Bank Dataset

Aaron Abromowitz, Stephanie Duarte, Dammy Owolabi

Agenda

- Introduction
- EDA
- Objective 1: Interpretative Model
- Objective 2: Predictive Models
- Conclusion

Introduction

- Bank dataset from portuguese bank
- 41,188 clients, 19 explanatory variables, 1 variable of interest
  - 9 numeric variables, 10 categorical variables
- Categorical variables had “unknown” values as a separate class
- Variable of interest: has the client subscribed to a term deposit?
  - This is the ‘y’ variable in the dataset
- Bank is interested in the effectiveness of their campaign
- Random 80/20 split
  - Training Data: 32,950 rows, used for analysis and model creation
  - Test Data: 8,238, used for model validation

Note on Duration Variable

- Duration = last contact duration, in seconds (numeric)
- This attribute highly affects the output target (e.g., if duration=0 then y="no").
- This input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
- We decided to remove the variable from our analysis

EDA

Dataset Columns (subset)

Variable Name	Variable Description	Variable Category	Data Class
default	Has credit in default?	bank client data	categorical
loan	Has personal loan?	bank client data	categorical
month	last contact month of year	related with the last contact of the current campaign	categorical
day_of_week	last contact day of the week	related with the last contact of the current campaign	categorical
campaign	number of contacts performed during this campaign and for this client	other attributes for the campaign or past campaigns	numeric
pdays	number of days that passed by after the client was last contacted from a previous campaign, 999 means client was not previously contacted	other attributes for the campaign or past campaigns	numeric
previous	number of contacts performed before this campaign and for this client	other attributes for the campaign or past campaigns	numeric
emp.var.rate	employment variation rate - quarterly indicator	social and economic context attributes	numeric
euribor3m	euribor 3 month rate - daily indicator	social and economic context attributes	numeric
nr.employed	number of employees - quarterly indicator	social and economic context attributes	numeric

Summary Statistics - Numeric

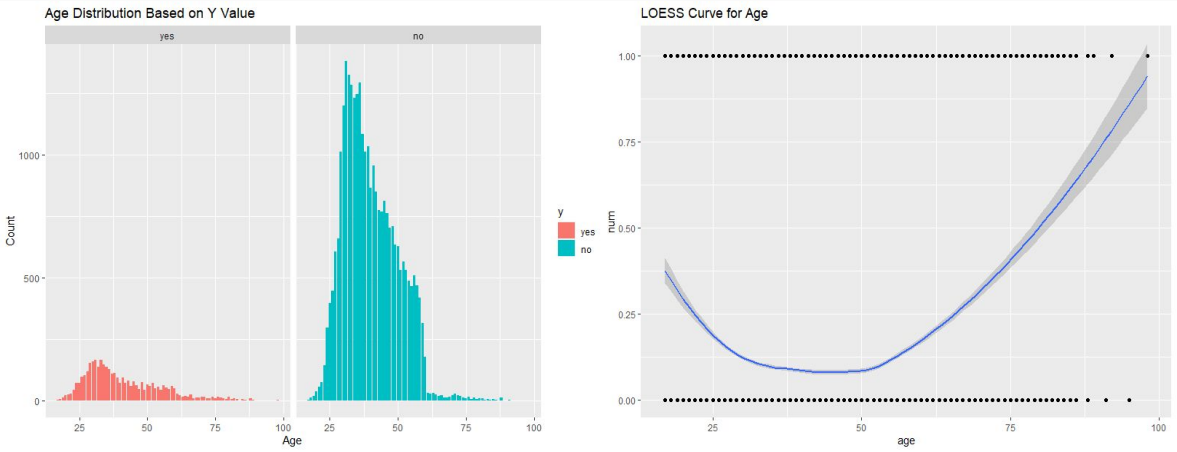
Variable Name	Min	1st Quartile	Median	Mean	3rd Quartile	Max
age	17	32	38	40.1	47	98
campaign	1	1	2	2.6	3	56
pdays	0	**999**	999	963	999	999
previous	0	0	0	0.2	**0**	7
emp.var.rate	-3.4	-1.8	1.1	0.075	1.4	1.4
cons.price.idx	92.20	93.08	93.75	93.57	93.99	94.77
cons.conf.idx	-50.8	-42.7	-41.8	-40.5	-36.4	-26.9
euribor3m	0.63	1.34	4.86	3.61	4.96	5.05
nr.employed	4,964	5,099	5,191	5,167	5,228	5,228

Summary Statistics - Categorical

Variable Name	Num Categories	Min Size	Median Size	Max Size
job	12	257 (unknown)	1,290	8,283 (admin)
marital	4	66 (unknown)	6,474	19,937 (married)
education	8	16 (illiterate)	3,758	9,738 (university.degree)
default	3	**2 (yes)**	6,888	26,060 (no)
housing	3	794 (unknown)	14,918	17,238 (yes)
loan	3	794 (unknown)	4,959	**27,197 (no)**
contact	2	11,961 (telephone)	16,475	20,989 (cellular)
month	**10**	138 (dec)	2,702	11,023 (may)
day_of_week	5	6,285 (fri)	6,513 (tue)	6,874 (thu)
poutcome	3	1,087 (success)	3,438	**28,425 (nonexistent)**

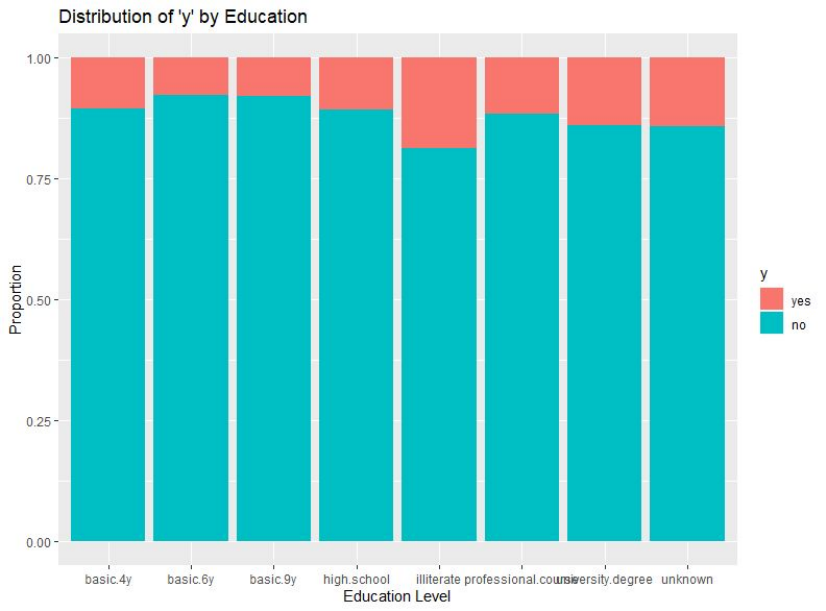
Ages

- Bank client data
- Numeric
- Similar distribution overall for both yes and no
- Far more no values than yes values
- People with age over 60 seem to be more likely to have a term deposit
- Probability of yes decreases, then increases
- Overall, it seems that this will not be a good variable for prediction



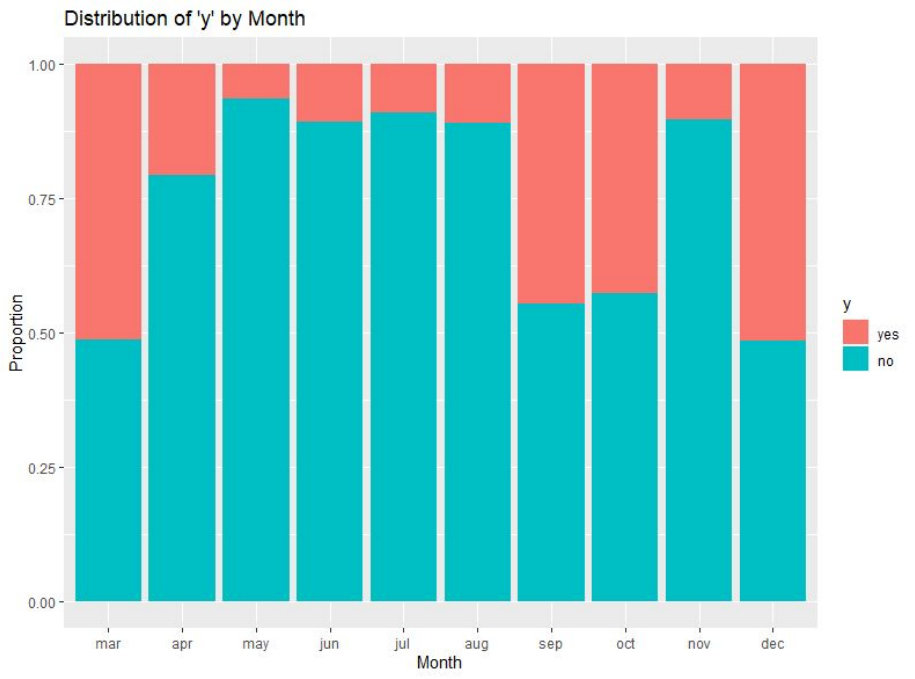
Education Data

- Bank client data
- Categorical
- Similar proportion of yes and no across all Education Levels
- This variable does not seem like it will be useful for prediction either



