



DataCorp

# Bank Targeted Marketing

John Paul Hernandez Alcala



Hello, my name is John Paul, and I am a data scientist with DataCorp. Today my goal is to give you a better understanding of which clients are more likely to subscribe to a term deposit. Let's get started!

# Problem Statement

- |    |   |
|----|---|
| 01 | Investigate how some client features affect term deposit subscription outcome |
| 02 | Derive models that can predict prospective subscribers                        |
| 03 | Propose 3 business recommendations for selecting the best clients             |



DataCorp

Before we dive into this, we have to present ourselves with three tasks that make up our problem statement: Investigate which client features affect term deposit subscription outcome, derive models that can predict prospective subscriber, and propose 3 business recommendations for selecting the best clients.

.

# Business Value

01	Investigate how some client features affect term deposit subscription outcome	<ul style="list-style-type: none"><li>• Bar Graphs for Outcome vs Client Personal Features</li><li>• Violin Plot for Outcome vs Client Personal Features</li><li>• Bar Graph and Point Plot for Outcome vs Client Interaction Features</li></ul>
02	Derive models that can predict prospective subscribers	<ul style="list-style-type: none"><li>• 5 Different Models</li></ul>
03	Propose 3 business recommendations for selecting clients	<ul style="list-style-type: none"><li>• Top Predictors from Final Model and Analysis<ul style="list-style-type: none"><li>◦ Interaction: month contacted, previous calls</li><li>◦ Personal: job, education, and age</li></ul></li></ul>



DataCorp

We accomplish these tasks to guide a financial institution or bank on which features a prospective term deposit subscriber would possess, so target marketing can be achieved. For the 1st task, we will look at multiple graphs, so we know which features result in a high or low subscriber count.

For the 2nd task, we will use our data to create 5 different models. Finally, based on the results of our final model and analysis, we can propose 3 business recommendations for selecting the best clients.

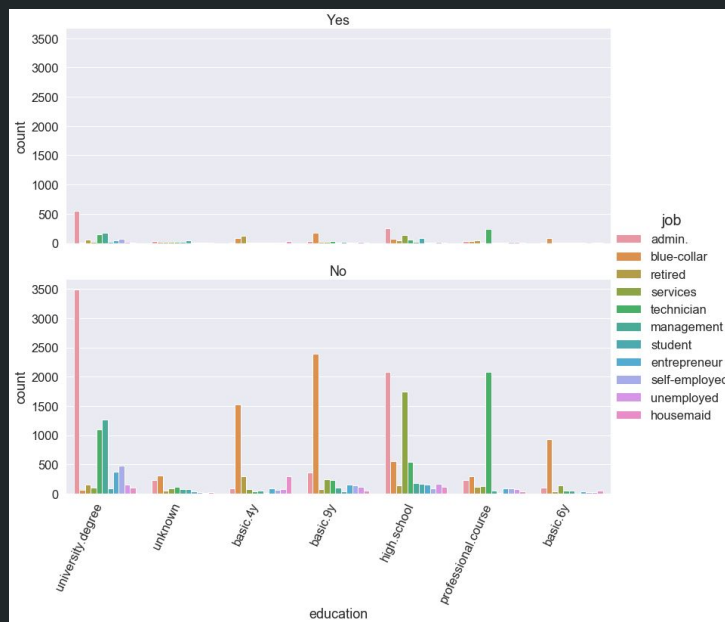
# Methodology



DataCorp

To start on our journey of completing our tasks, we first start off by collecting data from UCI, work our way up with statistical models such as violin plots, bar graphs, and point plots for categorical plots and decision tree and more ahead for supervised models. Finally, we will arrive to our three business recommendations.

