

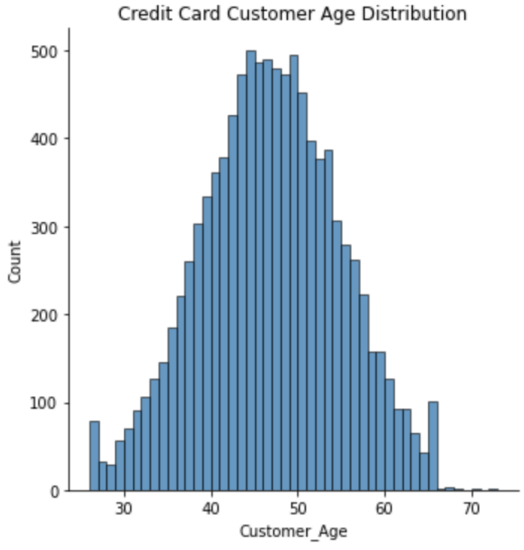
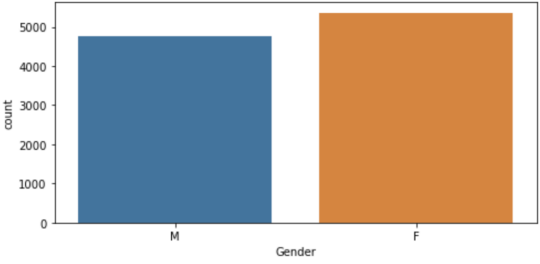
Data science capstone project

Credit card institutions use customer churning to predict who is going to stop using their credit card services. This helps to identify customers who are likely to churn and in turn allows them to, ideally, create solutions specifically to retain these customers.

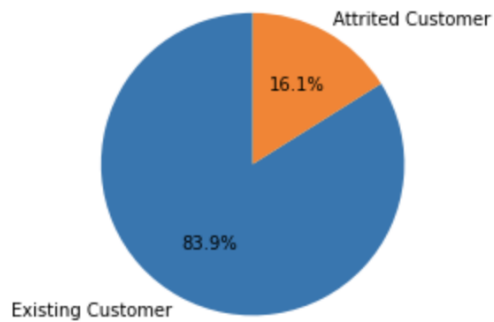
The dataset was taken from Kaggle [here](#), and consists of 10,000 customers mentioning their age, salary, marital_status, credit card limit, credit card category, etc. There are nearly 18 features.

Problem statement: To predict customer churn from the dataset and gain some insights on how the bank can reduce the customers who have churned.

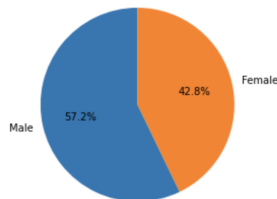
Findings from EDA:

	AGE <ul style="list-style-type: none">• Most customers were aged between 40 and 60 years old. The mean age of the dataset, 46, also falls within this range.
	GENDER <ul style="list-style-type: none">• There were not much difference between the total no. of male and female customers

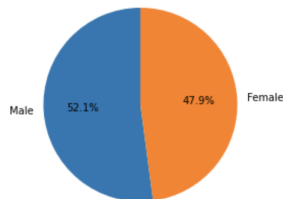
Proportion of Existing and Attrited Customer count



Attrited Customer vs Gender

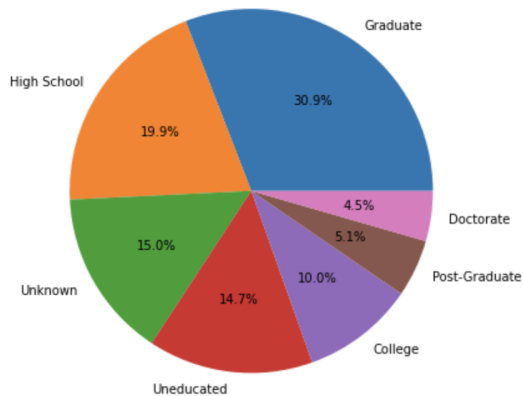


Existing Customer vs Gender

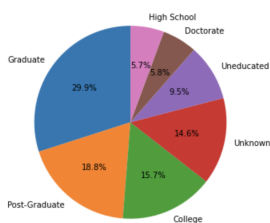


- There are a total of **16.1%** attrited customers.
- There were more existing customers than attrited customers.
- Out of these, there were more male customers who were attrited and existing compared to females.

