MOVEMBER FUNDRAISER'S PROJECT REPORT

DEFINITION

Project Overview

Movember is buildings a future where men live happier, healthier, longer lives. On average, men die 6 years earlier than women for largely preventable reasons. Rates of prostate cancer and testicular cancer are rising. Movember exists to stop men from dying too young. Movember is the leading charity tackling mental health and suicide prevention, prostate cancer, and testicular cancer on a global scale. In this project, I did the Exploration Data Analysis(EDA), Data Transformation, Features Engineering, and Model building to make predictions.

Problem Statement

- Identify the top behaviors and attributes that are likely to predict whether a Movember fundraiser will be retained in 2021 using Exploration Data Analysis.

You should aim to find or reveal all relevent properties, characteristics, patterns, statistics hidden in the dataset.

- Develop a binary classification Model to Identify which Movember fundraiser's in 2020 will be retained for 2021.

You'll need to build a model with whatever machine learning approaches you feel appropriate. You should evaluate your model on a range of metrics ; However, we also suggest that you use the F1 score to evaluate the performance of your final model. You should follow an industry recognized approach to Data Science problems (e.g CRISP-DM) and include a justificationfor your selected model.

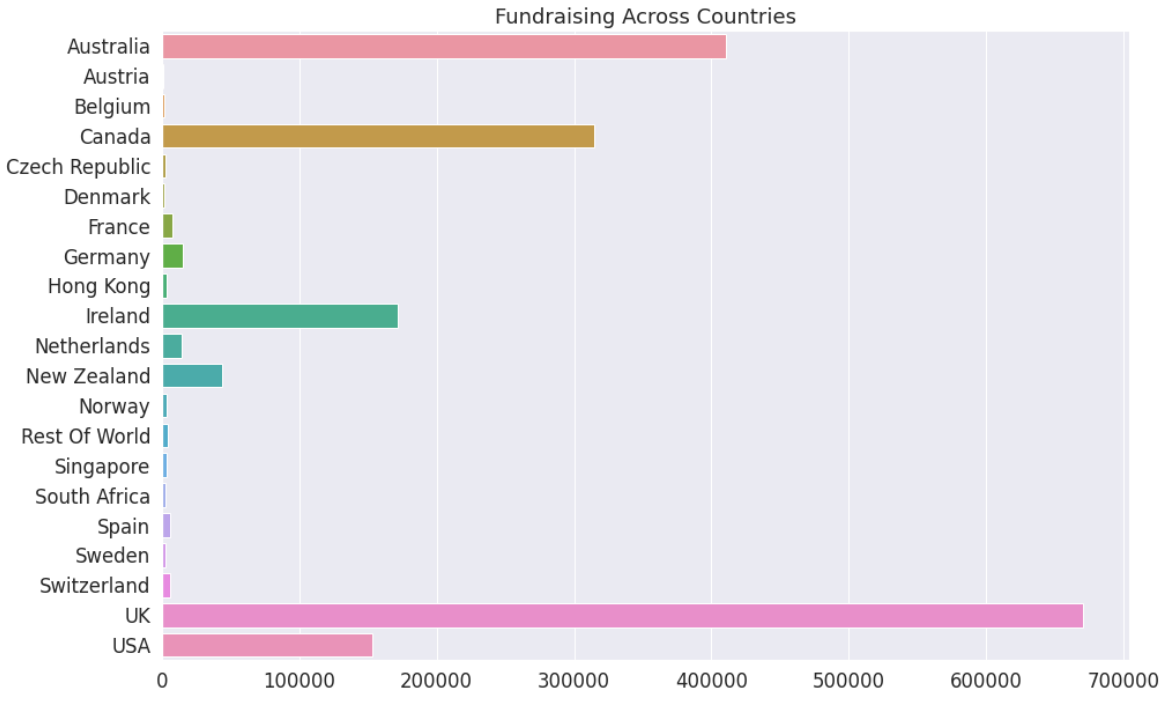
- Outline a Strategyon prioritising which 2020 fundraiser's Movmber should invest resources to retain based on their likelyhood to churnand their value to Movember(based on either 2020 fundraising value or predicted 2021 fundraising value.

DATA WRANGLING

The most important aspect of an machine learning projects is to first understand the nature of the data with the use of statistical methods in Numpy or Pandas. Such as loading the data, checking the columns summary, the info's, columns data types, missing values and many more as i did in this project.