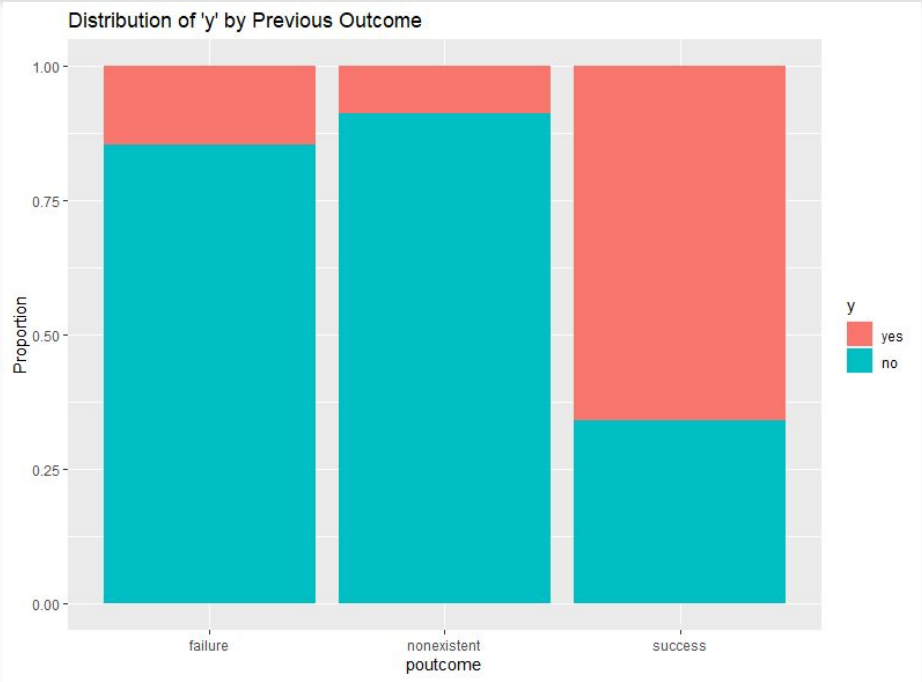
Month Data

- Related with the last contact of the current campaign
- Last contact month
- Categorical
- Some months have far more Yes values than other months
- Knowing the month could be useful for prediction



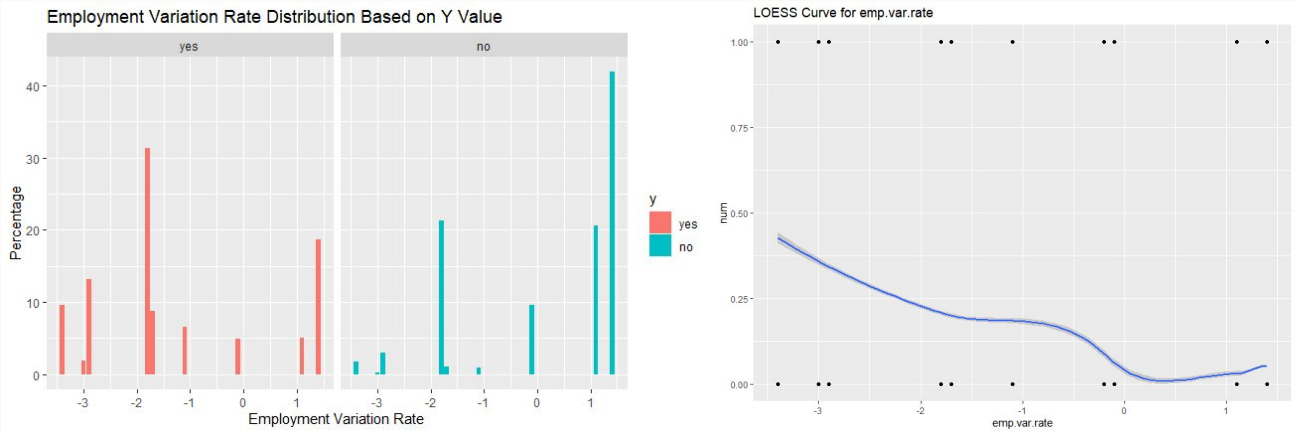
Previous Outcome Data

- Result of previous campaign
- Categorical
- If this person previously signed up for a Term Deposit, it is far more likely they will sign up again



Employment Variation Rate

- Socio Economic Data
- Numeric
- Quarterly Indicator
- When this value is higher, less people have a term deposit

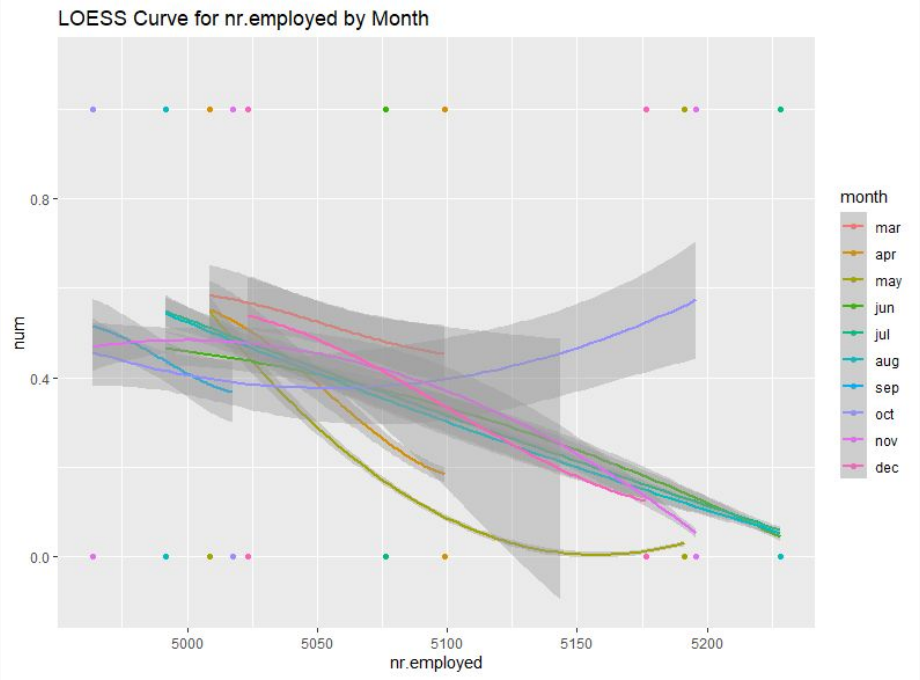


Socio-Economic Correlations

- Employment Variation Rate, Number Employed, and Euribor 3 month rate are all highly correlated

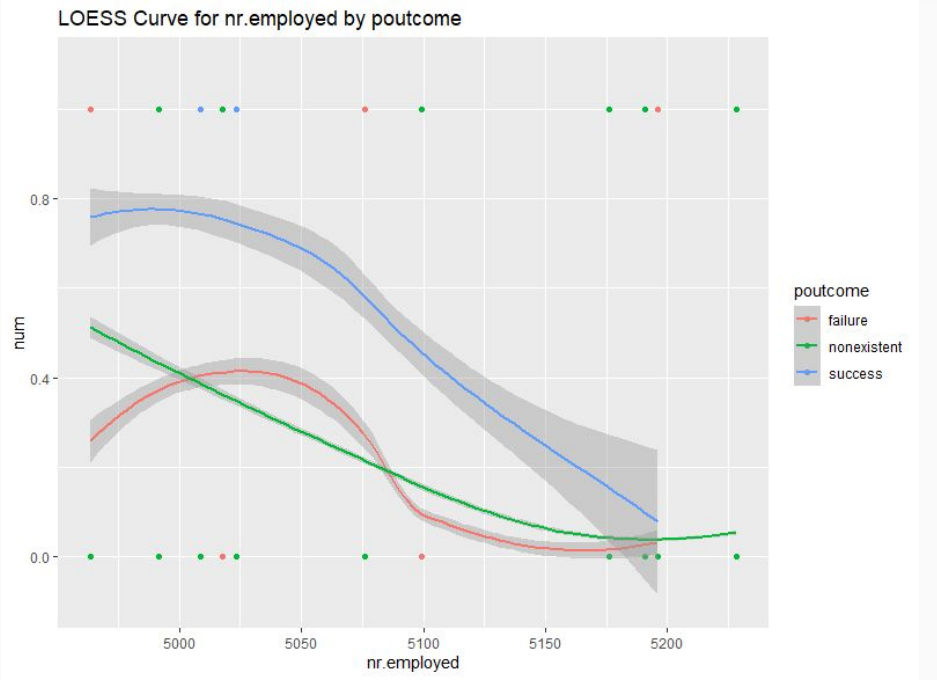
Number Employed by Month

- Nr.employed is highly correlated with emp.var.rate, so an increase will cause a decrease in Yes
- This seems to vary by Month though, where some months don't have very high nr.employed values at all
- This points to both these variables being useful in the model
    - Possible variable interaction



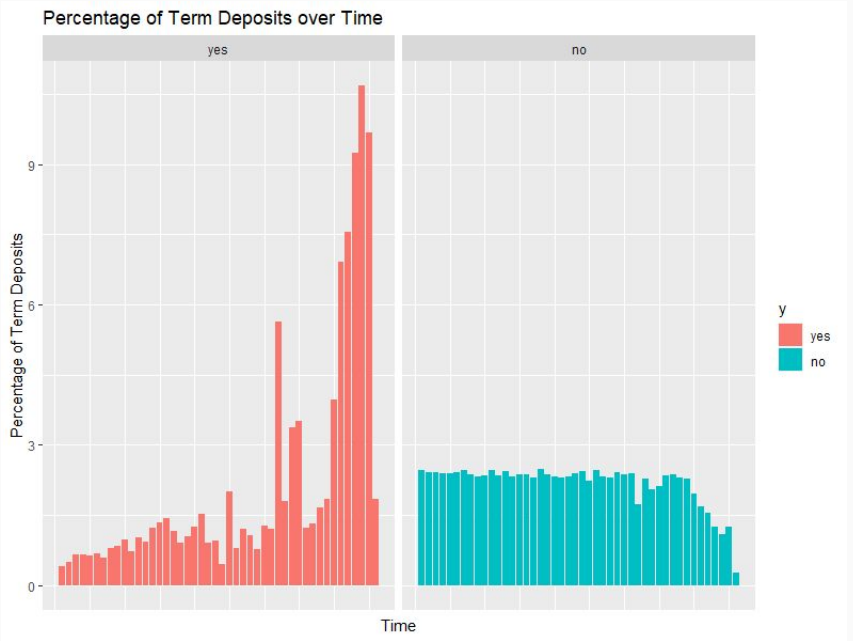
Number Employed by Poutcome

- All poutcome results show the same behavior, an increase in nr.employed leads to a decrease in Yes probability
- This decrease has the same rate roughly, but different starting points
- Points to both variables being useful in a model, but maybe no interaction



