Data Glacier

Your Deep Learning Partner

Bank Purchase Classification Case Study

Final Presentation

Farha Jabin Oyshee, Devika Chandnani, Dylan Huey, Camillo

28-Nov-2022

Background – Bank Purchase Classification case study

- ABC Bank wants to sell it's term deposit product to customers and before launching the product they want to develop a model which help them in understanding whether a particular customer will buy their product or not (based on customer's past interaction with bank or other Financial Institution).

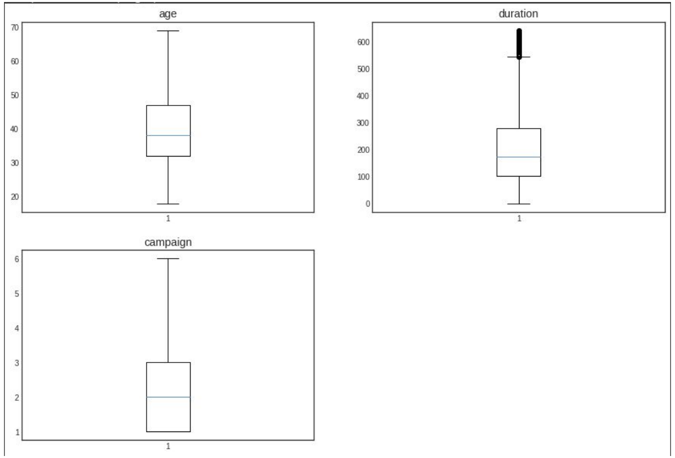
- Objective: Analyze previous bank customer data to propose an efficient solution for ABC banks upcoming marketing campaign. Identify trends in the data to ultimately create a model to help predict which customers will be most likely to purchase the new product

The analysis has been divided into four parts:

- Data Understanding
- Finding target groups
  - How we found the target groups
- Recommendations for model building

Background – Data Cleaning and Outlier Removal

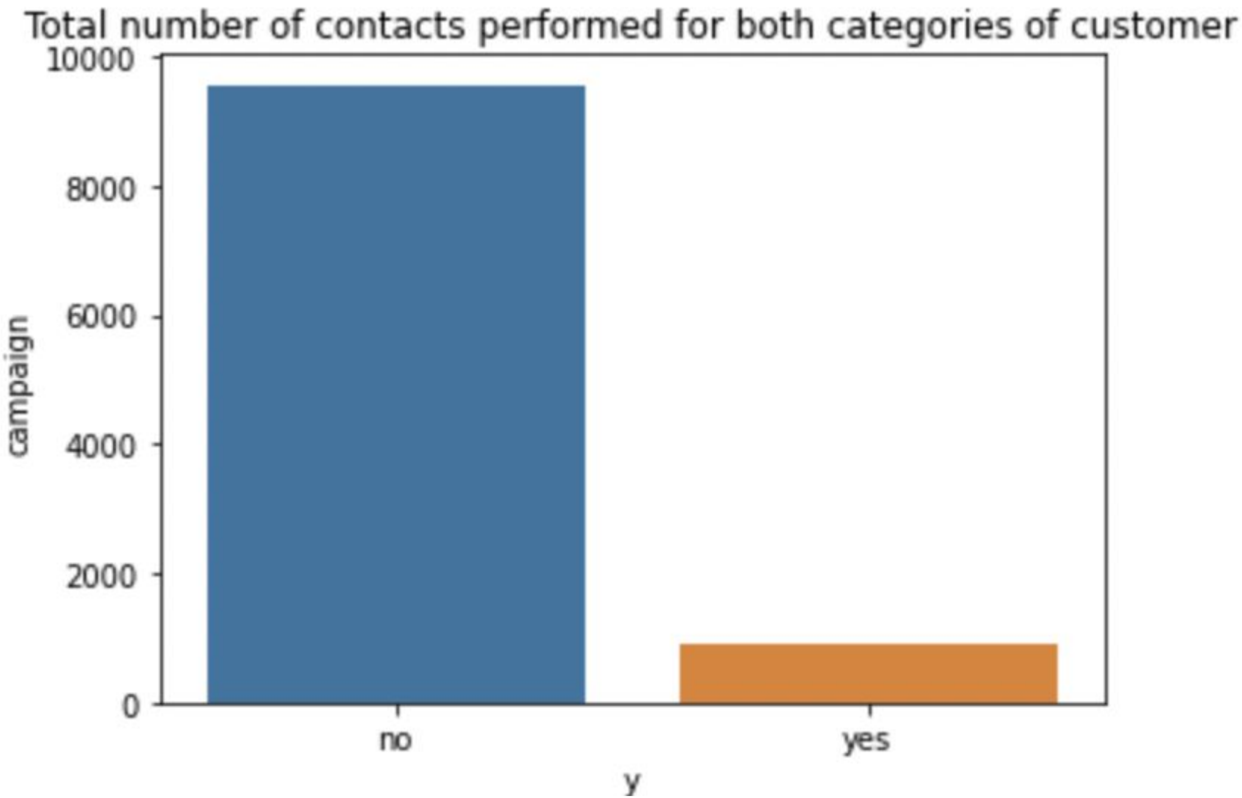
- The dataset already came without any unusable data points and was able to be used immediately
    - We have provided a screenshot of the number of null columns after importing the data
- There were some outliers within the age and campaign categories and replaced their values with the upper and lower IQR boundaries