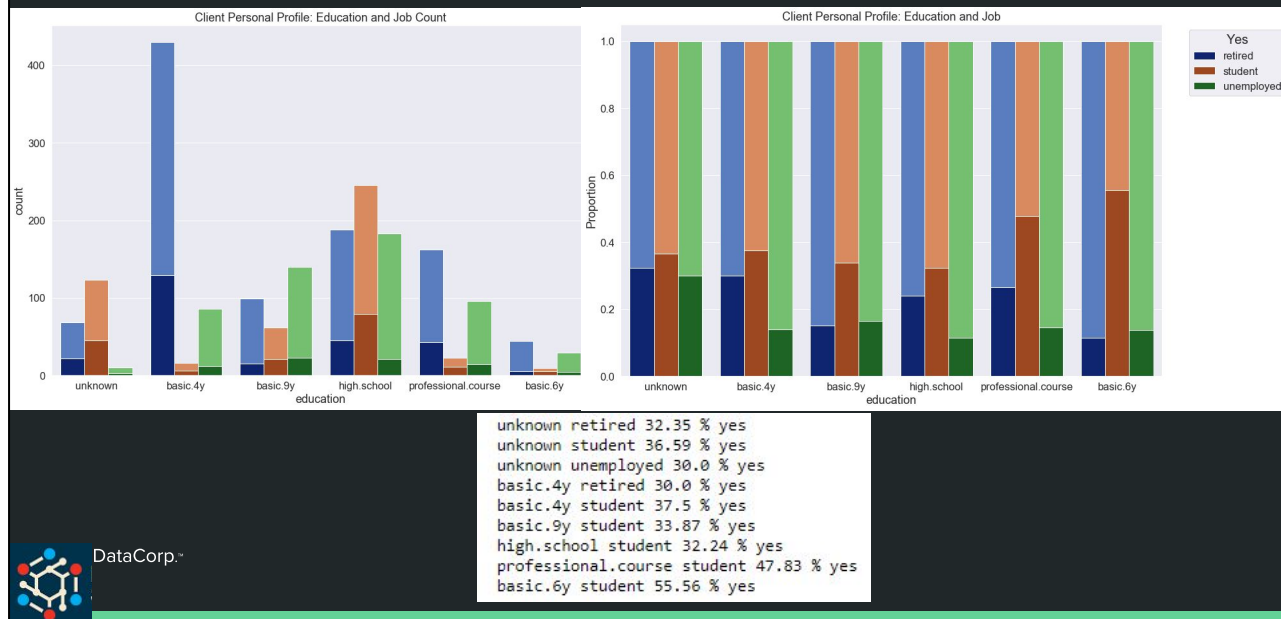
## Bar Graph of 'job', and 'education' Bar Graphs



DataCorp

Here we have a bar graph that shows all the 'education' values and 'job' values.

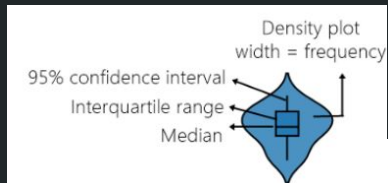
## Recommendation 1: Students, especially those who have an education level at 'basic.6y', 'professional.course', 'basic.4y'



We see from this graph that we have  $\geq 30\%$  chance of the listed individuals with the 'job'-'education' pair will sign up for a term deposit:

The recommendation here is to focus on students who have an education level at 'basic.6y', 'professional.course', 'basic.4y'

