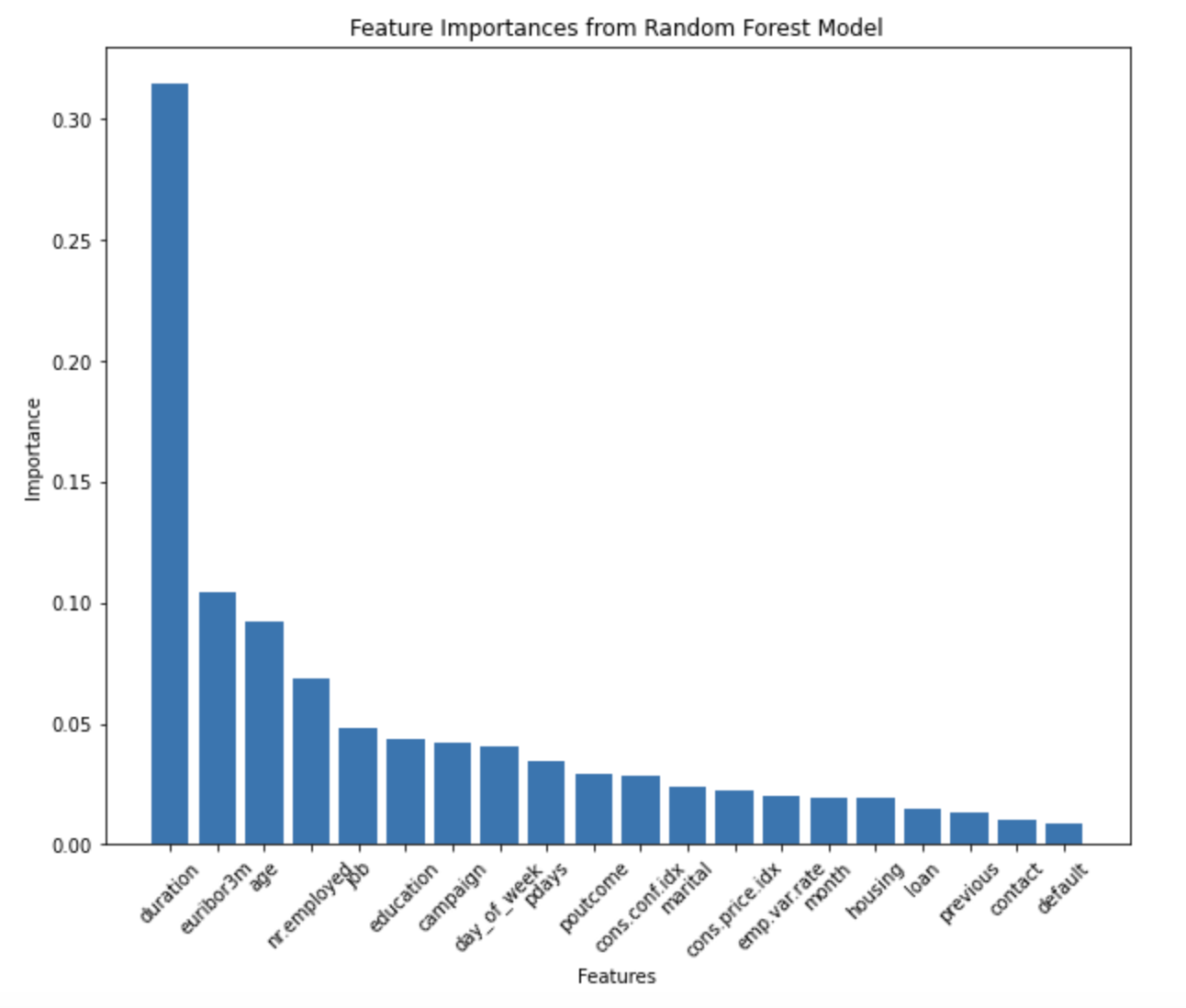
From the above graph, the insights of features that are significant are provided, these features have been shown to be responsive to the marketing campaign and have been predicted to be a better target for the marketing campaign, the features include –

* Duration, poutcome, consumer price index (cons.price.idx), marital, age and education

While in contrast, the features that are not significant for a better response to the marketing campaign or which are less likely to subscribe after the campaign are –

* Employee variation rate (emp.var.rate), Default, month and pdays.