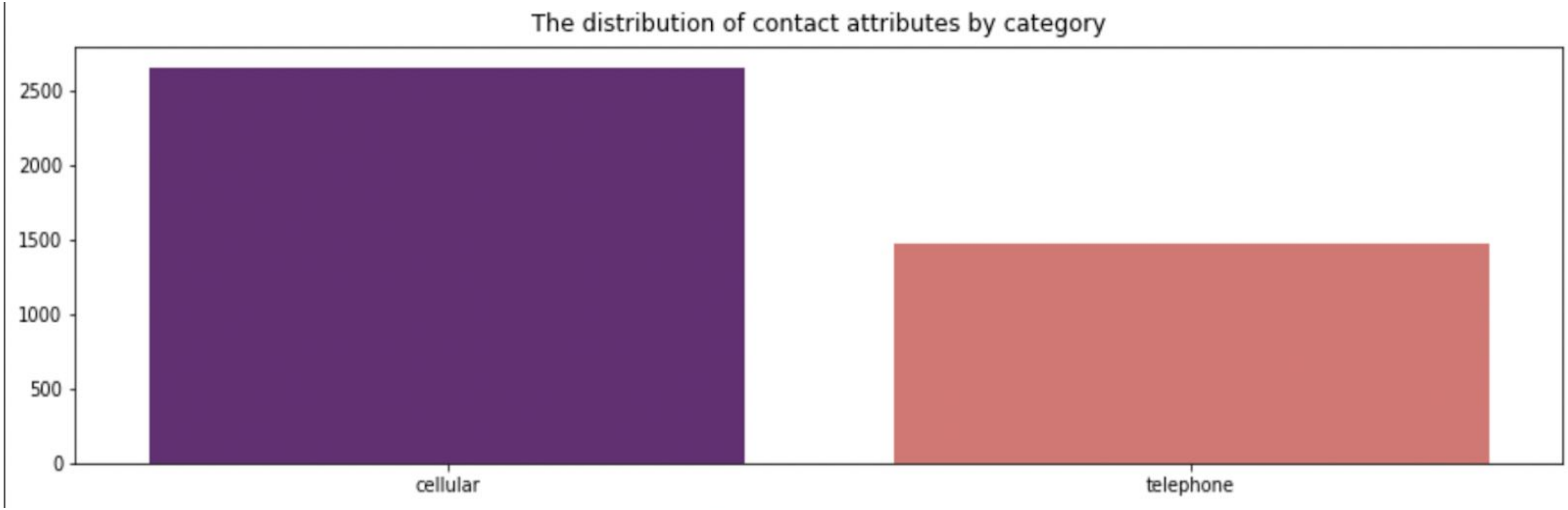

```
bank_additional_full.isnull().sum()
```

```
age               0
job               0
marital           0
education         0
default           0
housing           0
loan              0
contact           0
month             0
day_of_week       0
duration          0
campaign          0
pdays             0
previous          0
poutcome          0
emp.var.rate      0
cons.price.idx    0
cons.conf.idx     0
euribor3m         0
nr.employed       0
y                 0
dtype: int64
```