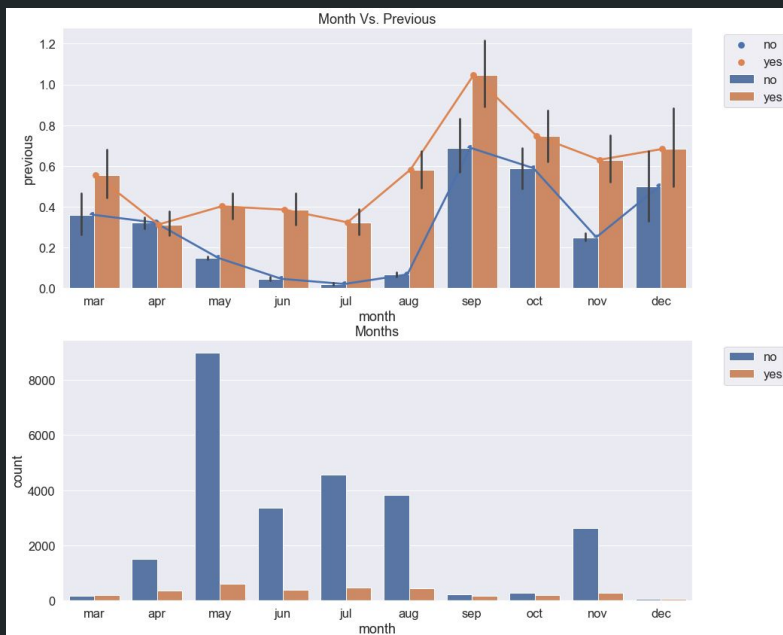
## Recommendation 2: students, retired and unemployed persons between 20-40 years old



DataCorp

Additionally, focus on students between 20-30 years old, or, in general, on clients between the age of 20-40 years old

## Recommendation 3: conduct more calls before a campaign and during Fall + March

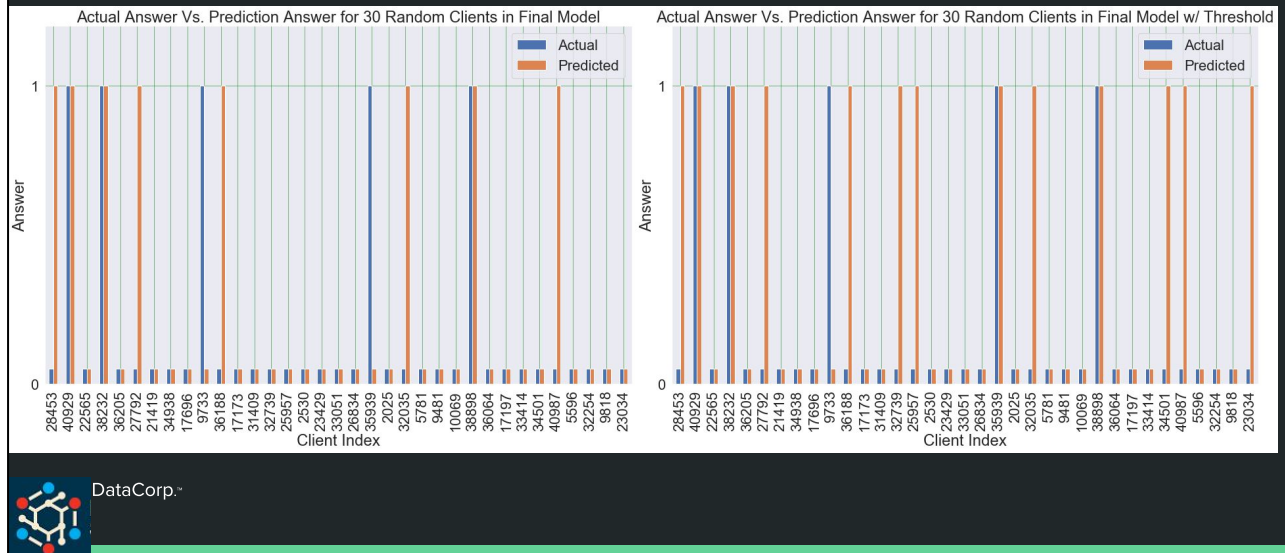


DataCorp

From above, we see that when more calls are made prior to the campaign of the CD, regardless of month, there is a noticeable likelihood that a client will request a term deposit. Additionally, the proportion of 'yes' in the Fall months and March is much higher than any other times. As such, the second recommendation is to conduct more calls before a campaign and during key months.

