PREDICTING BANK PRODUCT

BY NWAAMAKA IDUWE

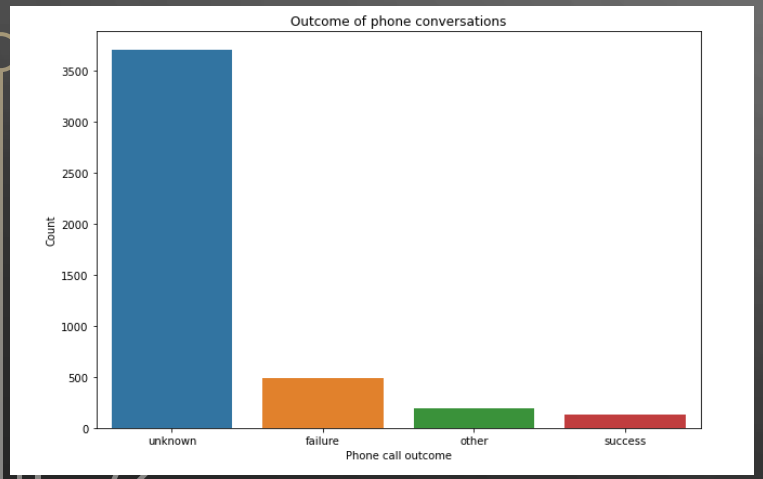
PROJECT DEFINITION

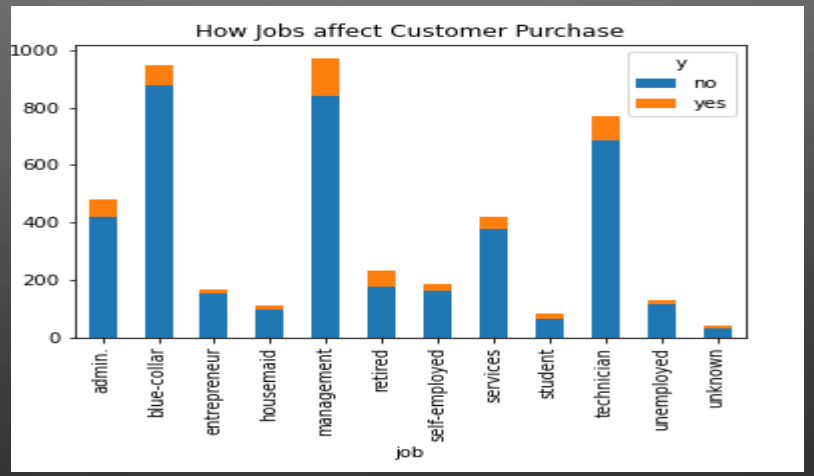
- The aim of this project is to use different machine learning algorithms to predict which customers will be the most likely to subscribe to the bank's new product –bank term deposit. Towards this aim, the objectives were:

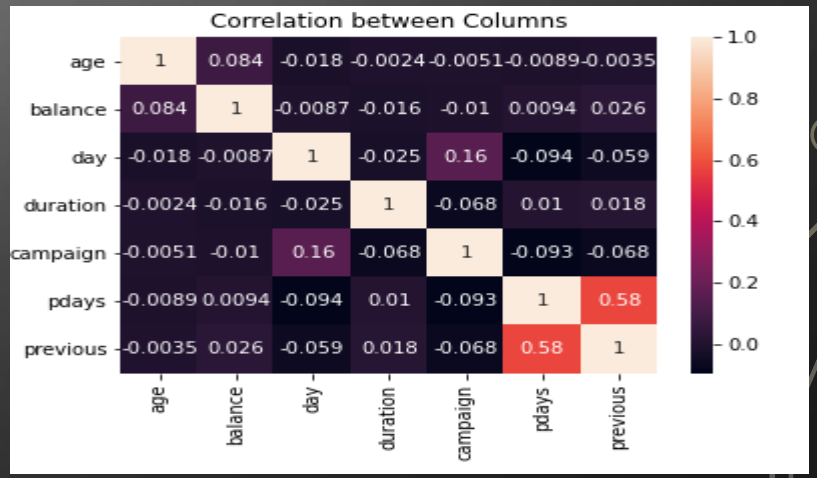
1. Exploratory Data Analysis
2. Feature Engineering
3. One-Hot Encoding
4. Model Selection
5. Model Training
6. Feature Importance
7. Model Accuracy, Precision and Recall
8. Cross-Validation

EXPLORATORY DATA ANALYSIS

- Before modelling, it is important to explore and visualize the raw data to ensure that I am familiar with its contents so that I can derive as much insights as possible from it.
- For this project, the first thing I did was to look at the data by columns so I can understand the kind of data I am working with in terms of data types, data size, data shape, etc. Following this, I conducted some univariate, bivariate and multivariate analysis to see what the relationships between columns are and how useful this might be to make sense of the important features that would come later.







PREPARING, SELECTING TRAINING THE MODEL AND FETCHING IMPORTANCES

PREPARING:

- To prepare the data for modelling, I feature engineered the column y so that its contents are integers and not strings. This is important as most models only work with numerical values. Added to this, I also one-hot encoded other columns containing strings by putting the corresponding columns in a list and then using the dummy feature on pandas to convert the contents in those columns to integers (0 or 1).

SELECTING & TRAINING:

- Here, we create the code that will train and test four models from which only one model will be chosen based on the score. We use a list and a loop for the list to test each model in the list with the train-test code. In this code, we set y to be our target variable which is coincidentally named ‘y’ in the data and we set x to be all other columns as they are the independent variables which y is dependent on. Our test size is set to 40%.
- With the highest accuracy score, the RandomForestClassifier (RF) was chosen as the desired model.

FETCHING IMPORTANCES:

- Using the selected model, we a code that will allow python tell us which of the columns in our data are the top 10 that we should be focusing on to ensure our target is met. According to the output, the top 10 are:
- 'job_entrepreneur', 'job_admin.', 'job_blue-collar', 'previous’, 'campaign', 'pdays', 'day', 'age', 'balance', 'duration’.

MODEL EVALUATION

ACCURACY, PRECISION AND RECALL:

- To ensure the model predictions are accurate and reliable, I conducted accuracy, precision and recall tests using tools from the sklearn metrics library. Accuracy tells how correct the model is, precision checks how precise the model's categorical prediction is and recall checks how well the model can detect a category.

As mentioned in the previous slide, the best performing model of the four is the RandomForestClassifier and the scores show that works very well, is stable and can generalize to new data. It has maintained an accuracy score between 89% and 91% which is a strong score. In addition to accuracy, more evaluations were conducted by testing precision and recall from which it was deduced that of the 1809 tested data, 96% were negative meaning that 96% of the people did not subscribe to the product while the 20% positive values are those who subscribed to the product. For recall, the model has an accuracy of 90% for those who would not subscribe to the product and a 41% accuracy for those who would subscribe to the product.

CROSS VALIDATION USING K-FOLD:

- In addition to these, I conducted a final test called cross validation using a tool called K-fold tool from the Sklearn metrics library. The cross-validation test tells me how well my model can generalize to new data by testing multiple trainings and tests. In this case, I used 10 splits and with a score of 88% the results showed that indeed the model can generalize to new data.

	precision	recall	f1-score	support
0	0.96	0.90	0.93	1709
1	0.20	0.41	0.27	100
accuracy			0.88	1809
macro avg	0.58	0.66	0.60	1809
weighted avg	0.92	0.88	0.89	1809

THANK YOU