Data Understanding



Data Understanding - Campaign types

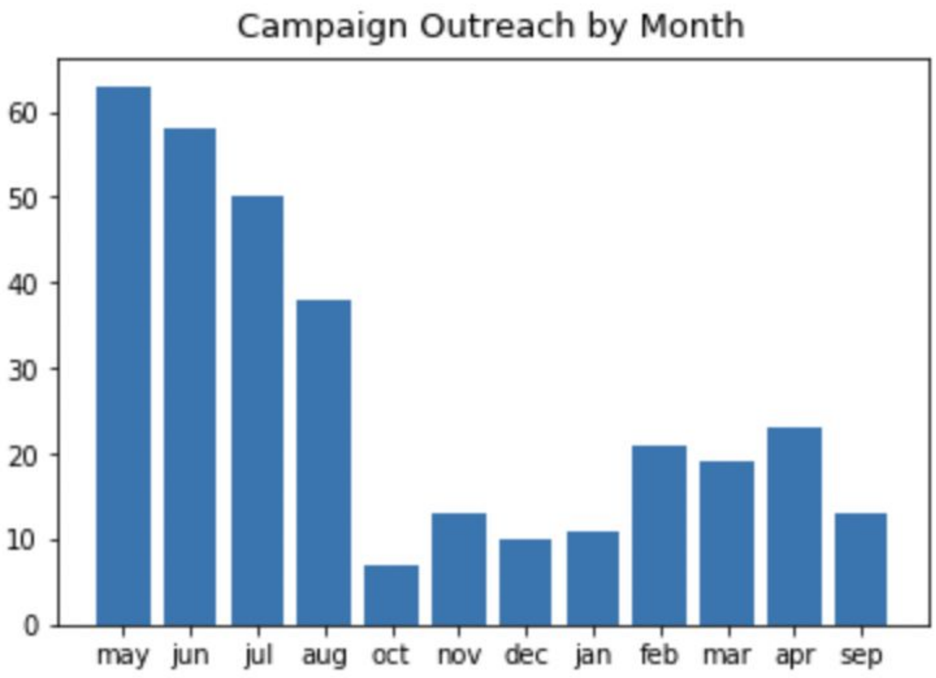


This graph shows the comparison of campaign reach by category and that the campaign reaches more than 50% more customers on a mobile phone compared to a telephone. This will help when determining what the target demographic will be.

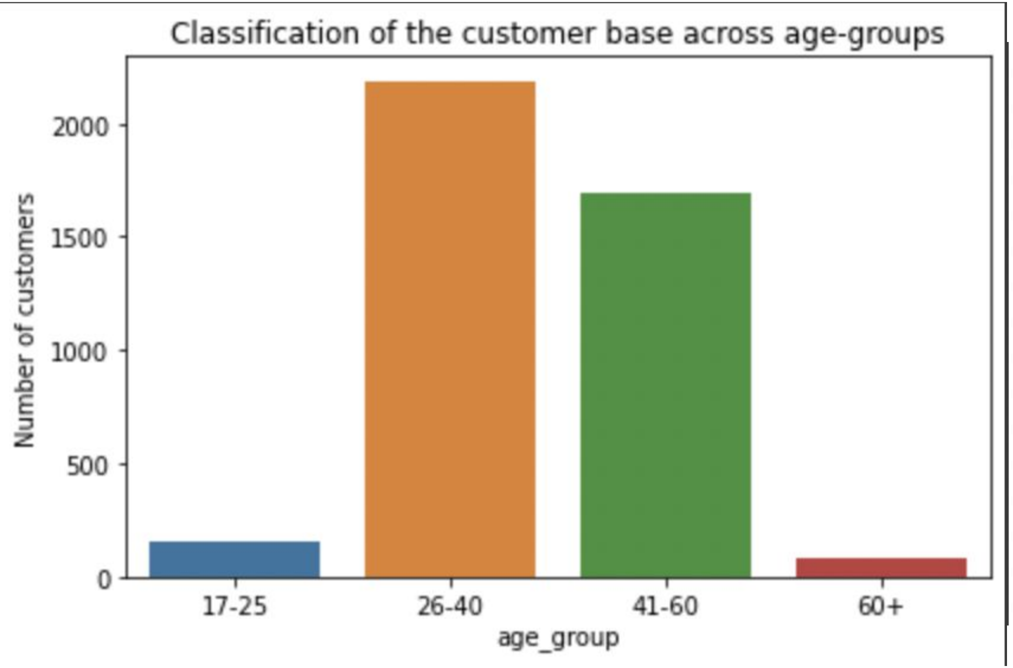
Data Understanding - Campaign types

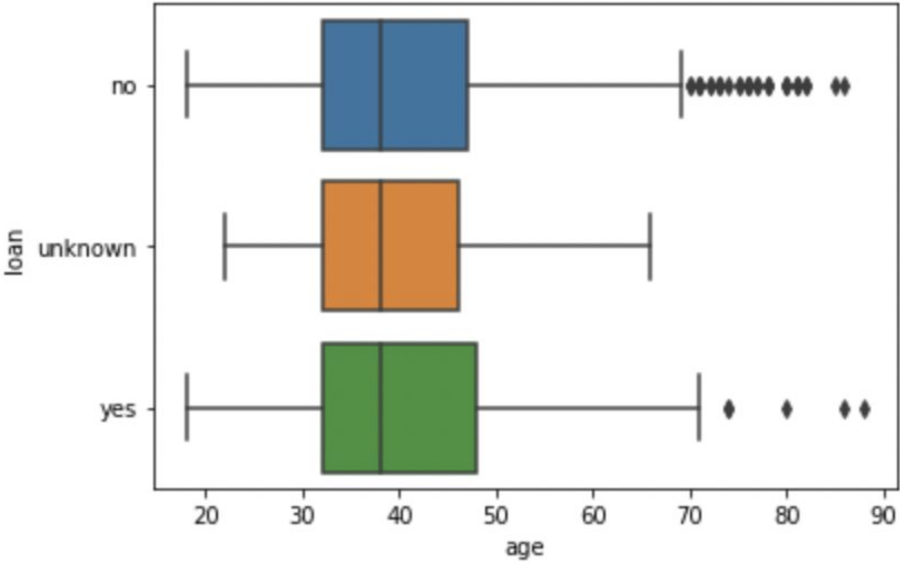
Campaign Outreach by Month:

- Best performing month: May
- Campaign performed the best during the summer months (May-Aug)
- Campaign performed the worst during winter months (Oct-Jan)
- Focus on campaign success early on as it quickly drops in effectiveness



Target Group Identification - Age

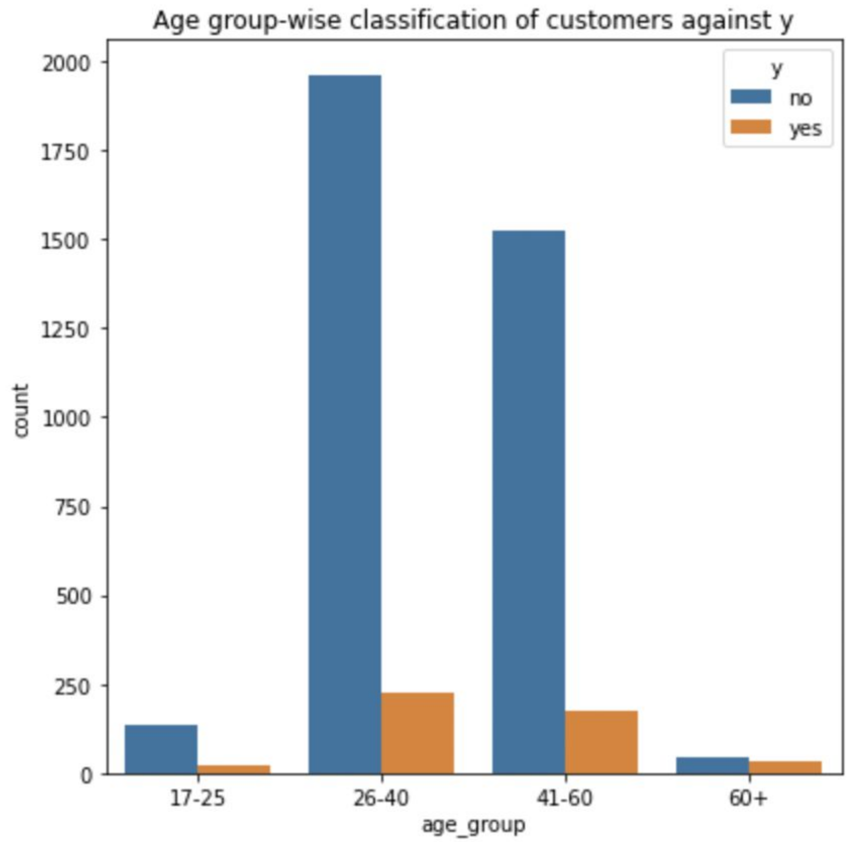




Target Group Identification - Age

Age:

- Most popular age groups:
  - 26-40 y/o
  - 41-60 y/o
- No significant trends between age group and loans
- Highest number of “yes” from the two most popular age groups
  - This may be caused by larger sample size
- We cannot recommend age as a target group on its own