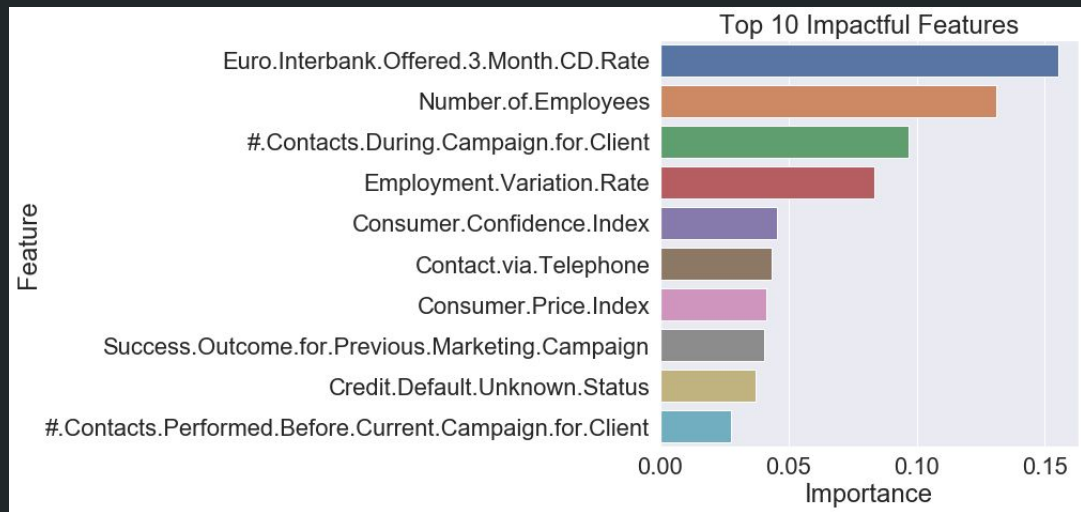


# Final Model vs Final Model with Threshold



Here we see our model with and without a threshold implemented. This threshold makes our model more sensitive, but less specific and precise than the default model; that is, our threshold model allows for more false positives in order to reduce false negatives which is our goal in this scenario. With the threshold model, we are able to predict ~72% of clients as subscribers who were indeed subscribers, and it also predicted 73% of clients as non-subscribers who were indeed not subscribers. This is compared to our original model that was able to predict 56% of clients as subscribers who were indeed subscribers and 92% of clients as non-subscribers who were indeed not subscribers.




# Top 10 Features



DataCorp

Here we see our model's top 10 features it employed.

# Recommendations Derived from Model and Analysis

-  1
  - Subscriber outcome is high for students who have an education level at 'basic.4y', 'basic.9y', 'high.school'
-  2
  - Subscriber outcome is high for certain age ranges for students, retired, and unemployed
-  3
  - Subscriber outcome is high for client contact efforts and timing: month contact, and previous calls.



DataCorp.™

With our model, we can see that subscriber outcome is high for client's job, specific education level and age. Additionally, it is high for contact efforts.