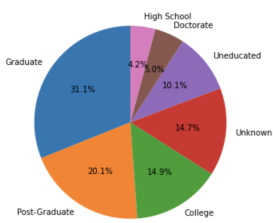
Proportion of Education Levels



Attrited Customer vs Education Level

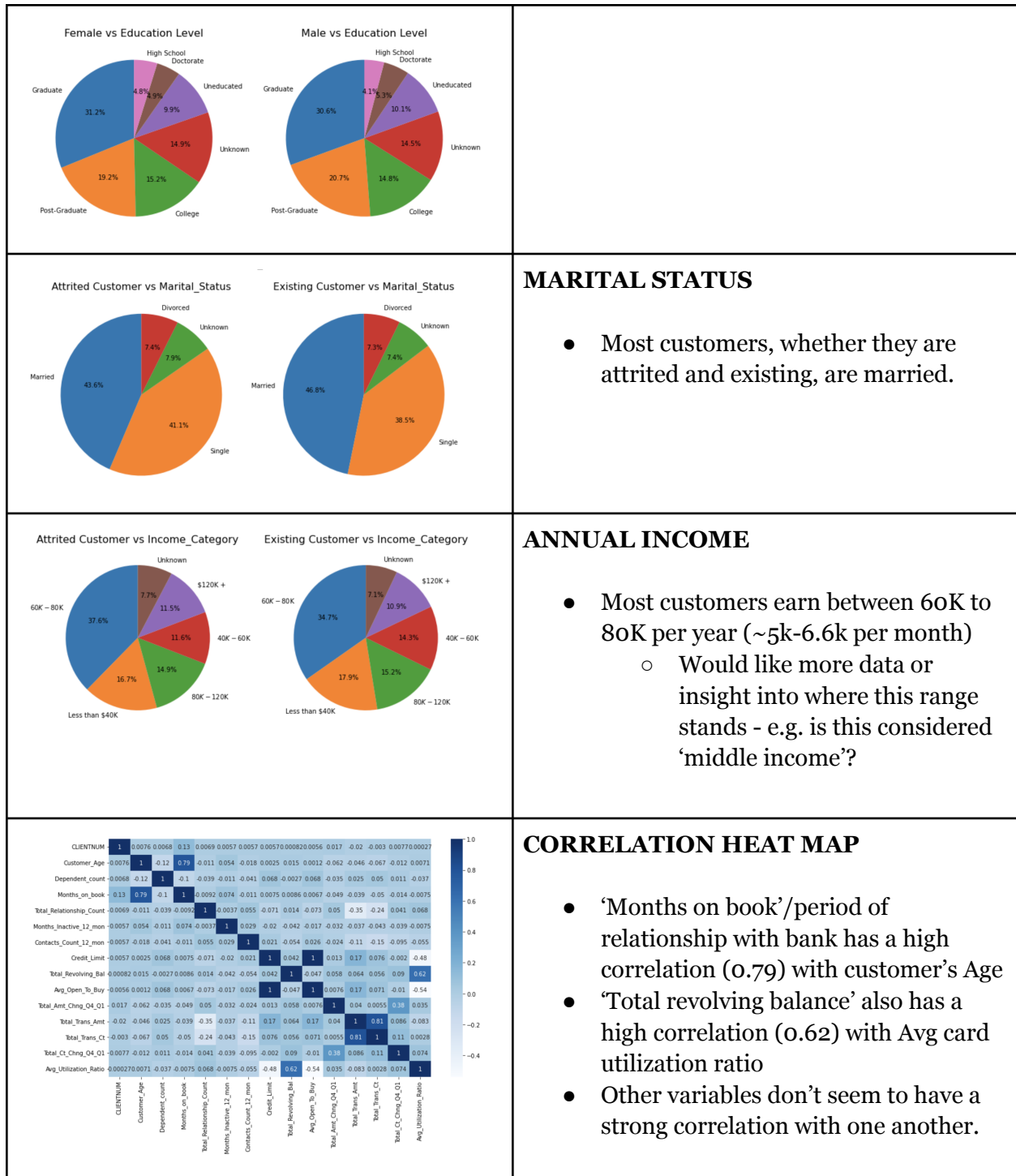


Existing Customer vs Education Level



EDUCATION LEVELS

- Most customers in the database were Graduates.
- **40.5%** of them are graduate levels and above (graduates, post-graduate and doctorate)
- On splitting them according to whether they are attrited or existing customers, the proportion breakdown according to education levels seems to be **comparable**.
- Graduate and Post-graduates are the two largest categories and make up more than 50% of customers for both Attrited and Existing customers
- When split by Gender, the trend/breakdown is largely the same.
- Most female and male customers are concentrated in the 'Graduate' level



Attrited Customer vs Income_Category

Income_Category	Percentage
60K - 80K	37.6%
Less than \$40K	16.7%
80K - 120K	14.9%
40K - 60K	11.6%
\$120K +	11.5%
Unknown	7.7%

Existing Customer vs Income_Category

Income_Category	Percentage
60K - 80K	34.7%
Less than \$40K	17.9%
80K - 120K	15.2%
40K - 60K	14.3%
\$120K +	10.9%
Unknown	7.1%

Attrited Customer vs Income_Category

Income_Category	Percentage
60K - 80K	37.6%
Less than \$40K	16.7%
80K - 120K	14.9%
40K - 60K	11.6%
\$120K +	11.5%
Unknown	7.7%

Existing Customer vs Income_Category

Income_Category	Percentage
60K - 80K	34.7%
Less than \$40K	17.9%
80K - 120K	15.2%
40K - 60K	14.3%
\$120K +	10.9%
Unknown	7.1%

	CLIENTNUM	Customer_Age	Dependent_count	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