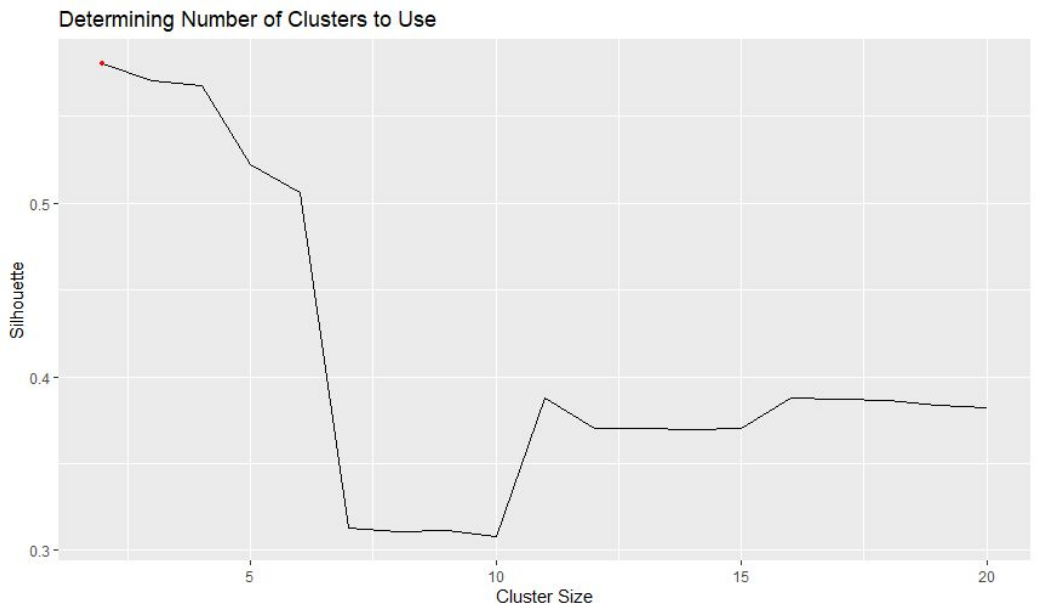
Data over time

- The data rows were in order that they were received
- We wanted to see if the distribution of data changed over time
- For Term Deposits, those became more common as time went on
- Wanted test data to have similar characters to training data, so used random 80/20 split



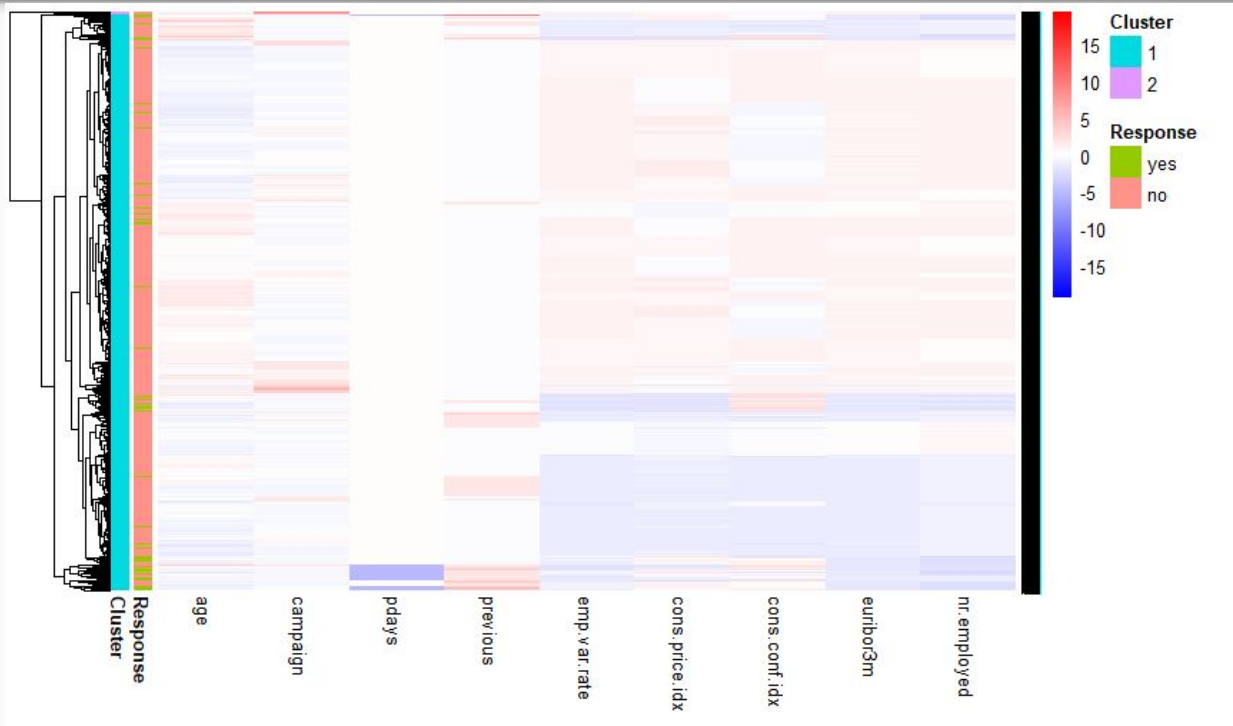
Clustering

- Attempted to cluster numeric variables
- Used Silhouette Statistic as metric for scoring clusters
- 2 clusters scored the highest, but there was an increase from 10 to 11 clusters
- Proceeded to investigate cluster sizes of 2 and 11



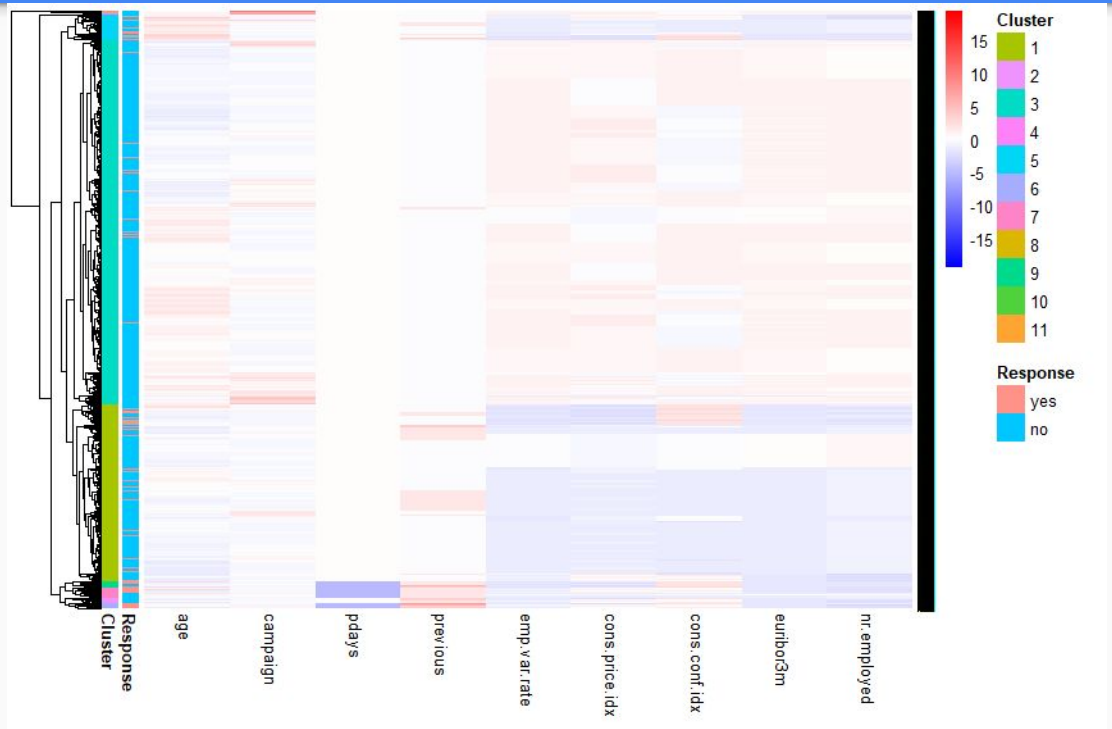
Heatmap - 2 Clusters

- Cluster is too small to be very useful



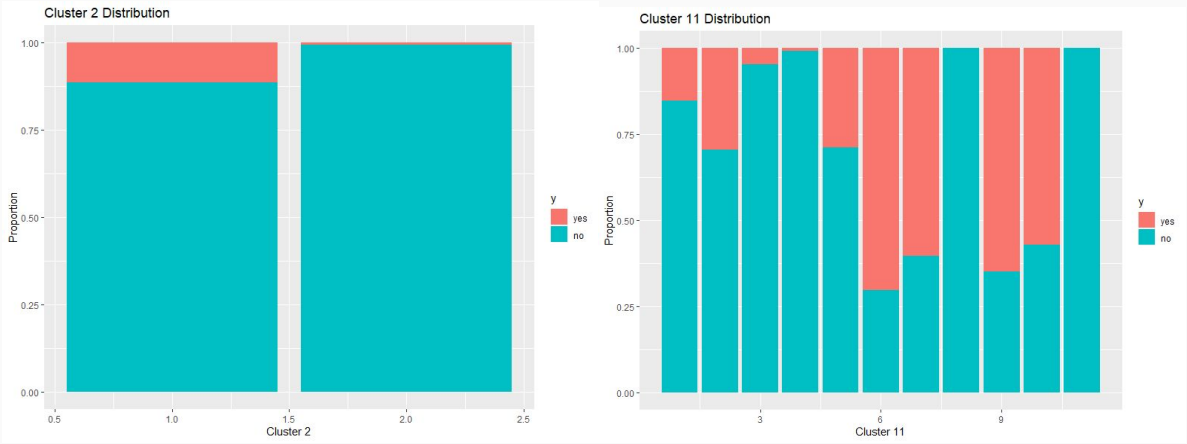
Heatmap - 11 Clusters

- 11 Clusters is more interesting
- Some clusters contain many ‘yes’ responses
- Many clusters seem to have a similar number of ‘yes’ to ‘no’ as the dataset as a whole does