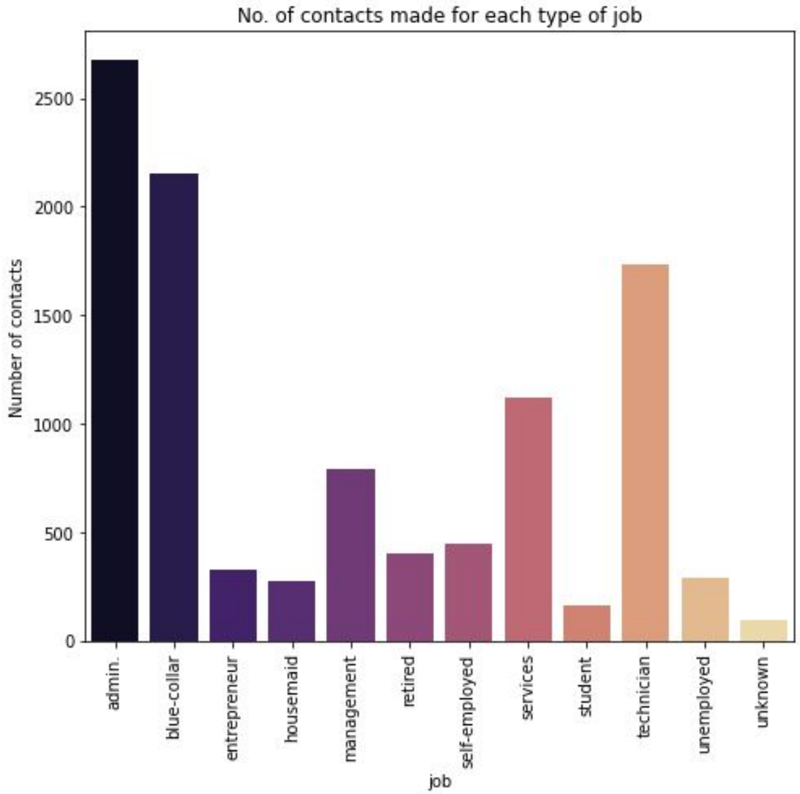


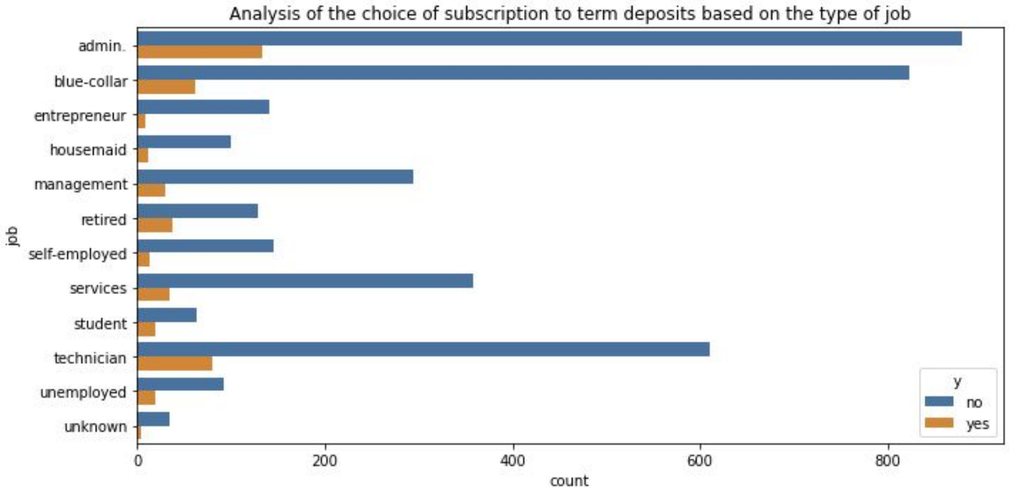
Target Group Identification - Economic Perspectives

Correlation between attributes:

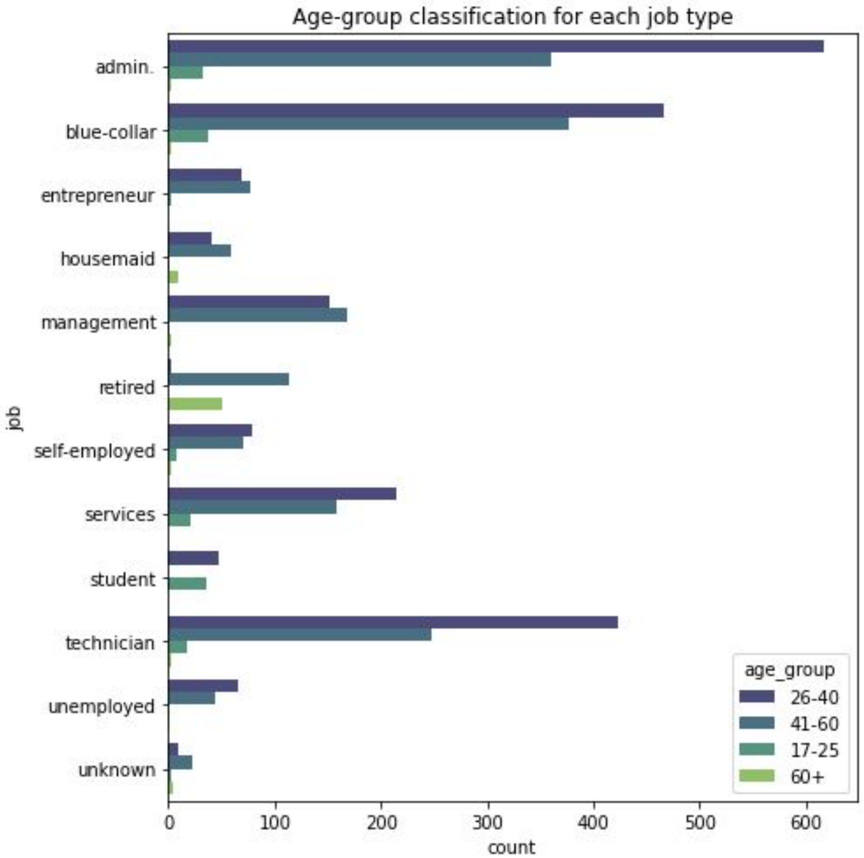
- Employment rate, consumer confidence index, and consumer price index all had high correlations
- These factors may give more insight about target client groups
- We may find that clients who have higher confidence and price index are more likely to purchase the product

