**CREDIT CARD CHURNERS**

With around 19 features per user, we need to predict which user can drop off.

**Business Problem**: - A business manager of a consumer credit card portfolio is facing problem of customer attrition. They want to analyze the data to find out the reason behind this and leverage the same to predict customers who are likely to drop off.

**Data Dictionary: -**

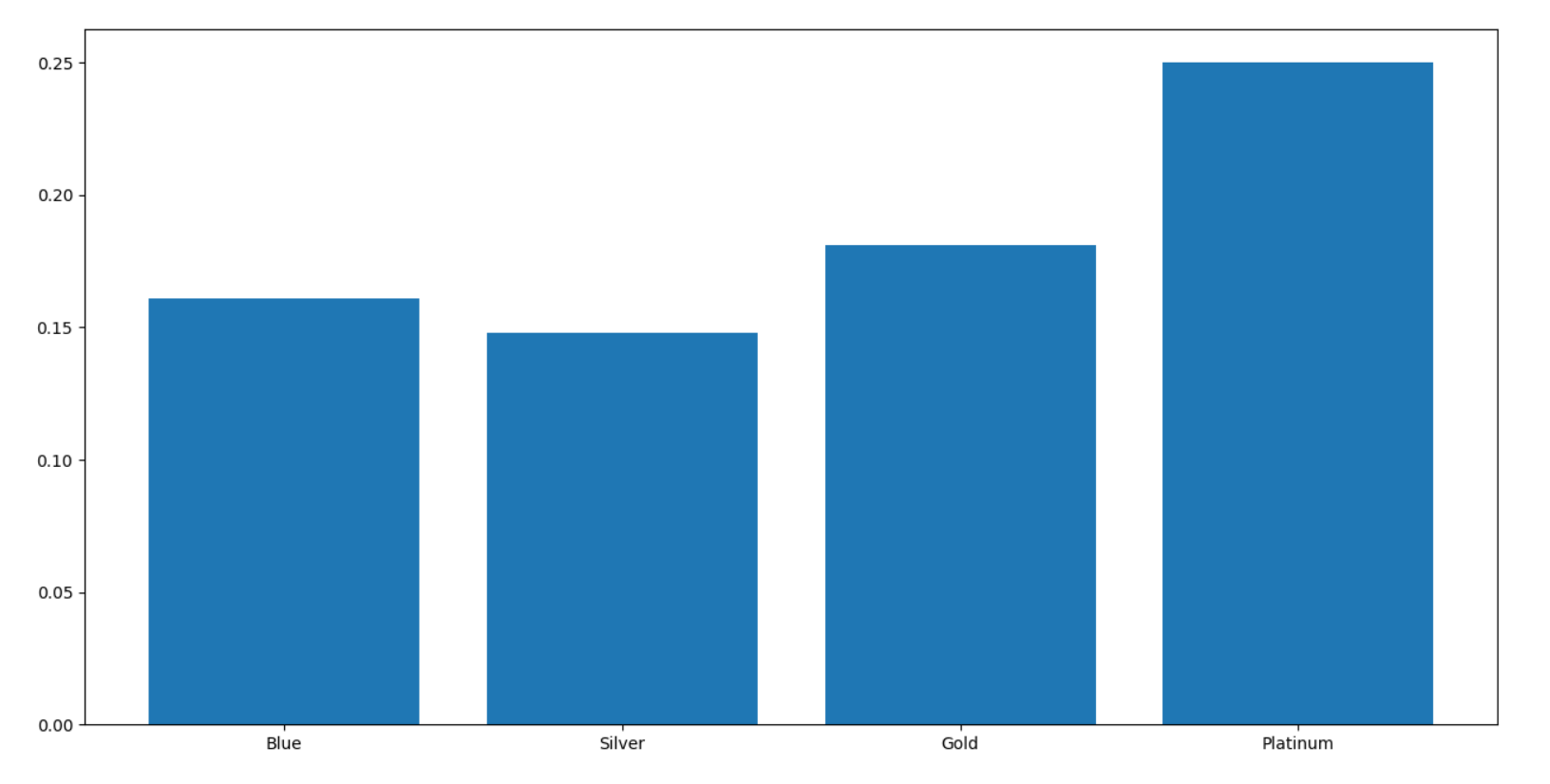
|  |  |  |
| --- | --- | --- |
| Variable | Type | Description |
| Clientnum | Num | Client number. Unique identifier for the customer holding the account |
| Attrition\_Flag | char | Internal event (customer activity) variable - if the account is closed then 1 else 0 |
| Customer\_Age | Num | Demographic variable - Customer's Age in Years |
| Gender | Char | Demographic variable - M=Male, F=Female |
| Dependent\_count | Num | Demographic variable - Number of dependents |
| Education\_Level | Char | Demographic variable - Educational Qualification of the account holder (example: high school, college graduate, etc.) |
| Marital\_Status | Char | Demographic variable - Married, Single, Unknown |