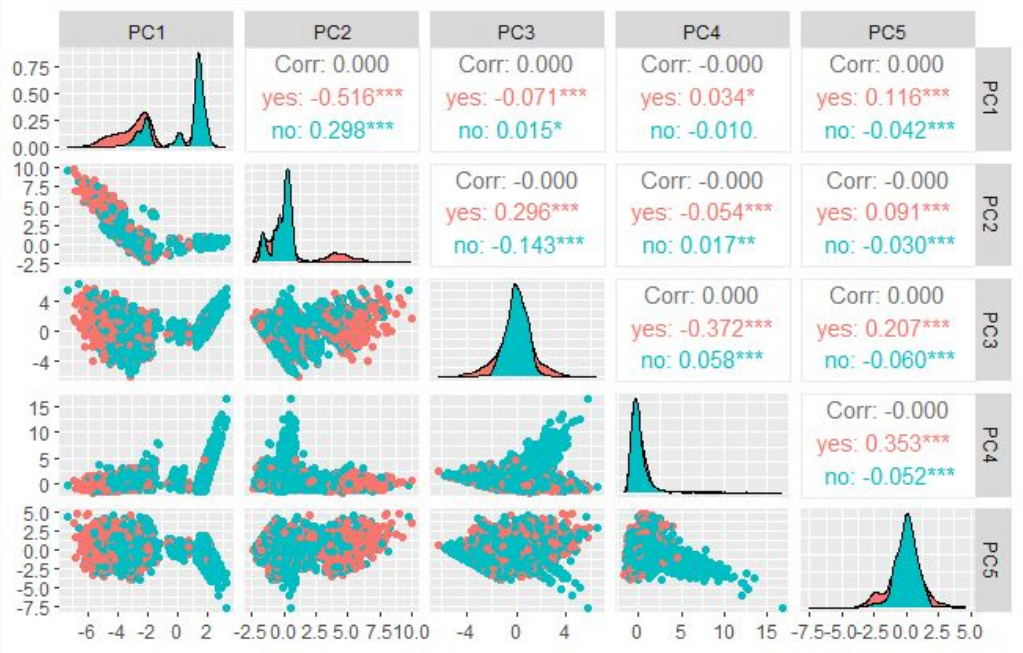
Heatmap - y Distribution

- Percentage of y varies by cluster
- Points to the ability of clusters to reasonably predict well



PCA Analysis

- Includes only numeric variables
- In order to retain at least 90% of the total variance 5 principal components were necessary to effectively represent the original data.
- PC1 had the best correlation with the response variable "y"
    - Even PC1 doesn't display a clear decision boundary



Objective 1

Simple Model Creation

- Forward/Backward Variable Selection
- CV using 10 folds
- Tried to maximize mean AUC (Area under the ROC Curve)
- y ~ month + poutcome + emp.var.rate + euribor3m + contact + cons.price.idx
- nr.employed removed due to high correlation

Variable	AUC
(add) nr.employed	0.7495
(add) month	0.7806
(add) poutcome	0.7874
(add) emp.var.rate	0.7885
(add) euribor3m	0.7897
(add) contact	0.7909
(add) cons.price.idx	0.7924
(remove) nr.employed	0.7929

Simple Model Creation - Correlation

- To reduce correlation, removed euribor3m
  - Correlation between emp.var.rate and euribor3m was 0.972
- Mean AUC is still reasonable, 0.7918
- All p values are extremely significant
- All VIF values are below 5
- y ~ month + poutcome + emp.var.rate + contact + cons.price.idx

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-108.53001	5.14668	-21.087	< 2e-16	***
monthapr	-1.36684	0.11411	-11.979	< 2e-16	***
monthmay	-1.88593	0.10846	-17.388	< 2e-16	***
monthjun	-1.47245	0.11859	-12.416	< 2e-16	***
monthjul	-1.02367	0.11750	-8.712	< 2e-16	***
monthaug	-0.69115	0.11645	-5.935	2.94e-09	***
monthsep	-1.01545	0.14351	-7.076	1.49e-12	***
monthoct	-1.03319	0.13677	-7.554	4.21e-14	***
monthnov	-1.44456	0.11940	-12.098	< 2e-16	***
monthdec	-0.55440	0.21012	-2.638	0.00833	**
poutcomenonexistent	0.43221	0.05976	7.232	4.75e-13	***
poutcomesuccess	1.81937	0.08629	21.083	< 2e-16	***
emp.var.rate	-0.82587	0.02213	-37.325	< 2e-16	***
contacttelephone	-0.43524	0.06009	-7.244	4.37e-13	***
cons.price.idx	1.14550	0.05491	20.863	< 2e-16	***

	2.5 %	97.5 %
(Intercept)	3.027764e-52	1.757006e-43
monthapr	2.037687e-01	3.187688e-01
monthmay	1.226234e-01	1.876321e-01
monthjun	1.817472e-01	2.893578e-01
monthjul	2.852998e-01	4.522838e-01
monthaug	3.987062e-01	6.294739e-01
monthsep	2.732696e-01	4.797154e-01
monthoct	2.720738e-01	4.651517e-01
monthnov	1.865509e-01	2.979492e-01
monthdec	3.804638e-01	8.678971e-01
poutcomenonexistent	1.371372e+00	1.733440e+00
poutcomesuccess	5.212077e+00	7.310388e+00
emp.var.rate	4.192491e-01	4.572424e-01
contacttelephone	5.748268e-01	7.275151e-01
cons.price.idx	2.823383e+00	3.501570e+00

	GVIF
month	6.253389
poutcome	1.288981
emp.var.rate	76.465448
contact	1.990009
cons.price.idx	11.218806
euribor3m	51.138803

	GVIF
month	2.394736
poutcome	1.274785
emp.var.rate	3.656646
contact	1.506607
cons.price.idx	3.284604

Model Summary

Variable Name	Odds Ratio Coefficient Value	95% Confidence Interval	P Value
monthapr	0.2549	(0.2038, 0.3188)	4.60 e-33
monthmay	0.1517	(0.1226, 0.1876)	1.02 e-67
monthjun	0.2294	(0.1817, 0.2894)	2.14 e-35
monthjul	0.3593	(0.2853, 0.4523)	2.99 e-18
monthaug	0.5010	(0.3987, 0.6295)	2.94 e-09
monthsep	0.3622	(0.2733, 0.4797)	1.49 e-12
monthoct	0.3559	(0.2721, 0.4652)	4.21 e-14
monthnov	0.2358	(0.1866, 0.2979)	1.08 e-33
monthdec	0.5744	(0.3805, 0.8679)	0.0083
Poutcomenonexistent	1.5407	(1.3714, 1.7334)	1.13 e-98
Poutcomesuccess	6.1679	(5.2121, 7.3104)	1.33 e-71
emp.var.rate	0.4379	(0.4192, 0.4572)	6.48 e-305
contacttelephone	0.6471	(0.5748, 0.7275)	4.37 e-13
cons.price.idx	3.1440	(2.8234, 3.5016)	1.16 e-96

Model Interpretation

month (Mar v Apr): The odds of convincing the clients to subscribe to a term deposit in the month of April decreases by a factor of 0.25 when compared to the month of March, all other variables remaining constant. The 95% Confidence interval of the decrease is (0.20, 0.32).

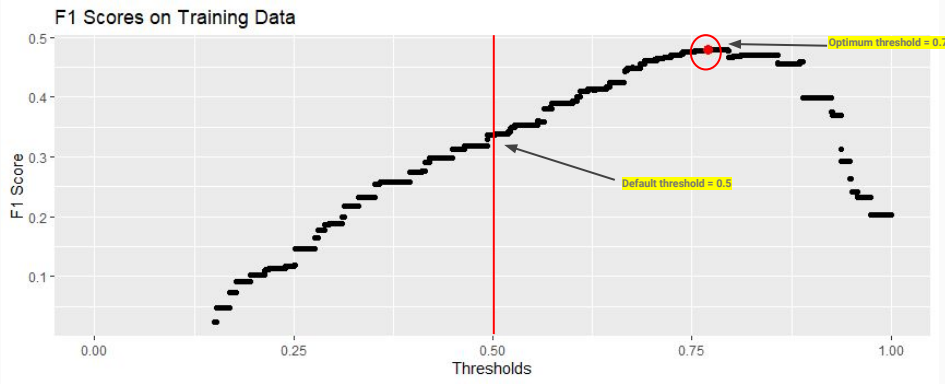
emp.var.rate: For every 1 unit increase in the employment variation rate, the odds of convincing the customer to subscribe to a term deposit decrease by a factor of 0.44, all other variables remaining constant. The 95% Confidence interval of the decrease is (0.42, 0.46).