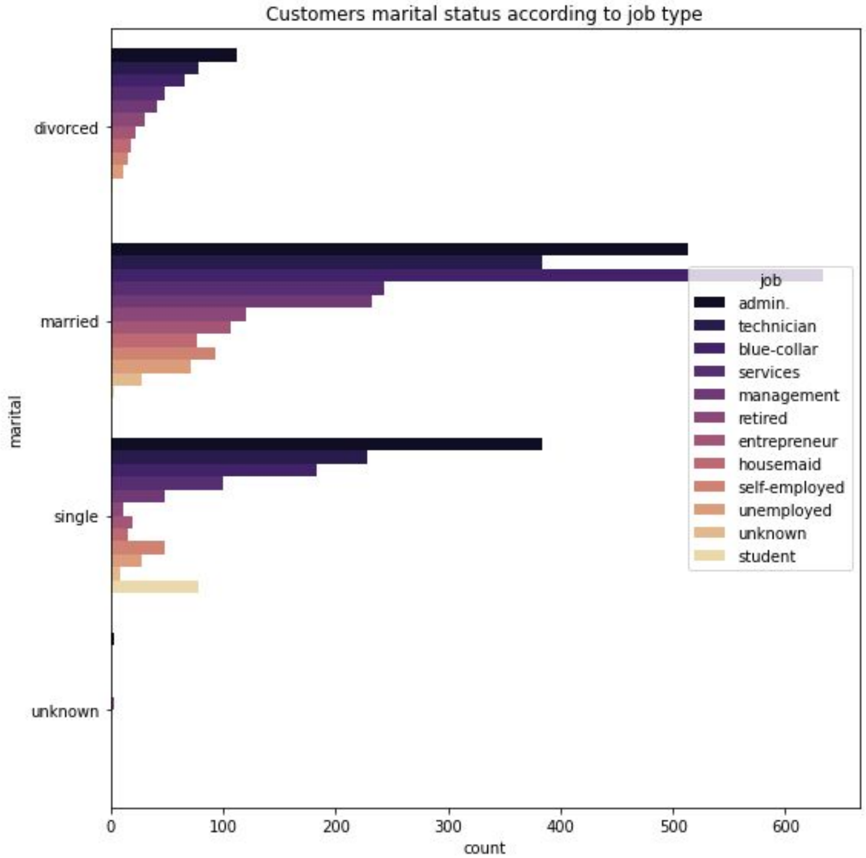
Target Group Identification - Employment and Occupation