| Income\_Category | Char | Demographic variable - Annual Income Category of the account holder (< $40K, $40K - 60K, $60K - $80K, $80K-$120K, > $120K, Unknown) |
| Card\_Category | Char | Product Variable - Type of Card (Blue, Silver, Gold, Platinum) |
| Months\_on\_book | Num | Months on book (Time of Relationship) |
| Total\_Relationship\_Count | Num | Total no. of products held by the customer |
| Months\_Inactive\_12\_mon | Num | No. of months inactive in the last 12 months |
| Contacts\_Count\_12\_mon | Num | No. of Contacts in the last 12 months |
| Credit\_Limit | Num | Credit Limit on the Credit Card |
| Total\_Revolving\_Bal | Num | Total Revolving Balance on the Credit Card |
| Avg\_Open\_To\_Buy | Num | Open to Buy Credit Line (Average of last 12 months) |
| Total\_Amt\_Chng\_Q4\_Q1 | Num | Change in Transaction Amount (Q4 over Q1) |
| Total\_Trans\_Amt | Num | Total Transaction Amount (Last 12 months) |
| Total\_Trans\_Ct | Num | Total Transaction Count (Last 12 months) |
| Total\_Ct\_Chng\_Q4\_Q1 | Num | Change in Transaction Count (Q4 over Q1) |
| Avg\_Utilization\_Ratio | Num | Average Card Utilization Ratio |
|  |  |  |