Evaluation Metrics

- Sensitivity, Specificity, PPV, NPV, and AUROC/AUC were calculated for each model
- Since Term Deposit = ‘yes’ means more money for the bank, making sure the model predicts ‘yes’ values well is the most important
- Sensitivity = Given the value is ‘yes’, model predicted ‘yes’
- PPV = Given model predicted ‘yes’, value is ‘yes’
- Used F1 since that is a useful way to measure positive prediction efficacy
    - F1 = 2 * Sensitivity * PPV / (Sensitivity + PPV)
    - Best value is 1

Threshold Value

- 0.5 is the default Threshold to use for the yes/no decision, but isn't always the best
- Got threshold value to use on Training data, but used it on Test data
- Threshold to maximize F1 Score
- Plot shows the F1 Scores for different Thresholds for the Simple Logistic Regression Model
- Sensitivity decreases with threshold, PPV increases



Simple Model Training Confusion Matrix

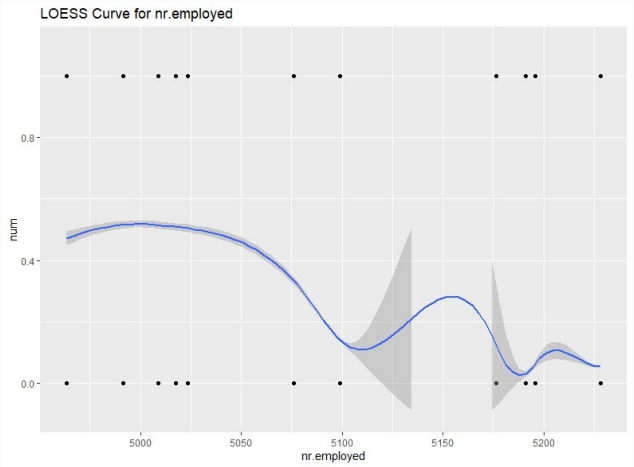
- AUC for training data was 0.7868
- Threshold for maximum F1 was 0.7703
  - Confusion matrix was obtained from this threshold
- Highest F1 score was 0.4798

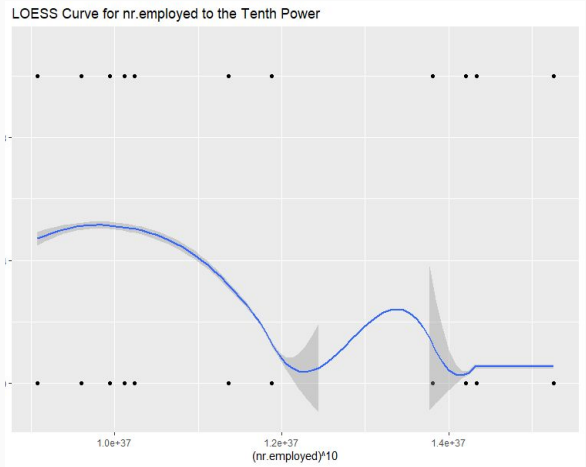
```
          Reference
Prediction   yes    no
       yes  1902  2297
       no   1827 26924

               Accuracy : 0.8748
                 95% CI : (0.8712, 0.8784)
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : 1

                  Kappa : 0.409

 Mcnemar's Test P-Value : 2.81e-13

            Sensitivity : 0.51006
            Specificity : 0.92139
         Pos Pred Value : 0.45296
         Neg Pred Value : 0.93645
             Prevalence : 0.11317
         Detection Rate : 0.05772
   Detection Prevalence : 0.12744
      Balanced Accuracy : 0.71572

       'Positive' Class : yes
```

Objective 2

Adding Polynomial Terms to Model

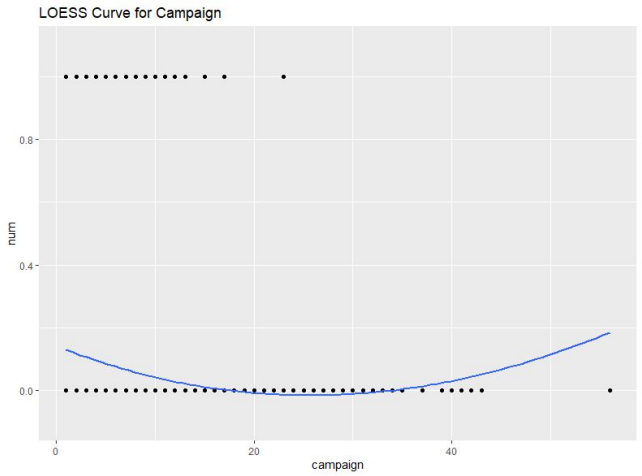
- Taking polynomial terms for nr.employed seems to lead to more stability when nr.employed is increased
- Decided to investigate polynomials in variable selection

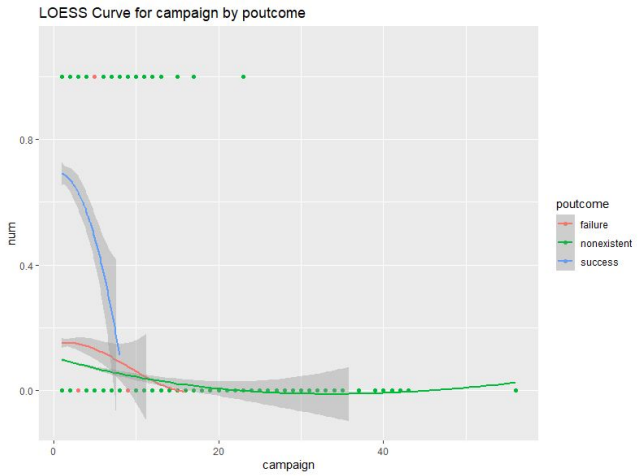




Campaign by poutcome

- Campaign by itself seems that there is no constant trend
- Certain poutcome values lead to a more stable behavior though
- Points to a potential interaction term





Complex Logistic Regression Model Creation

- Similar approach to simple model (Forward/Backward Variable Selection, CV using 10 folds, Mean AUC as error metric)
- Added single variable, polynomial variables, and interaction variables
- y ~ poly(cons.conf.idx,10) + pdays + day_of_week*month + month*contact + cons.conf.idx*housing + poutcome*previous + poly(campaign,5) + poly(euribor3m,8) + campaign*month + cons.conf.idx*age + poly(previous,6) + campaign*contact + poly(age,3)

Variable	AUC
(add) poly(nr.employed,9)	0.7724
(add) poly(euribor3m,6)	0.7832
(add) contact	0.7893
(add) poutcome	0.7949
(add) poly(cons.conf.idx,10)	0.7965
(add) pdays	0.7973
(add) poly(age,2)	0.7979
…	…
(add) campaign*month	0.8018
(remove) poutcome*campaign	0.8023
(add) cons.conf.idx*age	0.8024
(add) poly(previous,6)	0.8026
(add) campaign*contact	0.8030
(add) poly(age,3)	0.8036

Complex Logistic Regression Confusion Matrix

- AUC for training data was 0.8013
- Threshold for highest F1 was 0.7354
- Highest F1 score was 0.5073
- All relevant metrics improved from simple logistic model

```
          Reference
Prediction   yes    no
       yes  2008  2179
       no   1721 27042

               Accuracy : 0.8816
                 95% CI : (0.8781, 0.8851)
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : 0.9985

                  Kappa : 0.4403

 Mcnemar's Test P-Value : 2.52e-13

            Sensitivity : 0.53848
            Specificity : 0.92543
         Pos Pred Value : 0.47958
         Neg Pred Value : 0.94017
             Prevalence : 0.11317
         Detection Rate : 0.06094
   Detection Prevalence : 0.12707
      Balanced Accuracy : 0.73196

       'Positive' Class : yes
```

LDA/QDA Confusion Matrix

LDA model

- AUC for training data was 0.7555
- Threshold for highest F1 was 0.2306
- Highest F1 score was 0.4561
- Used only numeric variables–y ~ emp.var.rate + cons.price.idx + euribor3m

```
Confusion Matrix and Statistics

          Reference
Prediction   yes    no
       yes  1683  1968
       no   2046 27253

               Accuracy : 0.8782
                 95% CI : (0.8746, 0.8817)
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : 1.0000

                  Kappa : 0.3875

 Mcnemar's Test P-Value : 0.2242

            Sensitivity : 0.45133
            Specificity : 0.93265
         Pos Pred Value : 0.46097
         Neg Pred Value : 0.93017
             Prevalence : 0.11317
         Detection Rate : 0.05108
   Detection Prevalence : 0.11080
      Balanced Accuracy : 0.69199

       'Positive' Class : yes
```

LDA/QDA Confusion Matrix

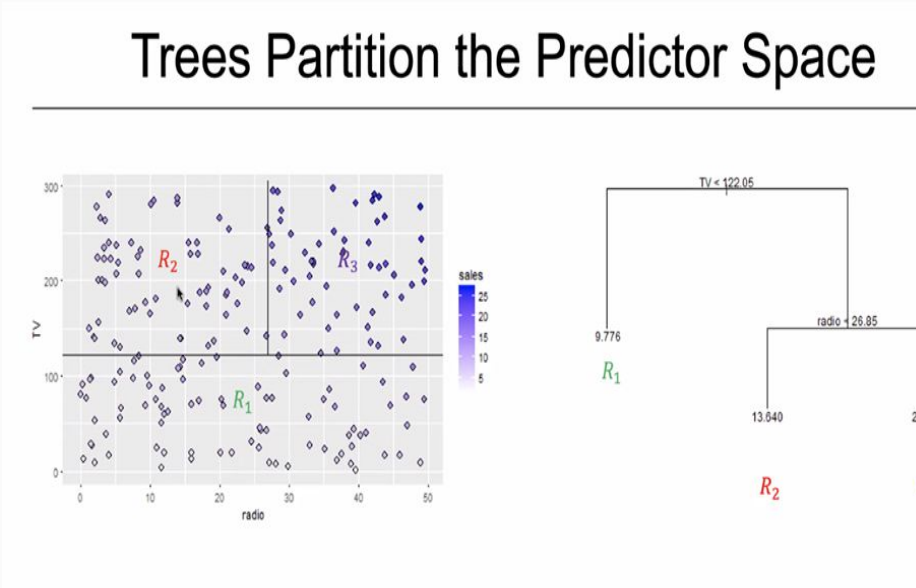
QDA model

- AUC for training data was 0.7694
- Threshold for highest F1 was 0.1264
- Highest F1 score was 0.4681
- Used only numeric variables–y ~ emp.var.rate + cons.price.idx + euribor3m

```
Confusion Matrix and Statistics

          Reference
Prediction   yes    no
       yes  1833  2270
       no   1896 26951

               Accuracy : 0.8736
                 95% CI : (0.8699, 0.8771)
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : 1

                  Kappa : 0.3965

 Mcnemar's Test P-Value : 7.517e-09

            Sensitivity : 0.49155
            Specificity : 0.92232
         Pos Pred Value : 0.44675
         Neg Pred Value : 0.93427
             Prevalence : 0.11317
         Detection Rate : 0.05563
   Detection Prevalence : 0.12452
      Balanced Accuracy : 0.70693

       'Positive' Class : yes
```