EXPLORATION DATA ANALYSIS (EDA)

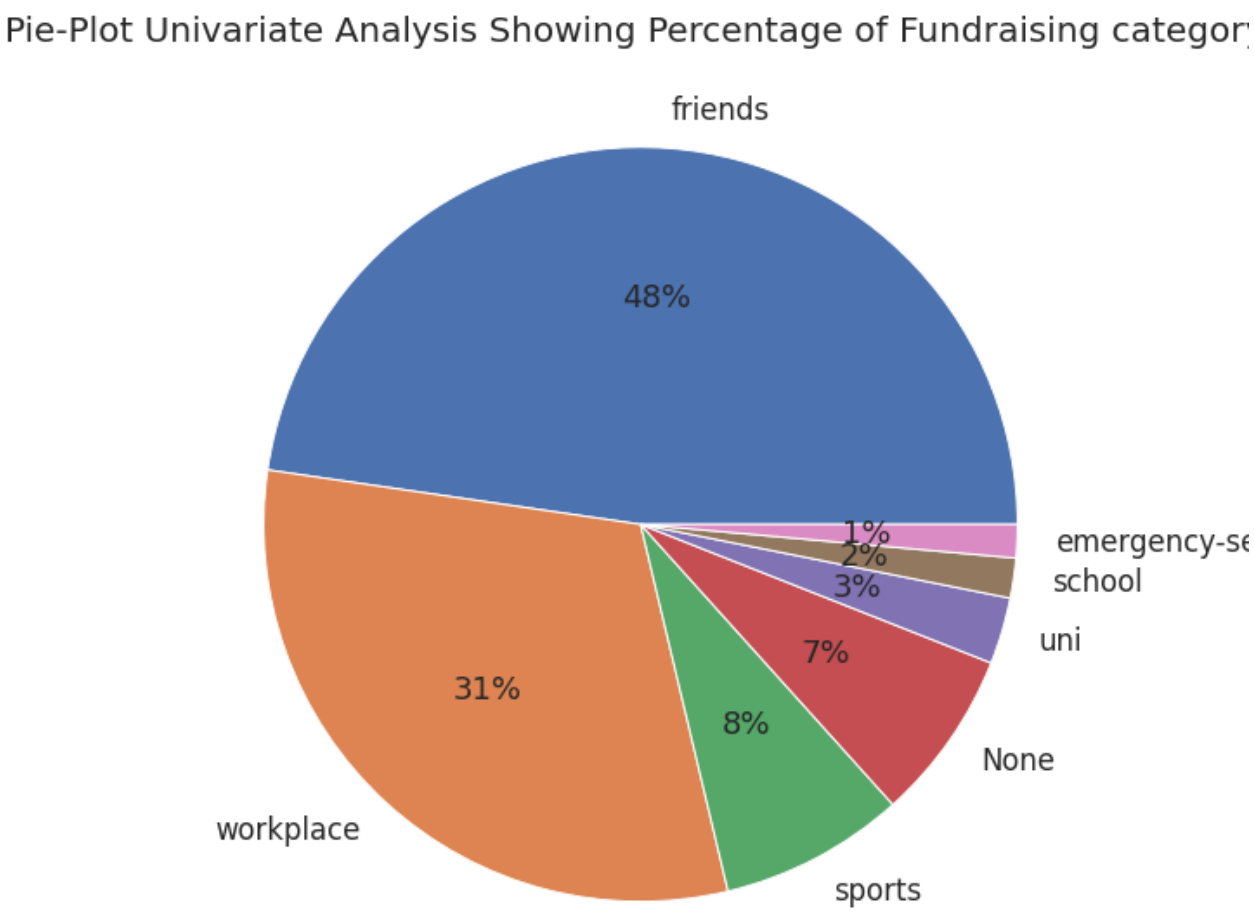
Exploration data analysis is the process of representing the reports graphically. In this project, I did a few visualizations with deep meanings. Those EDA's are as follows:



The above visualization indicates the total fundraising across each

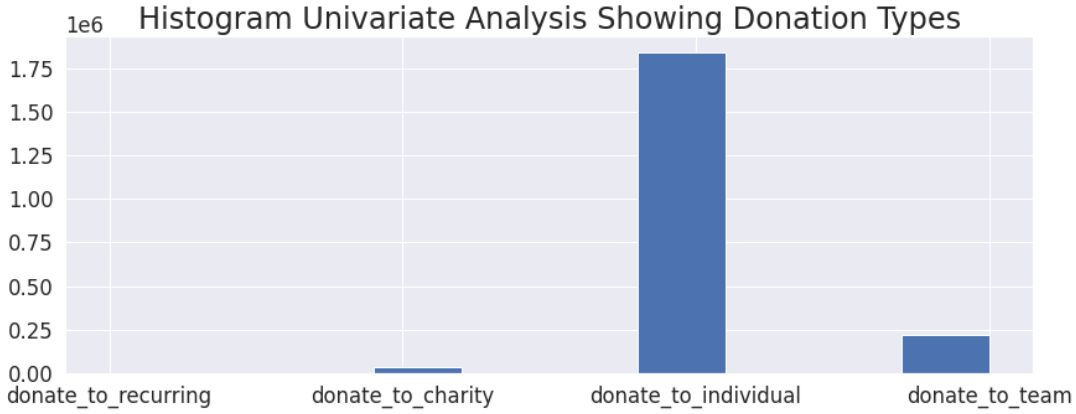
country. The UK tends to have the highest rate of fundraiser's, followed by Austria.

2.



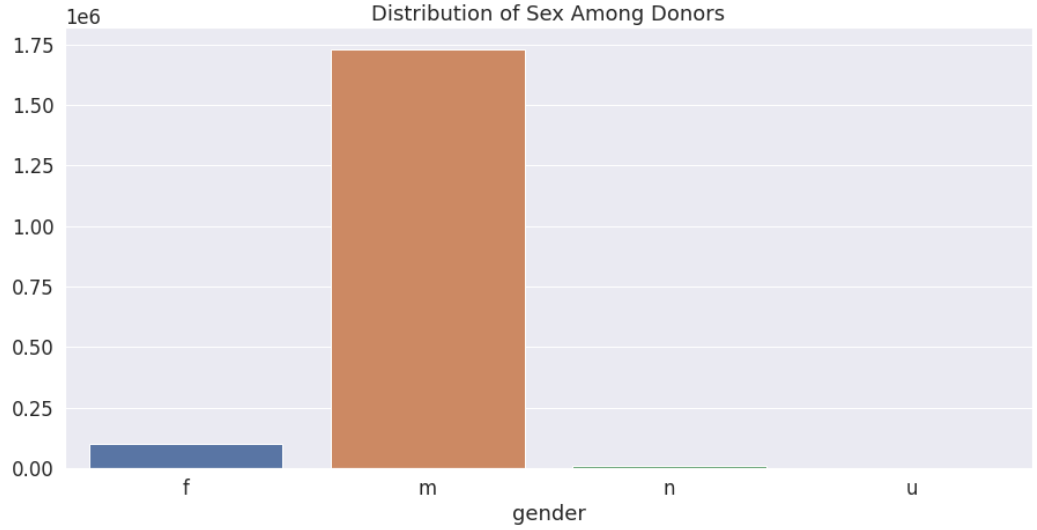
The above beautiful pie-plot gives information about fundraising in this category. Taking a deep look at the graph, people from the Friends & Workplace categories have more than 75% of the fundraisings.

3.