**Hypothesis Generation:-**

Factors that can influence a drop off:-

1. Users not feeling beneficial from the credit card policies such as limit, rewards etc
2. People with low income or people who are unemployed do not tend to continue with credit cards
3. People who don’t use credit card much shall eventually drop off, since they might not find worthy to pay annual fee when they are not using it much

* First, let’s see if variations in existing & former customers are varying as card category changes. A card category tells us about the limit & rewards a customer gets.



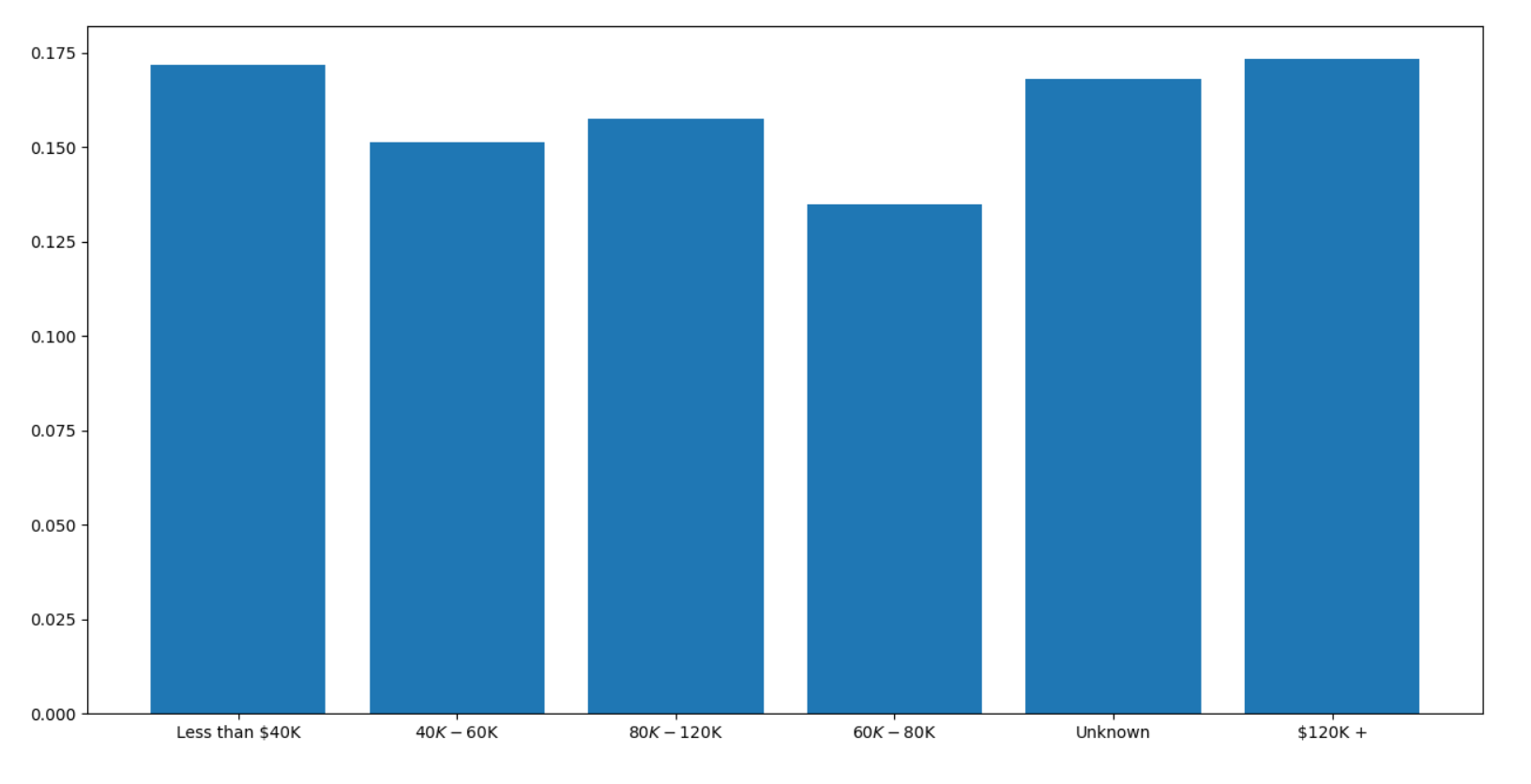
Each bar represents the ratio of number of customers who left to that of total customers under that category

We can see the category of platinum card has least retaining power.