Random Forest Introduction

- Decision Trees split up Explanatory Variables into partitions
- Random Forest is a collection of Decision Trees
- Non-parametric Model
- Hyper-parameters
  - Mtry
    - Amount of variables randomly selected each split
  - Ntree
    - Amt of decision trees in the model
- Random Forests are generated differently each time



Random Forest: Confusion Matrix

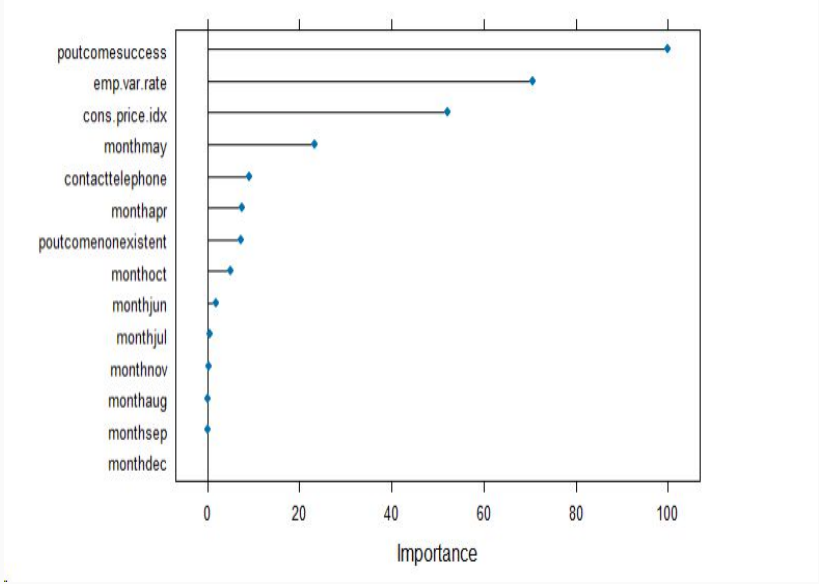
Caret Package

- Finds the optimum mtry and ntree value
- AUC for training data was 0.6685
- Threshold for highest F1 was 0.0040 (really low)
- Highest F1 score was 0.4593
- We used the simple Logistic model equation: y ~ month + poutcome + emp.var.rate + contact + cons.price.idx

```
Confusion Matrix and Statistics

          Reference
Prediction  yes   no
       yes  348  330
       no   563 6997

               Accuracy : 0.8916
                 95% CI : (0.8847, 0.8982)
    No Information Rate : 0.8894
    P-Value [Acc > NIR] : 0.2703

                  Kappa : 0.3795

 Mcnemar's Test P-Value : 8.256e-15

            Sensitivity : 0.38200
            Specificity : 0.95496
         Pos Pred Value : 0.51327
         Neg Pred Value : 0.92553
             Prevalence : 0.11059
         Detection Rate : 0.04224
   Detection Prevalence : 0.08230
      Balanced Accuracy : 0.66848

       'Positive' Class : yes
```

Random Forest: Contribution Plots

- The 3 most important variables based on the overall accuracy to the models are Poutcome, emp.var.rate, cons.price.idx.



Random Forest: Confusion Matrix

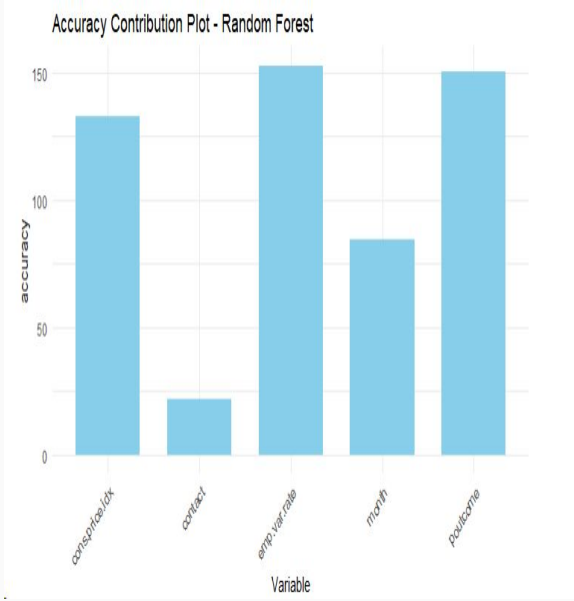
randomForest

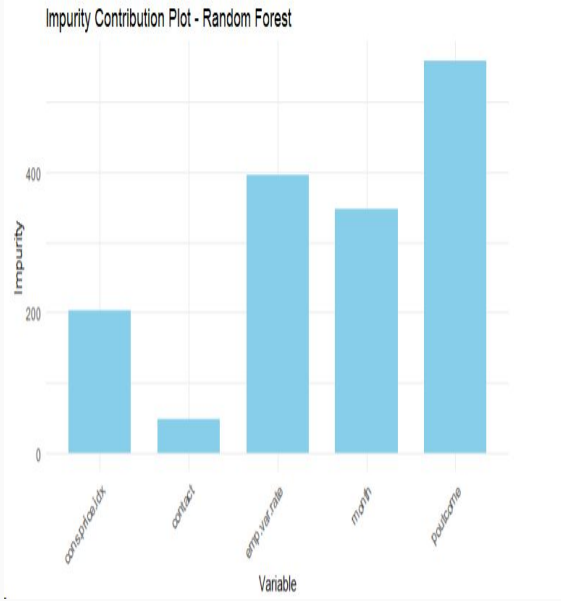
- Hyperparameters
  - Ntree = 3
  - Mtry = 5000
- We used the simple Logistic model equation: y ~ month + poutcome + emp.var.rate + contact + cons.price.idx
- Threshold for highest F1 was 0.0034 (really low)
- Highest F1 score was 0.4605
- AUC for training data was 0.6703
  - AUC is 0.3% > AUC of caret package.

```
Confusion Matrix and Statistics

          Reference
Prediction  yes   no
       yes  352  335
       no   559 6992

               Accuracy : 0.8915
                 95% CI : (0.8846, 0.8981)
    No Information Rate : 0.8894
    P-Value [Acc > NIR] : 0.2821

                  Kappa : 0.3818

 Mcnemar's Test P-Value : 8.769e-14

            Sensitivity : 0.38639
            Specificity : 0.95428
         Pos Pred Value : 0.51237
         Neg Pred Value : 0.92597
             Prevalence : 0.11059
         Detection Rate : 0.04273
   Detection Prevalence : 0.08339
      Balanced Accuracy : 0.67033

       'Positive' Class : yes
```

Random Forest: Contribution Plots

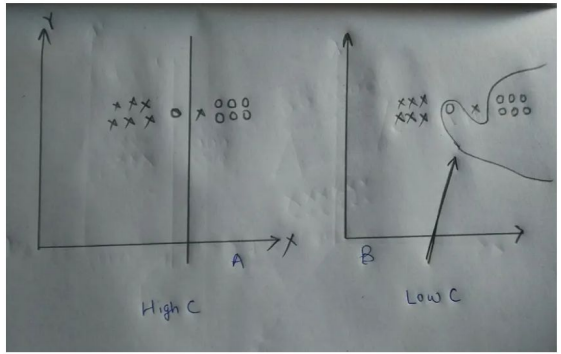
- The 3 most important variables based on the overall accuracy and impurity to the models are Poutcome, emp.var.rate, cons.price.idx.

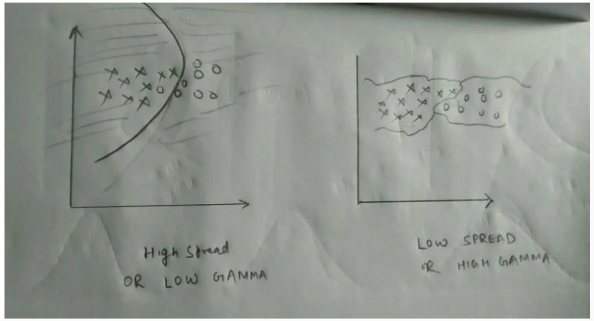




Support Vector Machines (SVM)

- Non-parametric model for both regression and classification
- Creates a boundary (line, plane, or hyper-plan) of best fit to classify the data points
- Hyper-parameters for model creation:
  - Kernel = type of curve (line, polynomial, radial, etc.)
  - Gamma = affects the fluctuations in the boundary
  - Cost = affects the bias / variance trade-off





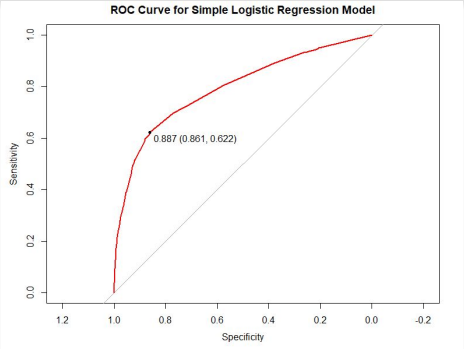
- Pictures from: https://medium.com/@myselfaman12345/c-and-gamma-in-svm-e6cee48626be

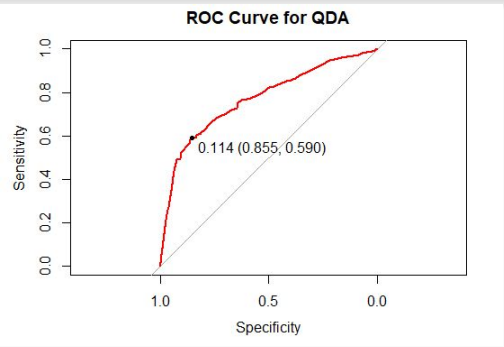
SVM Training Confusion Matrix

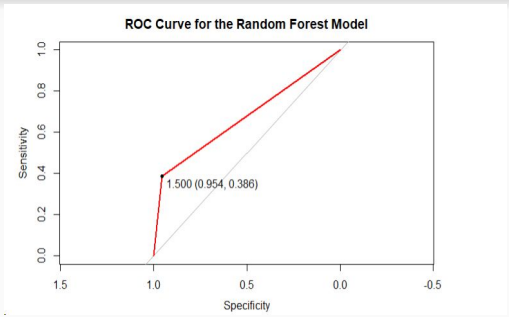
- We used y ~ euribor3m + month + poutcome + contact + cons.conf.idx + campaign + previous + age + housing + day_of_week
  - All the variables from the complex model, just none of the complexity
- Cost = 10 (less variance)
- Gamma = 1 (more fluctuations)
- Thresholds were very low (0.08-0.10)
- F1 = 0.65, AUC = 0.8513, highest values we've seen on the training data
- Sensitivity = 0.6163, PPV = 0.6901
  - Usually Sensitivity > PPV