CLIENTNUM	1	0.0076	0.0068	0.13	0.0069	0.0057	0.0057	0.0057	0.00082	0.0056	0.017	-0.02	-0.003	0.0077	0.00027
Customer_Age	-0.0076	1	-0.12	0.79	-0.011	0.054	-0.018	0.0025	0.015	0.0012	-0.062	-0.046	-0.067	-0.012	0.0071
Dependent_count	-0.0068	-0.12	1	-0.1	-0.039	-0.011	-0.041	0.068	-0.0027	0.068	-0.035	0.025	0.05	0.011	-0.037
Months_on_book	0.13	0.79	-0.1	1	0.0092	0.074	0.011	0.0075	0.0086	0.0067	0.049	-0.039	-0.05	-0.014	-0.0075
Total_Relationship_Count	-0.0069	-0.011	-0.039	-0.0092	1	-0.0037	0.055	-0.071	0.014	-0.073	0.05	-0.35	-0.24	-0.041	0.068
Months_Inactive_12_mon	-0.0057	0.054	-0.011	0.074	-0.0037	1	0.029	-0.02	-0.042	-0.017	-0.032	-0.037	-0.043	-0.039	-0.0075
Contacts_Count_12_mon	-0.0057	-0.018	-0.041	-0.011	0.055	0.029	1	0.021	-0.054	0.026	-0.024	-0.11	-0.15	-0.095	-0.055
Credit_Limit	-0.0057	0.0025	0.068	0.0075	-0.071	-0.02	0.021	1	0.042	1	0.013	0.17	0.076	-0.002	-0.48
Total_Revolving_Bal	-0.00082	0.015	-0.0027	0.0086	0.014	-0.042	-0.054	0.042	1	-0.047	0.058	0.064	0.056	0.09	0.62
Avg_Open_To_Buy	-0.0056	0.0012	0.068	0.0067	-0.073	-0.017	0.026	-0.047	1	0.0076	0.17	0.071	-0.01	-0.54	
Total_Amt_Chng_Q4_Q1	-0.017	-0.062	-0.035	-0.049	0.05	-0.032	-0.024	0.013	0.058	0.0076	1	0.04	0.0055	-0.38	0.035
Total_Trans_Amt	-0.02	-0.046	0.025	-0.039	-0.35	-0.037	-0.11	0.17	0.064	0.17	0.04	1	0.081	0.086	-0.083
Total_Trans_Ct	-0.003	-0.067	0.05	-0.05	-0.24	-0.043	-0.15	0.076	0.056	0.071	0.0055	0.81	1	0.11	0.0028
Total_Ct_Chng_Q4_Q1	-0.0077	-0.012	0.011	-0.014	0.041	-0.039	-0.095	-0.002	0.09	-0.01	-0.38	0.086	0.11	1	0.074
Avg_Utilization_Ratio	-0.00027	0.0071	-0.037	0.0075	0.068	-0.0075	0.055	-0.48	0.62	-0.54	0.035	-0.083	0.0028	0.074	1