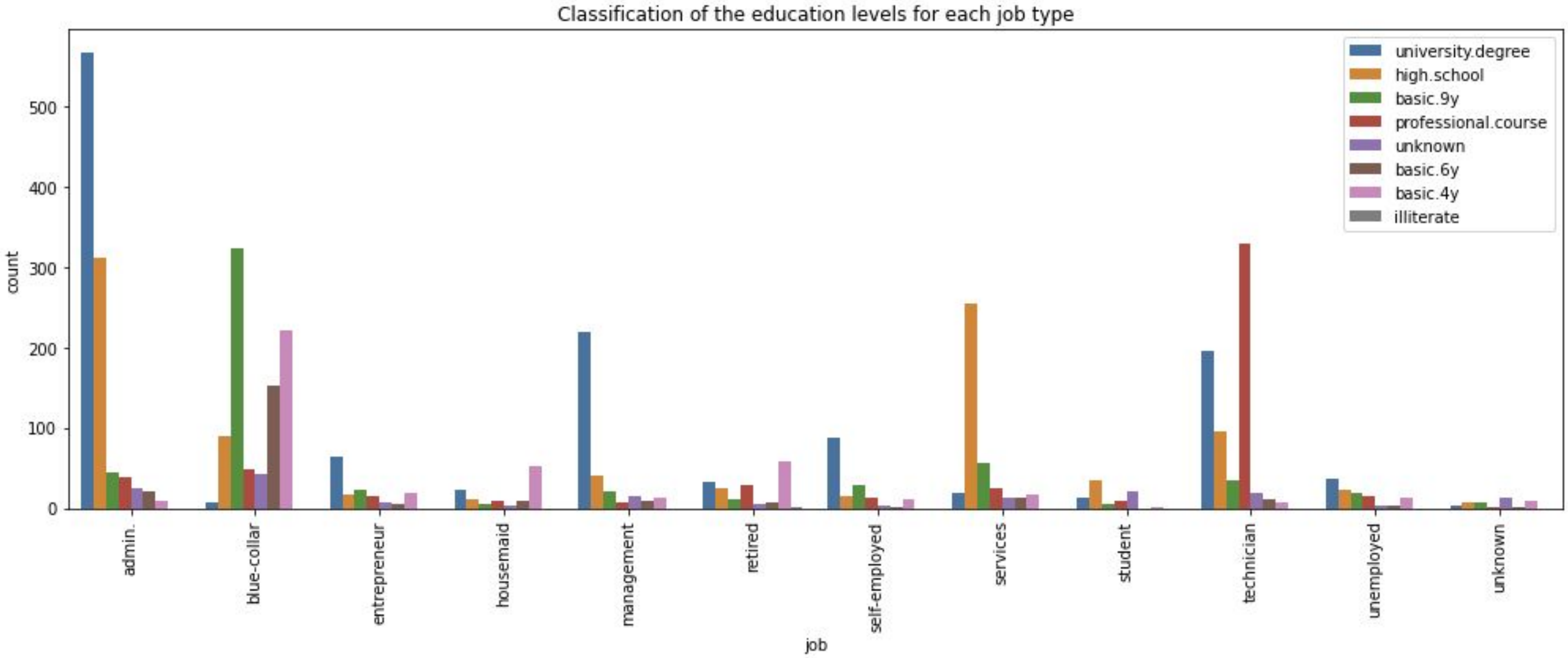


Target Group Identification - Occupation and Other Factors





Target Group Identification - Education and Occupation



Target Group Identification - Final Thoughts

Final Thoughts and Recommendations:

- After exploring many factors and groups, the bank should choose highly efficiency target groups and dates for their ad campaign
  - Suggested date: January - April
  - Suggested groups:
    - Occupation: admin, blue-collar, student, technician
    - Age: 26-40, 17-25
    - Marital Status: married, single (top occupations only)
    - Education: University degree or professional course
- Many useful target groups, but occupation has the largest impact on predicting the purchase rate

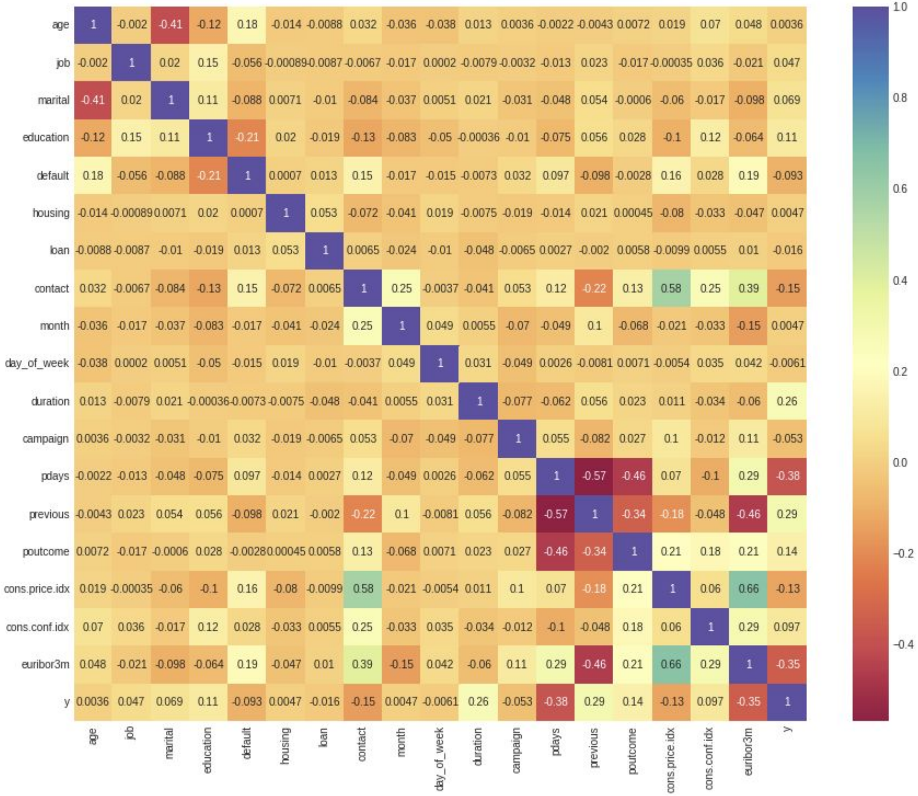
Model Selection and Execution

Thoughts and Recommendations for ML Model Selection:

- Model should predict whether a client will purchase the new product based on a variety of different data inputs
- We will test 6 different algorithms and choose the best
  - Linear algorithms: logistic regression, linear discriminant analysis
  - Nonlinear algorithms: classification and regression trees, support vector machines, Gaussian Naive Bayes, K-nearest neighbors
- Initial results are shown, but a further analysis of model building will be covered in the final report

```
ScaledLR: 0.860654 (0.034861)
ScaledLDA: 0.857459 (0.038983)
ScaledKNN: 0.715261 (0.037793)
ScaledCART: 0.649699 (0.045427)
ScaledNB: 0.826131 (0.038275)
ScaledSVM: 0.823826 (0.040493)
```

Feature Variable Correlation



Feature Variable Correlation:

- On the side we have provided a heatmap of all the correlations of the respective input variables
- As seen, there are no strong correlations between any two features
    - We could have potentially set a threshold value of 0.8 or -0.8