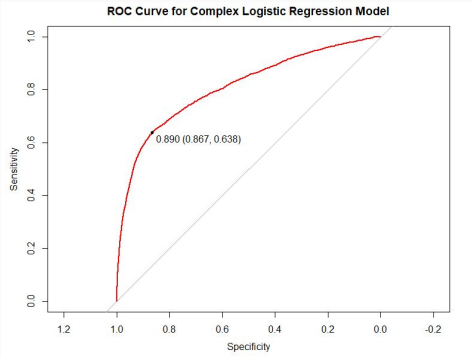
```
          Reference
Prediction   yes    no
       yes  2298  1032
       no   1431 28189

               Accuracy : 0.9253
                 95% CI : (0.9224, 0.9281)
    No Information Rate : 0.8868
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6094

 Mcnemar's Test P-Value : 1.061e-15

            Sensitivity : 0.61625
            Specificity : 0.96468
         Pos Pred Value : 0.69009
         Neg Pred Value : 0.95169
             Prevalence : 0.11317
         Detection Rate : 0.06974
   Detection Prevalence : 0.10106
      Balanced Accuracy : 0.79047

       'Positive' Class : yes
```

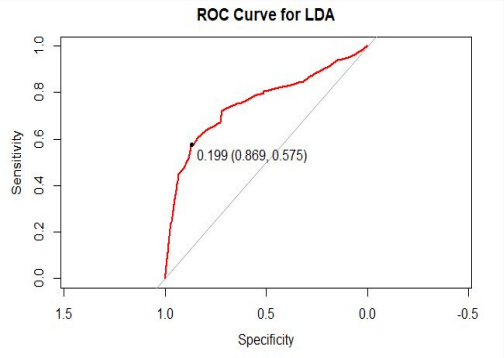
ROC Curves

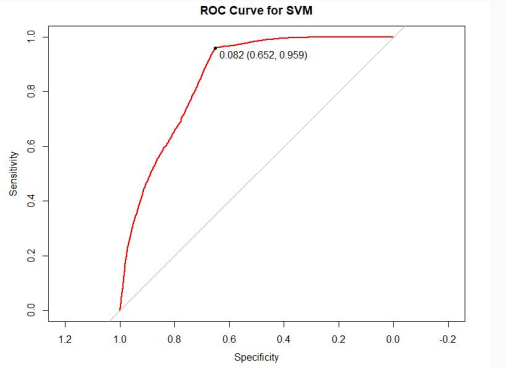

ROC Curve for Simple Logistic Regression Model
Sensitivity
Specificity
0.887 (0.861, 0.622)


ROC Curve for QDA
Sensitivity
Specificity
0.114 (0.855, 0.590)


ROC Curve for the Random Forest Model
Sensitivity
Specificity
1.500 (0.954, 0.386)


ROC Curve for Complex Logistic Regression Model
Sensitivity
Specificity
0.890 (0.867, 0.638)


ROC Curve for LDA
Sensitivity
Specificity
0.199 (0.869, 0.575)


ROC Curve for SVM
Sensitivity
Specificity
0.082 (0.652, 0.959)

Model Validations

Model	Sensitivity	Specificity	Prevalence	PPV	NPV	F1	AUROC
Simple Logistic Regression	0.4929	0.9221	0.1106	0.4402	0.9360	0.4650	0.7868
Complex Logistic Regression	**0.5280**	0.9249	0.1106	0.4665	**0.9403**	**0.4954**	**0.8013**
LDA Model	**0.4522**	0.9335	0.1106	0.4583	**0.9320**	0.4552	0.7506
QDA Model	0.4829	0.9255	0.1106	0.4463	0.9351	0.4639	0.7623
Random Forest Model	0.3863	0.9573	0.1106	0.5124	0.9259	0.4406	0.6704
Support Vector Machine (SVM)	0.5236	**0.8705**	0.1106	**0.3345**	0.9363	**0.4082**	**0.6979**

Conclusions

Model Comparisons

- Complex Logistic Regression had the highest Sensitivity, NPV, F1 score, and AUROC
- Logistic models easier to work with
- The Random Forest model had lowest Sensitivity but highest Specificity
- LDA has worse Sensitivity than Simple Logistic Model
- SVM had the worst Specificity, PPV, F1, and worst drop off between training and test results (overfit)
- Non-parametric models took the longest to train
- No models really did “well”, most had PPV < 50%
- Sensitivity was always much less than Specificity, due to the high No count

Model Validations on future data

- Tried training on older data and testing on newer data
  - Originally did a random 80/20 split of the data
- Training F1 was worse (~0.25, about 50% worse than previously)
- Needed to filter test data and simplify complex model due to missing values that were present in training
- Testing Sensitivity, PPV, and F1 was better, but Specificity, NPV, and AUROC was worse
- Simple model performed better than complex model

Model	Sensitivity	Specificity	Prevalence	PPV	NPV	F1	AUROC
Simple Logistic Regression	0.4929	0.9221	0.1106	0.4402	0.9360	0.4650	0.7868
Simple Logistic Regression (Future)	0.5661	0.7740	0.2979	0.5151	0.8079	0.5394	0.7095
Complex Logistic Regression	0.5280	0.9249	0.1106	0.4665	0.9403	0.4954	0.8013
Complex Logistic Regression (Future)	0.6287	0.6716	0.2979	0.4482	0.8100	0.5233	0.6502

Tips for next Campaign

- March is the best month to call in, but August and December aren't bad either
- Call people if they subscribed to a Term Deposit previously
- Call more when socio-economic indicators are low

Future Work

- Interpretive analysis with Duration variable
- Oversampling the Yes data to improve Sensitivity
- Variable Selection tailored for the Random Forest Model
- Re-visit SVM to reduce out of sample error
- Ensemble model

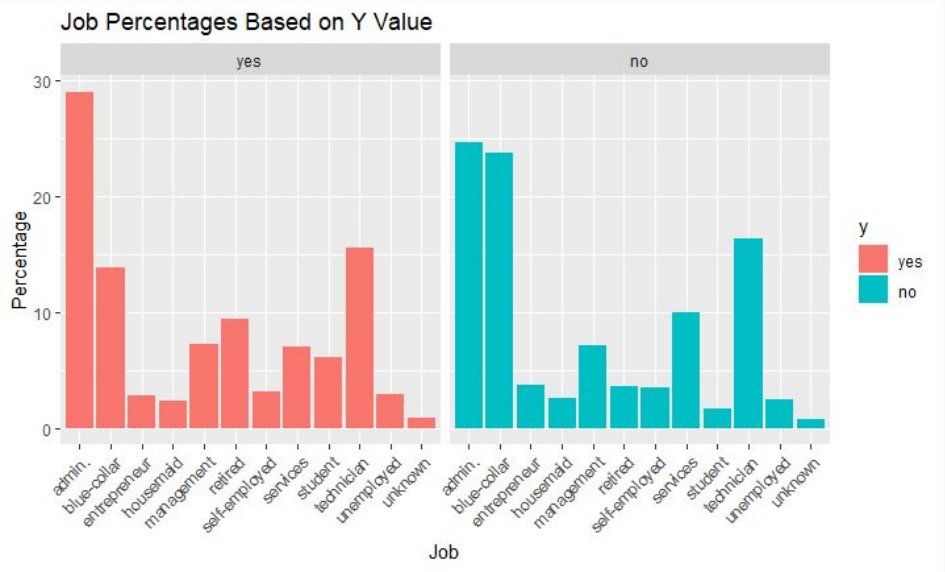
Thank you!

- Aaron Abromowtiz: aabromowitz@mail.smu.edu
- Stephanie Duarte: duartes@mail.smu.edu
- Dammy Owolabi: oowolabi@smu.edu

Backup

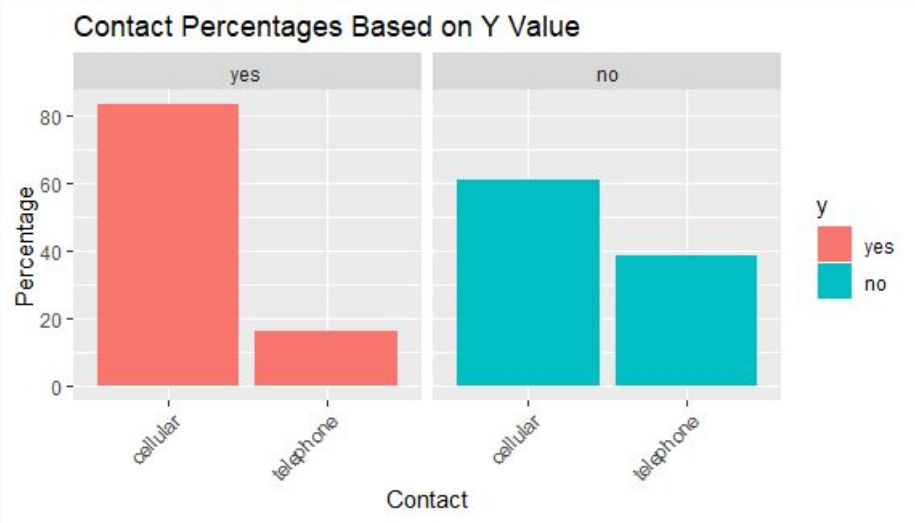
Job Data

- Bank client data
- Categorical
- There are jobs where the amount of yes vs no changes drastically
  - Ex: admin, blue-collar, retired, services, student
- Overall, not clear if this will be a good variable for prediction