MARITAL STATUS

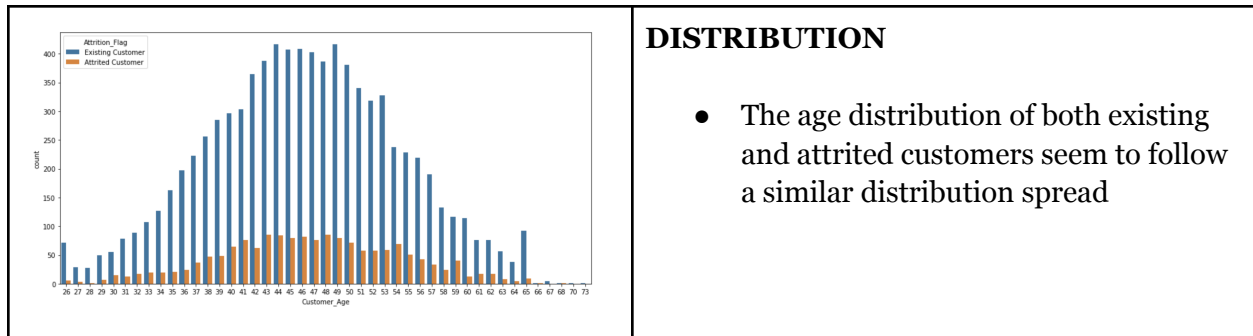
- Most customers, whether they are attrited and existing, are married.

ANNUAL INCOME

- Most customers earn between 60K to 80K per year (~5k-6.6k per month)
 - Would like more data or insight into where this range stands - e.g. is this considered ‘middle income’?

CORRELATION HEAT MAP

- ‘Months on book’/period of relationship with bank has a high correlation (0.79) with customer’s Age
- ‘Total revolving balance’ also has a high correlation (0.62) with Avg card utilization ratio
- Other variables don’t seem to have a strong correlation with one another.