1. The insights generated through the Random Forest model can be visualised from the feature importance of Random Forest.



The features that are predicted by the random forest model to be most responsive to the marketing campaign, resulting in higher subscription rate, are duration, age, job and education. In contrast, the features that are less responsive to the campaign are loan, housing, contact and default.



**Discussion**

# **Discussion**

This section outlines the results from the model; these models helped to acquire the insights by performing the predictions, and both have shown good accuracy results, implying that if the decisions are made using this insight, it can help to increase the subscription rate from the desired customer segment, these segments have been made using the features in the dataset. For instance, grouping the customers according to their educational background.

The business insights generated are as follows –

1. The attribute **Duration** has shown the most significant response towards the campaign, stating that the company should focus on increasing the duration of contact with the customers, The majority of the data states that the customer has contacted with very few seconds and thus doesn’t know about the subscription plan. Focusing on increasing and maintaining the contact will be positive decision and action.
2. **Job** feature has also shown importance in the random forest model, stating that certain jobs would have a higher potential to subscribe. Moreover, with the help of EDA, it was found that student, admin and management jobs are more likely to respond and subscribe to telecom services. So, the company should focus on targeting them as many people have admin jobs who are not subscribing.
3. The **Education** attribute showed a positive coefficient and higher importance, which implies that certain education features are more likely to subscribe. Additionally, with the help of EDA, it was found that people who have university degrees are more likely to subscribe. The company can also focus on people with high school education backgrounds as it has a higher customer base.
4. The **Default** feature has demonstrated negative coefficients, meaning this feature shouldn’t be focused on by the company as it is not significant in terms of response or the subscription rate.
5. **Contact** column/ feature is also negative coefficients, so it doesn’t matter which communication channel is being used to contact, so the company should not focus on taking this attribute while generating customer segmentation or data-driven decisions.
6. Some attributes like **Age** and **Marital** Status are also significant as they are positive coefficients. So, the company can target the customers on the basis of these attributes too.

## Limitation of the work

The limitation of this work can be that the model results can be biased due to lower precision, meaning the company should also focus on any external factors that are influencing.

The model has also been implemented using balancing the class, but due to lower changes in the results that haven’t been performed.

The other limitation can also be the assumptions made by the logistic regression model, which can be misleading for complex relations.

# **Conclusion**

This project started by looking at how well the logistic regression and random forest models can predict the marketing campaign will do, followed by finding the features that will be more important for targeting to get subscribed and improve the marketing results.

The analysis showed that certain features that are important to consider and target are Duration, Job, Education and Age, while some features like contact, default, loan and housing are not important to consider as these features do not show positive results.

# **Reference**

*Customer Segmentation Meaning & Analysis Models | Optimove*. (2023, July 6). Optimove. <https://www.optimove.com/resources/learning-center/customer-segmentation#:~:text=Using%20the%20large%20amount%20of,demographic%2C%20behavioral%20and%20other%20indicators>.

Chris, K. (2022, July 12). *With Open in Python – with statement syntax example*. freeCodeCamp.org. <https://www.freecodecamp.org/news/with-open-in-python-with-statement-syntax-example/#:~:text=To%20use%20the%20open%20function,file%20in%20a%20print%20function>

*General functions — pandas 2.1.2 documentation*. (n.d.). <https://pandas.pydata.org/docs/reference/general_functions.html>

S, R. A. (2023, May 18). *Map in Python: An overview on map function in Python*. Simplilearn.com. <https://www.simplilearn.com/tutorials/python-tutorial/map-in-python#:~:text=Map%20in%20Python%20is%20a,to%20the%20map%20in%20Python>.