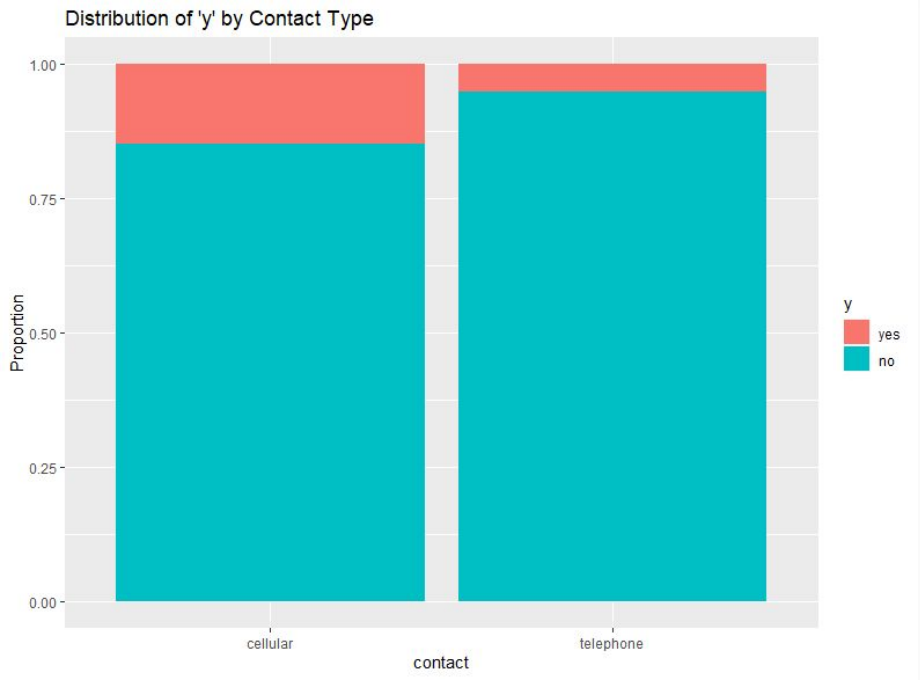
Contact Data

- Related with the last contact of the current campaign
- Categorical
- Last contact approach
- Contact over cell phone was ~20% more likely to lead to a term deposit



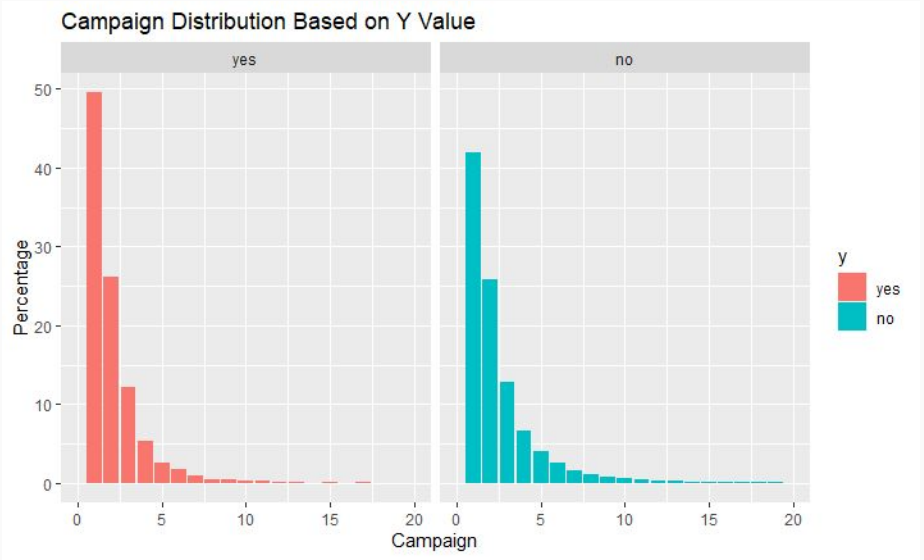
Contact Data

- Related with the last contact of the current campaign
- Categorical
- Last contact approach
- Contact over cell phone was ~20% more likely to lead to a term deposit



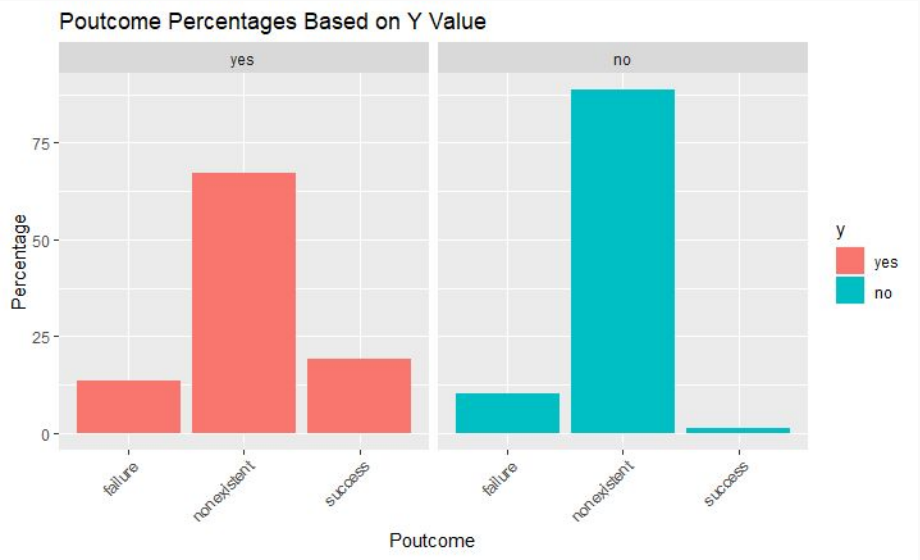
Campaign Data

- Number of contacts performed during this campaign and for this client
- Categorical
- Both have similar, exponential decay distributions



Poutcome Data

- Outcome of the previous marketing campaign
- Categorical
- Success is more common for clients with a Term Deposit



Consumer Price Index

- Socio Economic Data
- Numeric
- Monthly Indicator
- The value seems more evenly distributed for people that have a term deposit

PCA Analysis

- Removed numeric variables- age, campaign, and cons.conf.idx
- In order to retain at least 90% of the total variance 4 principal components were necessary to effectively represent the original data.
- PC1 had the best correlation with the response variable "y”

Model Interpretation

month (Mar v Apr): The odds of convincing the clients to subscribe to a term deposit in the month of April decreases by a factor of 0.25 when compared to the month of March, all other variables remaining constant. The 95% Confidence interval of the decrease is (0.20, 0.32).

poutcome (success v failure): The odds of convincing a client to subscribe to a term deposit who was successfully subscribed in the previous campaign is 6.17 times higher than that of someone who was not successfully subscribed, all other variables remaining constant. The 95% Confidence interval of the factor is (5.21, 7.31).

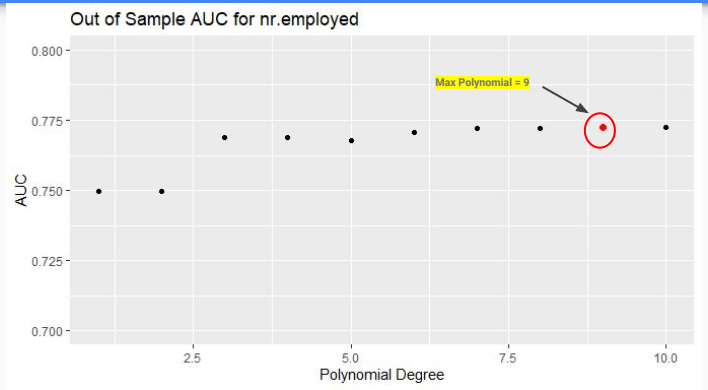
contact (cellular v telephone): The odds of convincing the client to subscribe to a term deposit when calling their landline phone decreased by a factor of 0.65 when compared to calling their cell phone, all other variables remaining constant. The 95% Confidence interval of the decrease is (0.57, 0.73).

emp.var.rate: For every 1 unit increase in the employment variation rate, the odds of convincing the customer to subscribe to a term deposit decrease by a factor of 0.44, all other variables remaining constant. The 95% Confidence interval of the decrease is (0.42, 0.46).

cons.price.idx: For every 1 unit increase in Consumer Price Index, the odds of convincing the customer to subscribe to a term deposit increase by 214%, all other variables remaining constant. The 95% Confidence of the increase is (182%, 250%).

Adding Polynomial Terms to Model

- Adding Polynomial terms to Model seemed to improve AUC
- This was observed for out of sample metrics
- Improvement seems to happen after 3 polynomial terms
- P values remaining significant
- Decided to investigate polynomials in variable selection




```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                2.44760    0.02338 104.676  < 2e-16 ***
poly(nr.employed, 10)1   156.34441    3.19323  48.961  < 2e-16 ***
poly(nr.employed, 10)2   -44.45416    3.14126 -14.152  < 2e-16 ***
poly(nr.employed, 10)3   -51.90203    4.32051 -12.013  < 2e-16 ***
poly(nr.employed, 10)4     4.92844    2.83743   1.737 0.082398 .
poly(nr.employed, 10)5    26.51619    2.34879  11.289  < 2e-16 ***
poly(nr.employed, 10)6    30.20002    3.42973   8.805  < 2e-16 ***
poly(nr.employed, 10)7     8.50557    3.68814   2.306 0.021100 *
poly(nr.employed, 10)8    -2.79958    2.14882  -1.303 0.192628
poly(nr.employed, 10)9    11.69125    3.00755   3.887 0.000101 ***
poly(nr.employed, 10)10   -7.20951    2.02859  -3.554 0.000379 ***
```

Adding Interaction Terms to Model

- Some variable interactions showed different distributions of Term Deposits depending on the combination
- Ex: Tue/Mar has more yes than Tue/May
- Ex: Thu/Apr has more yes than Mon/Apr
- Decided to try interactions in variable selection