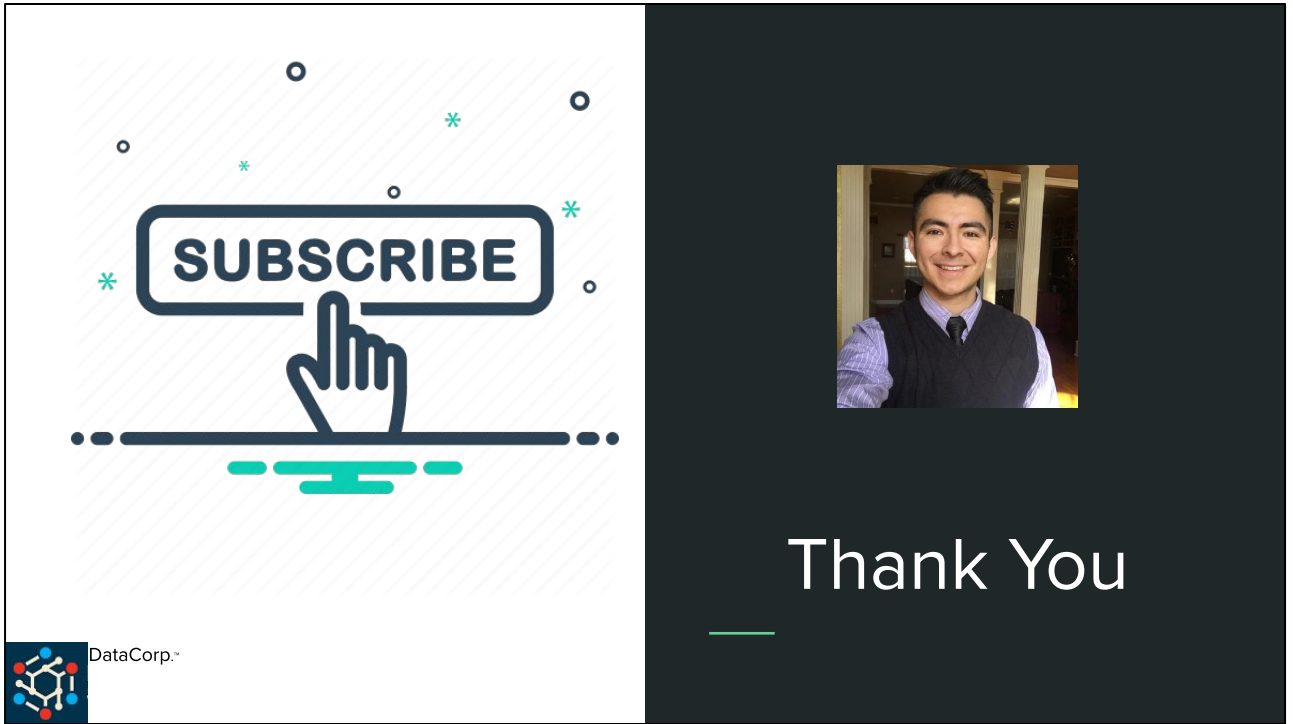
# Future Work

01	K Nearest Neighbor Imputation and Label Encoding	<ul style="list-style-type: none"><li>• Instead of median or mode for imputation</li><li>• Label encode some of the categorical variables that seemed ordinal</li></ul>
02	Employ Other Imbalanced Data Techniques	<ul style="list-style-type: none"><li>• Right evaluation metrics, resample the training set (under-sampling or over-sampling), cluster the abundant class, anomaly detection</li></ul>
03	More data	<ul style="list-style-type: none"><li>• Obtain more data from King County</li></ul>



DataCorp

Now that we have proposed our three business recommendations, we know there are ways we can have more confidence with our predictors. We can use K Nearest Neighbor Imputation and Label Encoding instead of median and mode. We can use other imbalanced data techniques to improve recall, f1 score, and AUC. Finally, we can get more data to train and test against our current model which will allow us to work towards a better model.



Thank you for your time! Please, feel free to ask me any questions at this time.

# Appendix 1a. Features from Dataset

1. age (numeric)  
2. job : type of job (categorical: "admin.", "blue-collar", "entrepreneur", "housemaid", "management", "retired", "self-employed", "services", "student", "technician", "unemployed", "unknown")  
3. marital : marital status (categorical: "divorced", "married", "single", "unknown"; note: "divorced" means divorced or widowed)  
4. education (categorical: "basic.4y", "basic.6y", "basic.9y", "high.school", "illiterate", "professional.course", "university.degree", "unknown")  
5. default: has credit in default? (categorical: "no", "yes", "unknown")  
6. housing: has housing loan? (categorical: "no", "yes", "unknown")  
7. loan: has personal loan? (categorical: "no", "yes", "unknown")

*Related with the last contact of the current campaign:*

8. contact: contact communication type (categorical: "cellular", "telephone")  
9. month: last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")  
10. day\_of\_week: last contact day of the week (categorical: "mon", "tue", "wed", "thu", "fri")  
11. duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