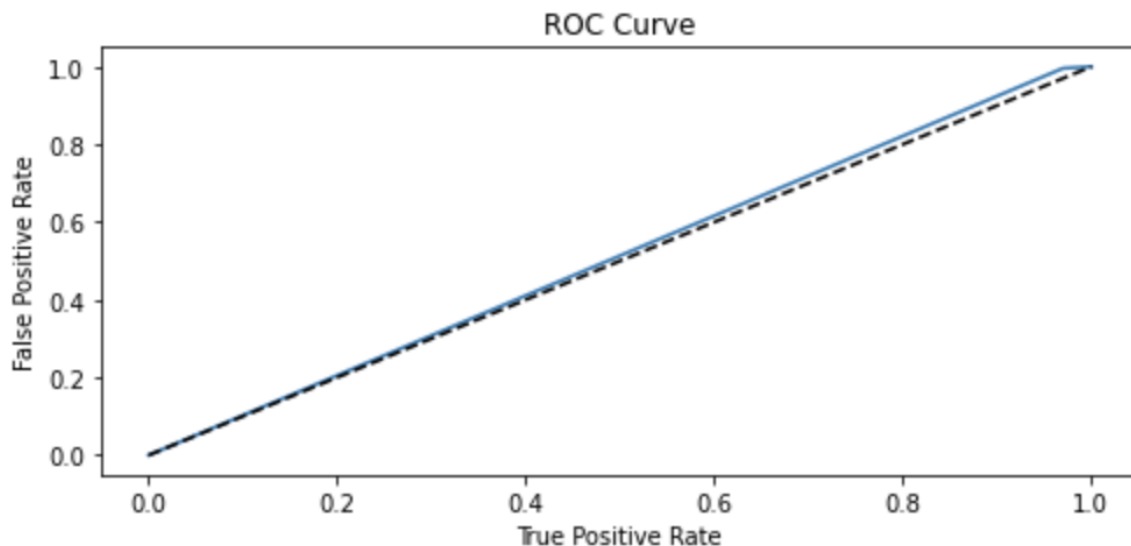


About the model: Used RandomForestClassifier and Linear Regression

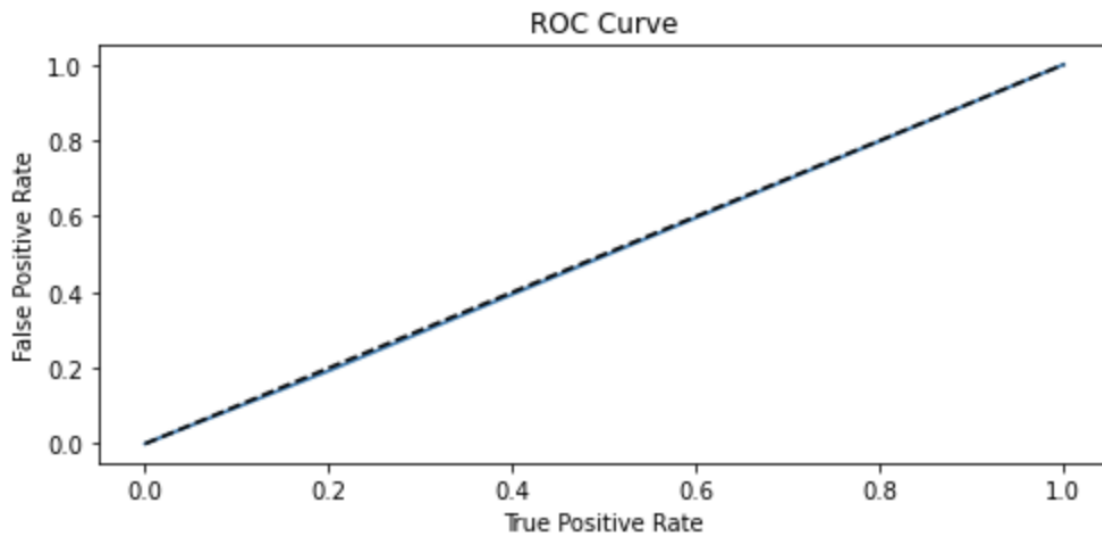
- RandomForestClassifier was chosen as the model for it's versatility. It can be used for both regression and classification tasks. The default hyperparameters it uses also often produce a good prediction result.
- The downsides to it is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. However in this case, the dataset wasn't too complex for the model.

First, since we require numerical values for the predictive model, the categorical columns need to be transformed. Hence **label encoding** is done.

In our initial setup of the model, the model produced results with **0.966 accuracy** and **0.907 AUC score**.



We also tried using the Linear Regression model which produced results with lower accuracy and AUC score at **0.905** and **0.732** respectively.



Results/Conclusion:

- In total, there are **16.07%** of customers who have churned.
- There were some interesting insights from the EDA.
- We found that the **proportion of gender count is almost equally distributed (52.9% male and 47.1%)** compared to the **proportion of existing and attributed customer count (83.9% and 16.1%) which is highly imbalanced.**
 - There is also a higher proportion of attrited customers who are male (57.2%) compared to female (42.8%).
 - Customers who have churned are highly educated - A high proportion of education level of attrited customer is Graduate level (29.9%), followed by Post-Graduate level (18.8%)
 - A high proportion of marital status of customers who have churned is Married (43.6%), followed by Single (41.1%) compared to Divorced (7.4%) and Unknown (7.9%) status - marital status of the attributed customers are highly clustered in Married status and Single
- However, while we gained more insight on customers who have churned, the findings are not very meaningful as we still don't seem to know **the reason why** they have churned.
 - When we split the data up between attributed and existing customers, and analysed each of these groups based on the different variables like gender, income, marital status, the trends were similar in these groups. So it's hard to find out the reason why a customer would have churned.