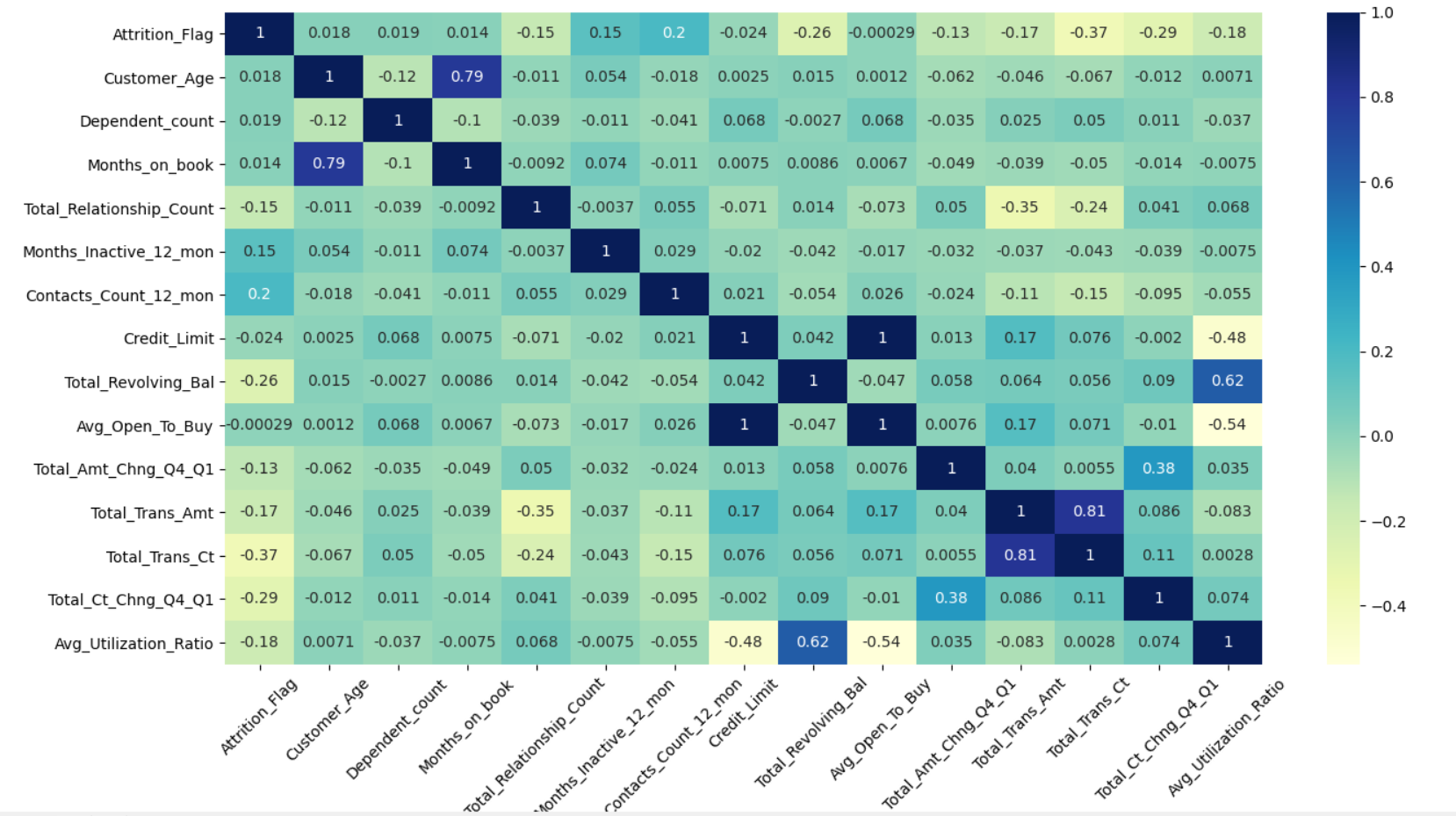
* Moving On, Let’s check if incomes provide a statistically significant variation.

Each bar represents the ratio of number of customers who left to that of total customers under that category

Though there’s not much variation, we can say that people who earn in range of 60-80K USD tend to stay.



Correlation b/w numerical variables.



Credit Limit is very highly correlated to Avg open to buy

Total Trans Ct is highly correlated to Total trans amt (Quite obvious)