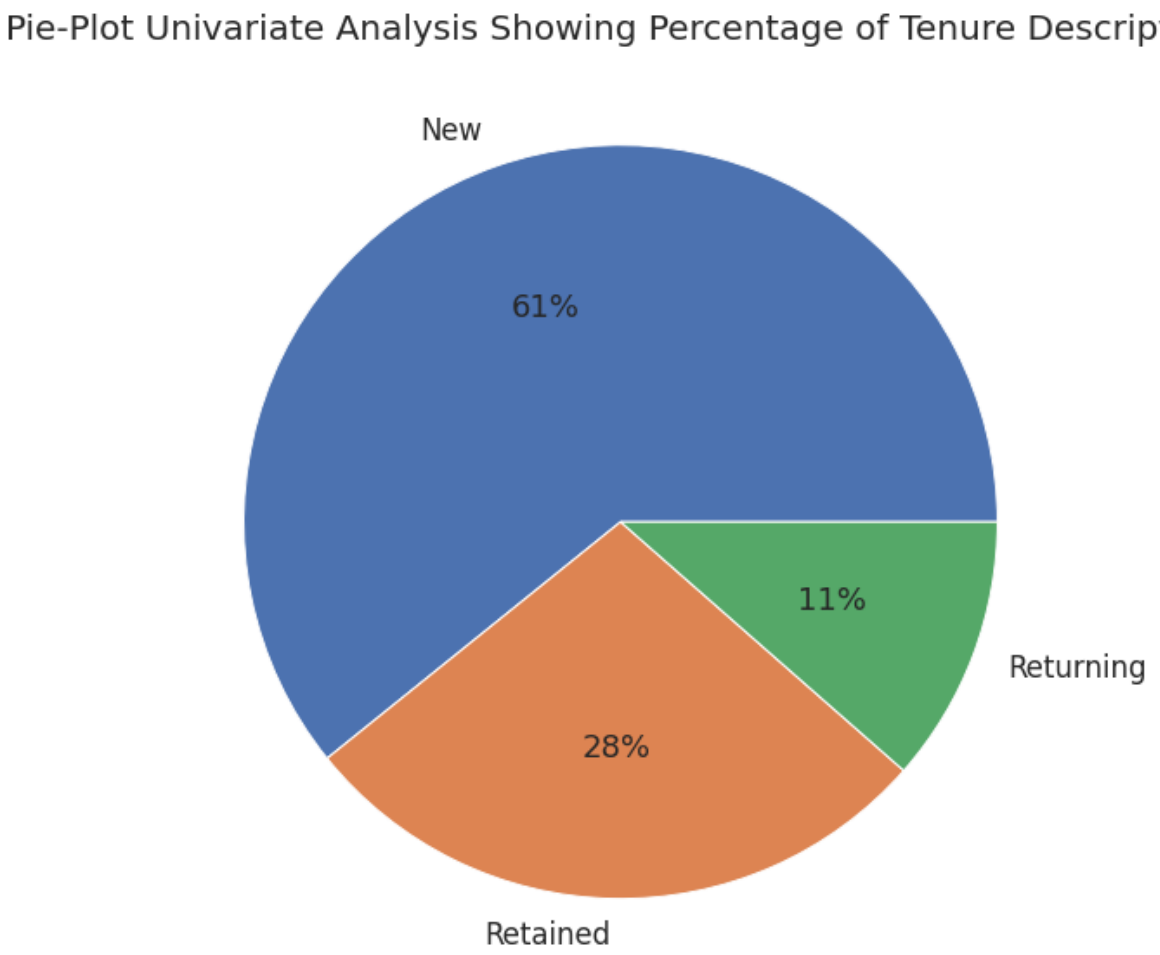
The Above histogram indicates that the fundraiser tends to donate more to the Individuals than to Teams.

4.



The graph above indicates the distribution of Donors among sex. Over 70% of the donors are males.

5.



The Above pie-plot indicates that over 60% ofFirst time fundraiser's donate the most.

DATA PREPROCESSING

Working with insignificant values in the dataset is also one of the most important steps to take while working on

machine learning projects. In these project, I performed a lot of preprocessing steps which I'll definitely explain how I did each of them.

- **Dropping Duplicate values: D**ropping duplicate values in the dataset tends to boost the performance of the model.

- **Missing Values : I**t's a great skill to be able to deal with any king of missing values in the dataframe. There are several ways to deal with missing values which includes filling the missing values with either the mean, median or mode of the column which is based on the data type of the column.

  In this project,there were missing values in the object columns so we filled them with the mode of each column by looping through them. And on the otherhand, we rename/replace mispelt columns or values in the dataframe.

- **Dealing with insignificant values: I** dealt with insignificant data in this project by dropping them as it adds nothing to the model performance, rather it causes noise.

- **Encoding: E**ncoding is also one of the most interesting part of a ML projects, which is often performed on categorical features in a dataframe. There are several ways to perform encoding, but in this project, I decided to use the PANDAS OneHotEncoding as I find this method suitable.

- **DATA SEGMENTATION: W**ith the help of the data dictionary, I was able to figure out which columns are the predictor and the column to be predicted as the dictionary comprises all columns and their meanings. With that, I also figured out that the

"**registered_2021**" column is the column to be predicted, so I renamed the column name to **"Churn".** Afterwards, I segmented the data frame into two parts **"X,y"** for further preprocessing.

- **DATA SPLITTING: A**fter the segmentation process, I used the method **"train_test_split"** provided by the scikit-learn library to split the data into training and Validation sets for checking model accuracy and performance.