Model Building

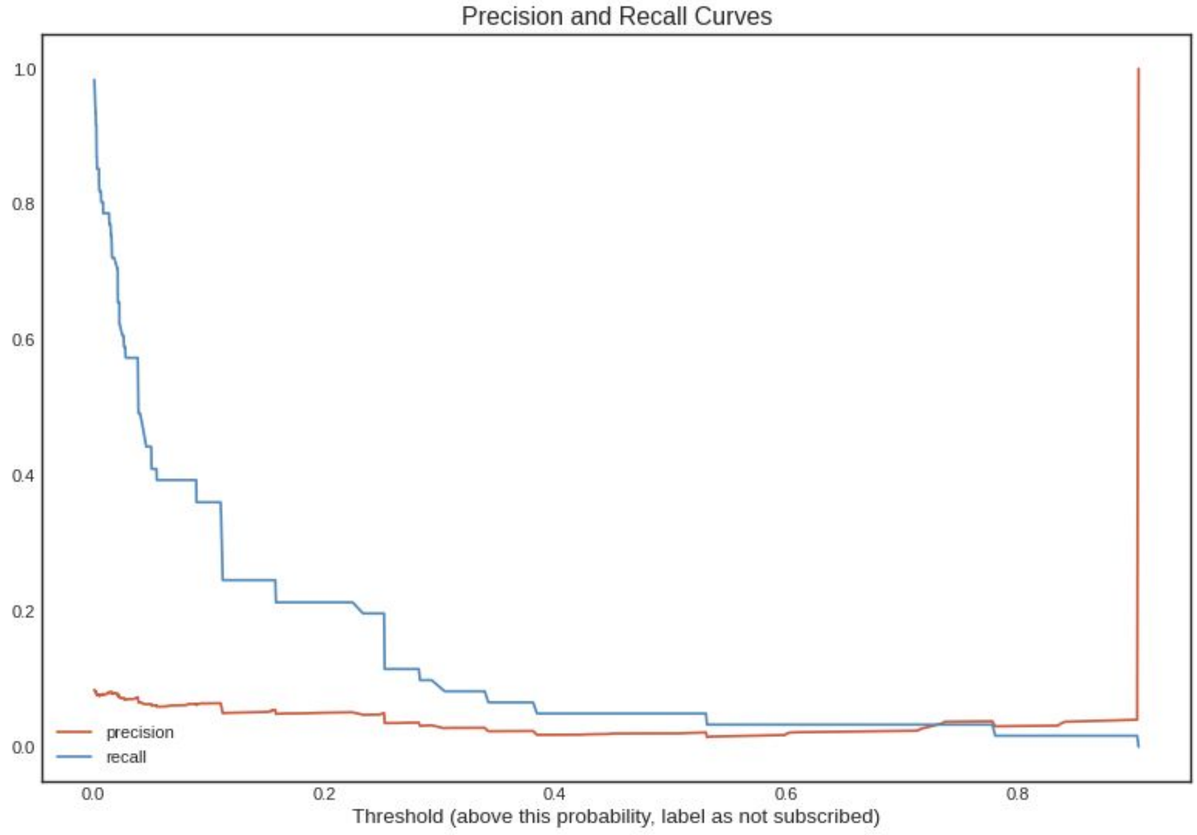
Building the Final Model:

- Before building out the final model, the training dataset was standardized
  - Inputs were scaled to generate predictions
- We chose to use a gradient boosting model for the final model selection
- Using the hold-out dataset the model scored an accuracy of 90%
- The next slide shows a graph of the precision and recall curves

```
[[642  12]
 [ 32  29]]
```

	precision	recall	f1-score	support
0	0.95	0.98	0.97	654
1	0.71	0.48	0.57	61
accuracy			0.94	715
macro avg	0.83	0.73	0.77	715
weighted avg	0.93	0.94	0.93	715

Precision and Recall Curves



Final Model

Building the Final Model:

- The model was fitted using logistic regression
- Parameter tuning was utilized to determine the models overall accuracy
  - The mean accuracy was 93%
- The classification report shows a precision value of 93%
  - No false positives were labeled
- In conclusion, we believe the bank would be able to confidently use our model to predict client purchase outcomes!

```
param_grid = {'C': np.logspace(-4, 4, 50),
              'penalty':['l1', 'l2']}
clf = GridSearchCV(LogisticRegression(random_state=0), param_grid
best_model = clf.fit(X_train,y_train)
print(best_model.best_estimator_)
print("The mean accuracy of the model is:",best_model.score(X_tes
```

```
LogisticRegression(C=51.79474679231202, random_state=0)
The mean accuracy of the model is: 0.9314685314685315
```

```
Confusion Matrix:
 [[652   2]
 [ 47  14]]
Classification Report:
```

	precision	recall	f1-score	support
0	0.93	1.00	0.96	654
1	0.88	0.23	0.36	61
accuracy			0.93	715
macro avg	0.90	0.61	0.66	715
weighted avg	0.93	0.93	0.91	715

Thank you!