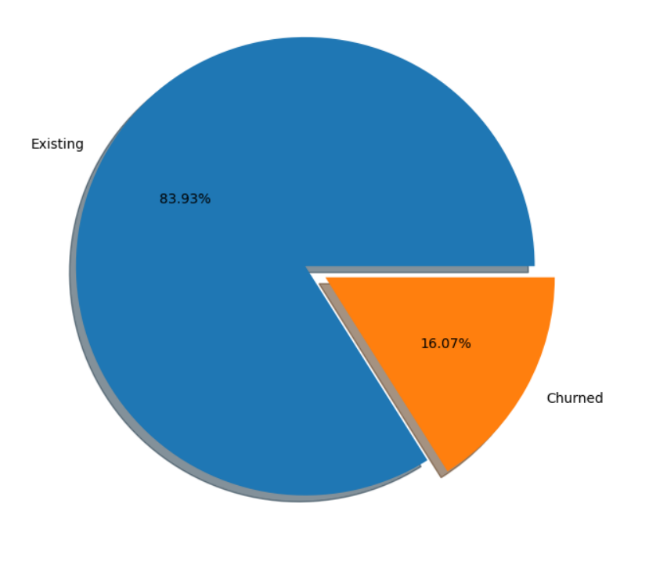
Customer Age is highly correlated to Months on book

We can either remove one of the variables in pair or linearly combine them:-

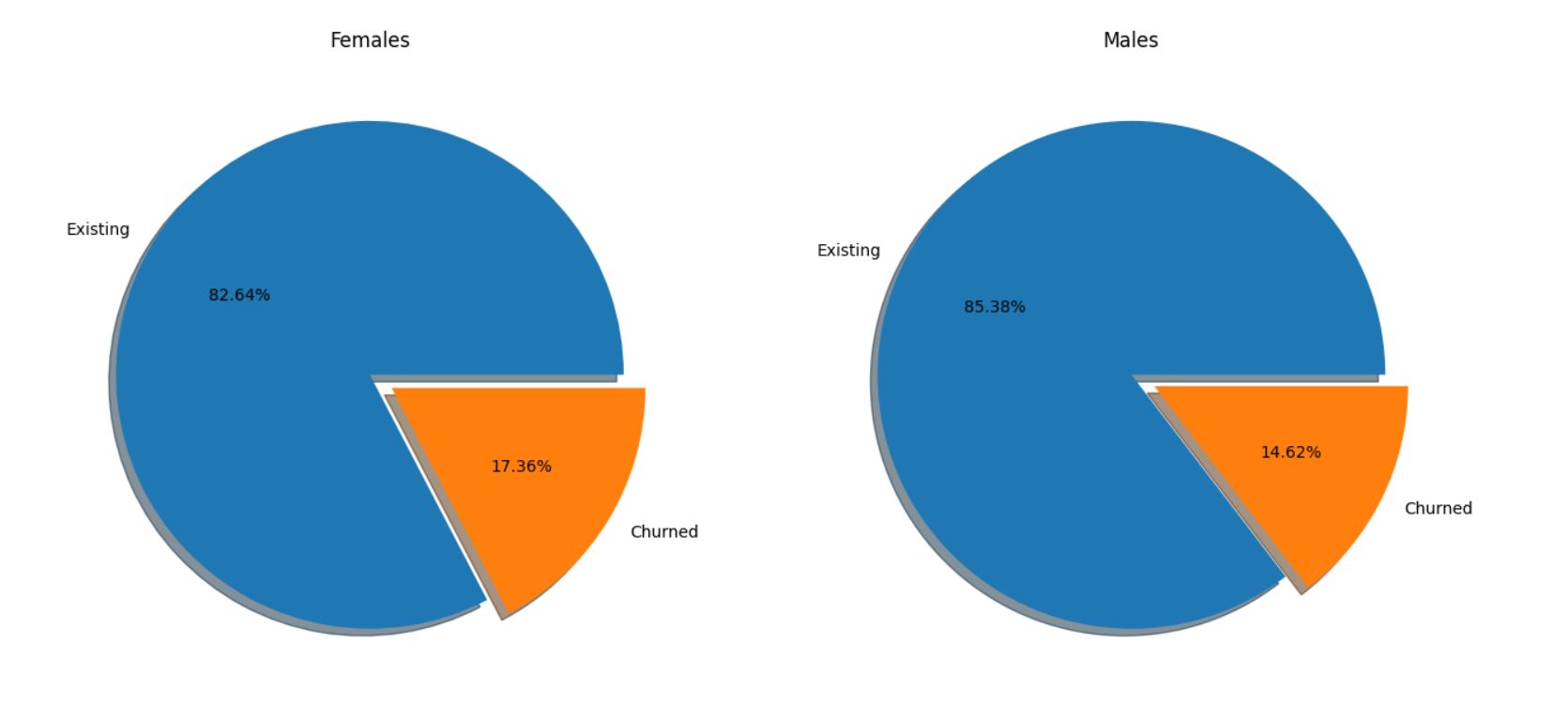
1. We can remove ‘’Avg open to buy” --- this operation didn’t make any difference, so just as well remove it.
2. Make a new variable which is ratio of total trans amt and total trans ct, which will basically be average transaction amount in last year