*Other attributes:*

12. campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)  
13. pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric, 999 means client was not previously contacted)  
14. previous: number of contacts performed before this campaign and for this client (numeric)  
15. poutcome: outcome of the previous marketing campaign (categorical: "failure", "nonexistent", "success")

*Social and economic context attributes*

16. emp.var.rate: employment variation rate - quarterly indicator (numeric)  
17. cons.price.idx: consumer price index - monthly indicator (numeric)  
18. cons.conf.idx: consumer confidence index - monthly indicator (numeric)  
19. euribor3m: euribor 3 month rate - daily indicator (numeric)  
20. nr.employed: number of employees - quarterly indicator (numeric)

*Output variable (desired target):*

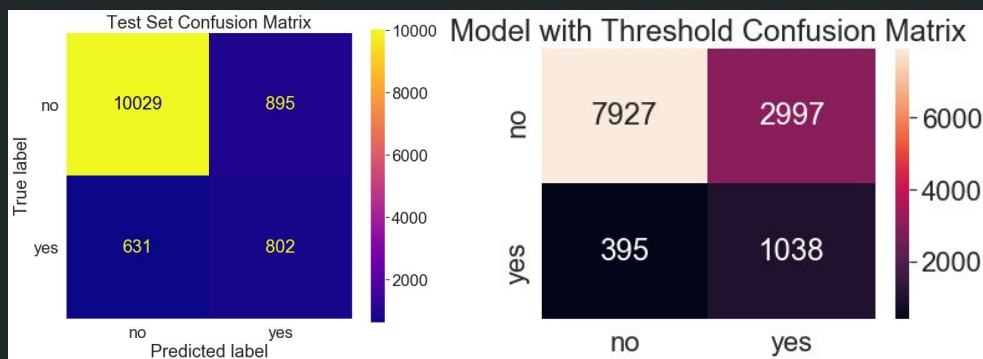
21. y - has the client subscribed a term deposit? (binary: "yes", "no")



DataCorp.

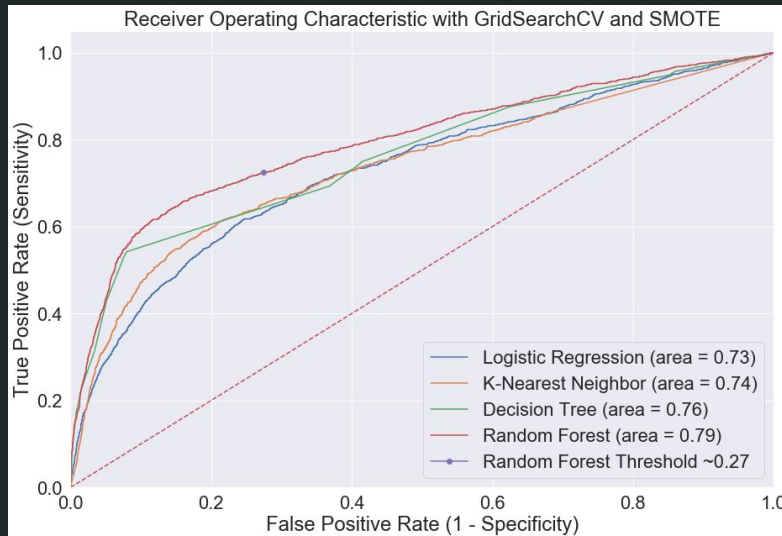
Key for features

## Appendix 1b. Confusion Matrix for Final Model and Final Model with Threshold



DataCorp

## Appendix 2. ROC-AUC Curve Graph



From the above, we see that the best model to focus on is the random forest one since it resulted in a AUC of 0.79 in our ROC Curve Graph. We have talked about F1 and recall score, but we have not discussed what AUC means for our model. Generally, the higher the AUC, better the model is at distinguishing between clients that will subscribe and will not. So an AUC of 0.79 means that there is 79% chance that the model will be able to distinguish between 'yes' class and 'no' class.