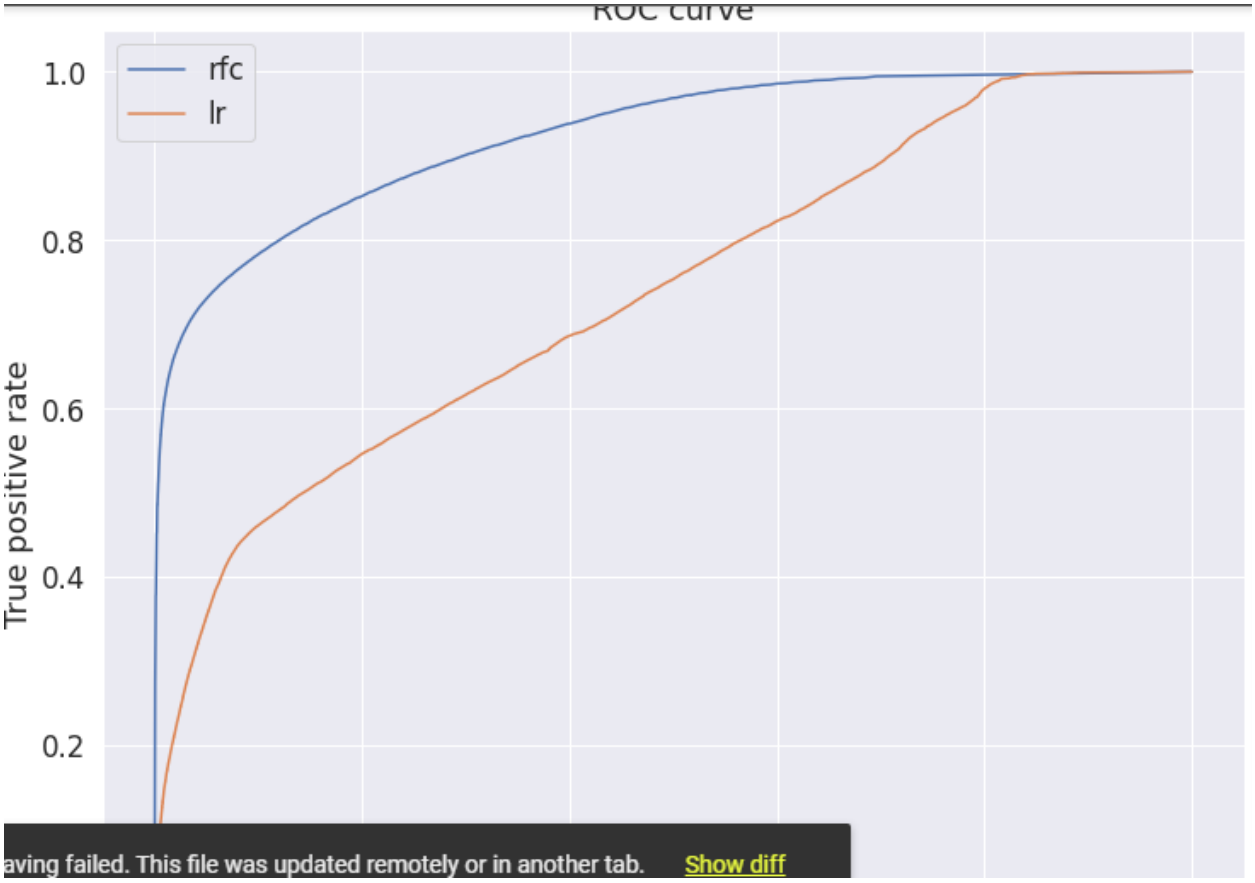
  Code :

```
X_train, X_test, y_train, y_test = train_test_split(X,y,
test_size=0.25, random_state=42, shuffle=True
```

  We used 75% for training our model and 25% for model validation.

MODEL BUILDING

This is the most interesting part i've been waiting for. In this project i used two different classification algorithms (Logistic Regression and Random Forest Classifier). Both models are good but it's compulsory to validate each model's performance. I performed alot of metrics and also plotted a **ROC_CURVE** to check which model performs better**.**



With the help of the above graph, I am able to convince myself that the RandomForest Model is the best for this project. Not only that, the F1-scores of each models are:

Random Forest : 80%

Logistic Regression : 52%

Outlining Strategies

With the help of the EDA's and preprocessing done in this project, I'll have to outline 5 bullet points which will help Movember to make

decisions on which fundraisers they should invest into and retain based on their likelihood to churn.

1. **The United Kingdom has the highest rate on fundraising, followed by Austria. Movember should invest into those countries as the churn percentage is over 60%.**

2. **Movember should also retain the friends & Workspace as they donates over 70% in the fundraising category.**

3. **Movember should also retain males in the gender category as they tend to donate over 70% in the category.**