Above operation downgraded my model, so I’m gonna undo these 2 point.

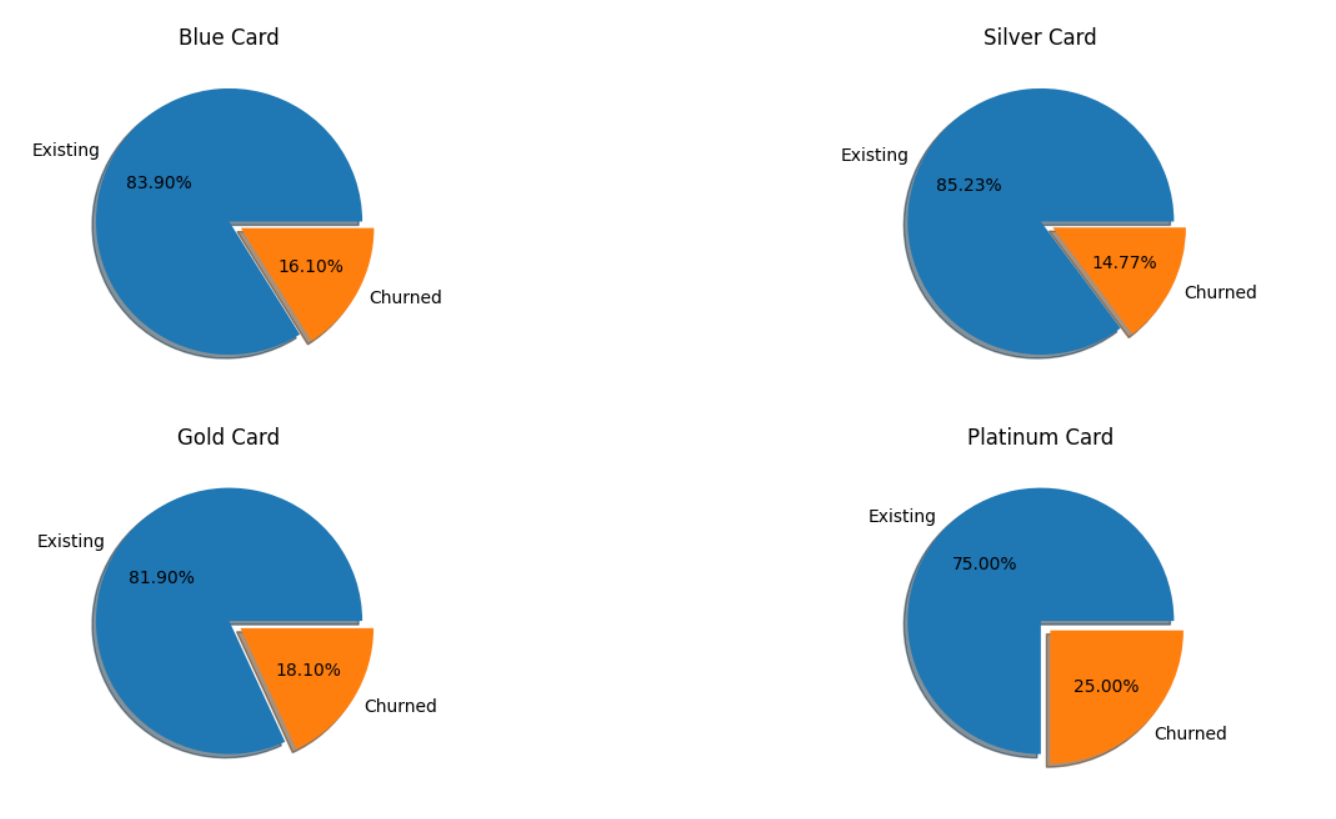
**Exploratory Data Analysis: -**



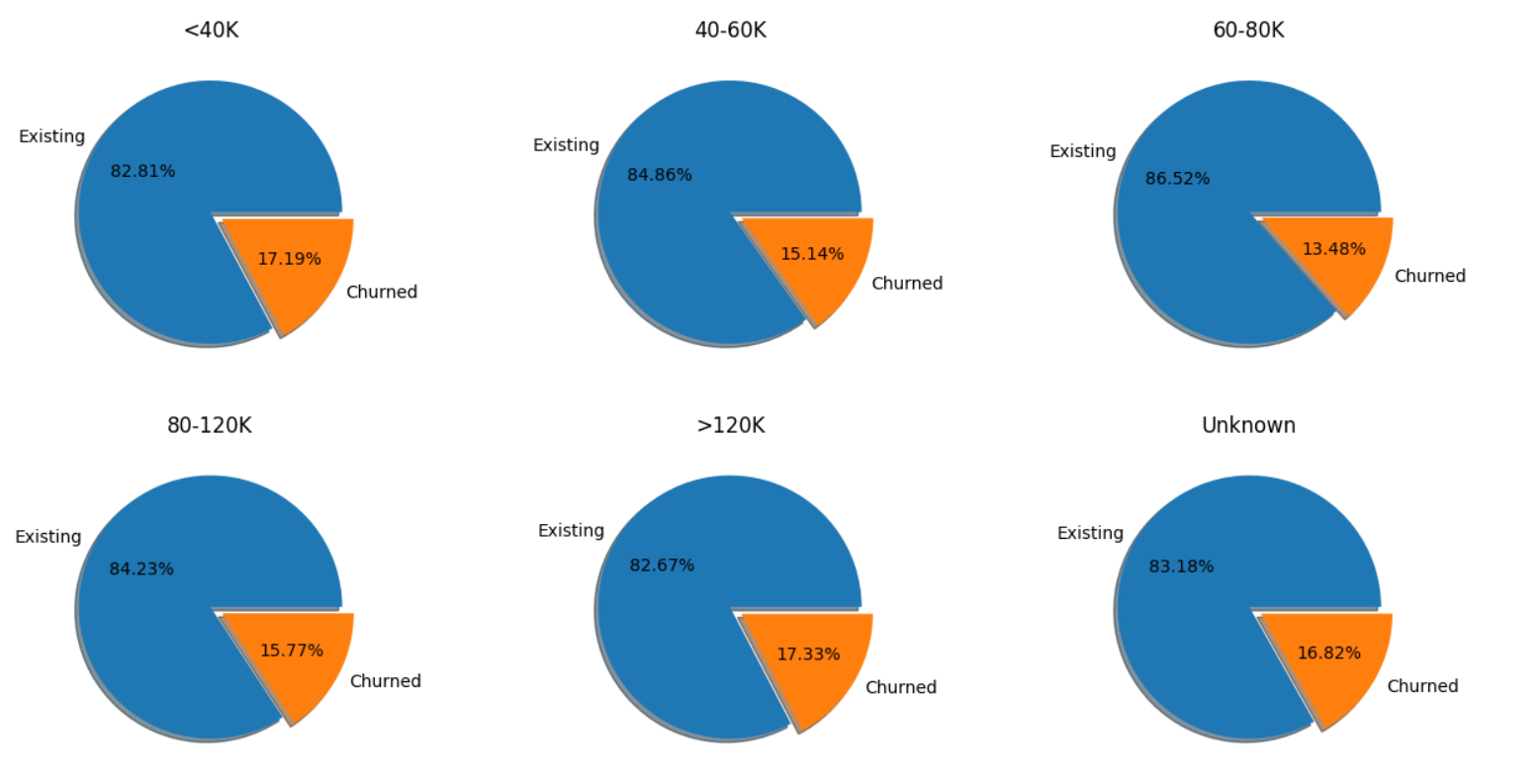
We have only 16% of customers in available dataset that got churned. So, this is gonna be bit difficult to train our model to identify churning customers.

