Week 3 – Perform Exploratory Data Analysis

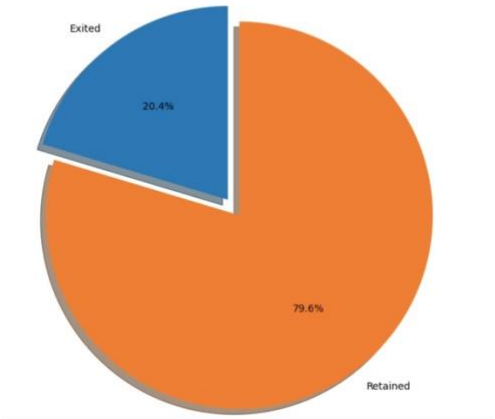
For this dataset, there was no missing values, but there were outliers. For the age attribute, there were outliers for the Exit 0 group. This meant elder people were more willing to stay at the same bank. I would keep theses outliers to see what happens. Also, I did a pie chart to find what percentage of customers churned among all customers. From the chart, I found that about 20% of the customers have churned. It is important for companies to find out what caused their clients churn.

In addition to that, I did box-plots to better understand the dataset and had some more interesting findings. Firstly, Germany were more likely to churn compared to France and Spain. Secondly, people who didn't have credit card were more likely to churn compared to the customers who had credit cards. Thirdly, there were customers who had exited but still had a balance in their account. My guess was they have exited from a product and not the bank. And lastly, by looking at the distribution of the data, I noticed that the minimum estimated salary was only 11.58, which was not make sense. The reason could be this customer entered the salary by per hour, not by per month or per year.

Before splitting the dataset and building models, I converted 2 categorical features (geography and gender) into numerical variables. I used one hot encoding to convert geography variable into numerical variable because there was no sequence or order in geography. Then, I used a different strategy, which was labelencoder to convert age gender attribute to numerical variable.

I used stratified sampling and the test size was 25%. So, the dataset had 75% of the training data and 25% of the test data. I split the dataset because I need training data to train the model and then applied models to do prediction on the test data. I chose these percentage because my dataset was relatively small and I wanted the training data to be a big portion of the whole dataset. I used stratified sampling to avoid extreme case. The data was basically ready for training and model implementation at this point.

Proportion of customer churned and retained



[9]:

	CreditScore	Age	Tenure	NumOfProducts	Balance	EstimatedSalary
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000
mean	650.528800	38.921800	5.012800	1.530200	76485.889288	100090.239881
std	96.653299	10.487806	2.892174	0.581654	62397.405202	57510.492818
min	350.000000	18.000000	0.000000	1.000000	0.000000	11.580000
25%	584.000000	32.000000	3.000000	1.000000	0.000000	51002.110000
50%	652.000000	37.000000	5.000000	1.000000	97198.540000	100193.915000
75%	718.000000	44.000000	7.000000	2.000000	127644.240000	149388.247500
max	850.000000	92.000000	10.000000	4.000000	250898.090000	199992.480000