****

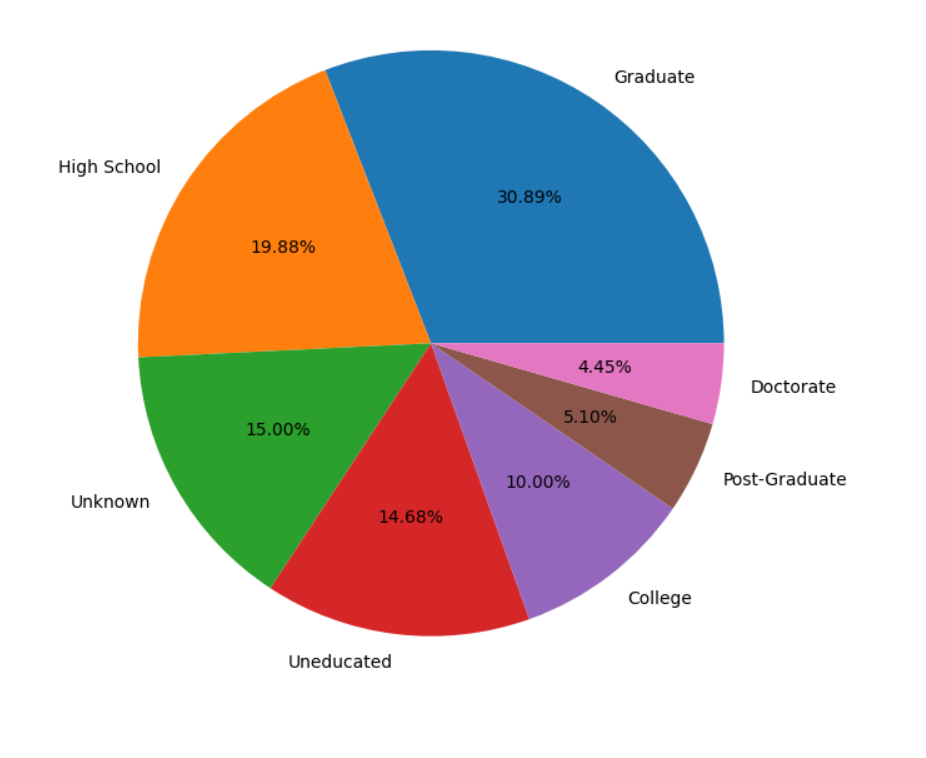
Females have a higher proportion of getting churned though by a very little margin.



Platinum> Gold> Blue> Silver (order of proportions of churned customers). Silver category is an average category with no so high annual charges and better benefits than lower categories. Thus, it has highest retaining power.



Surprisingly, people with highest salary are getting churned with highest proportion.



Distribution of education qualification among our given users. Considering unknown education as uneducated, it’s fair to say that >70% of our users have formal education.

**Data Processing:-**

* Converting categorical variables into numerical variables

Categorical Variables = ['Gender’, ’Education\_Level’, ’Marital\_Status’, ’Income\_Category', 'Card\_Category']

* We can simply hot encode Gender (2 levels) and Marital Status (3 levels)
* Education level, income category & Card category are ordinal variables, can be easily converted into numerical
* Dealing with missing values

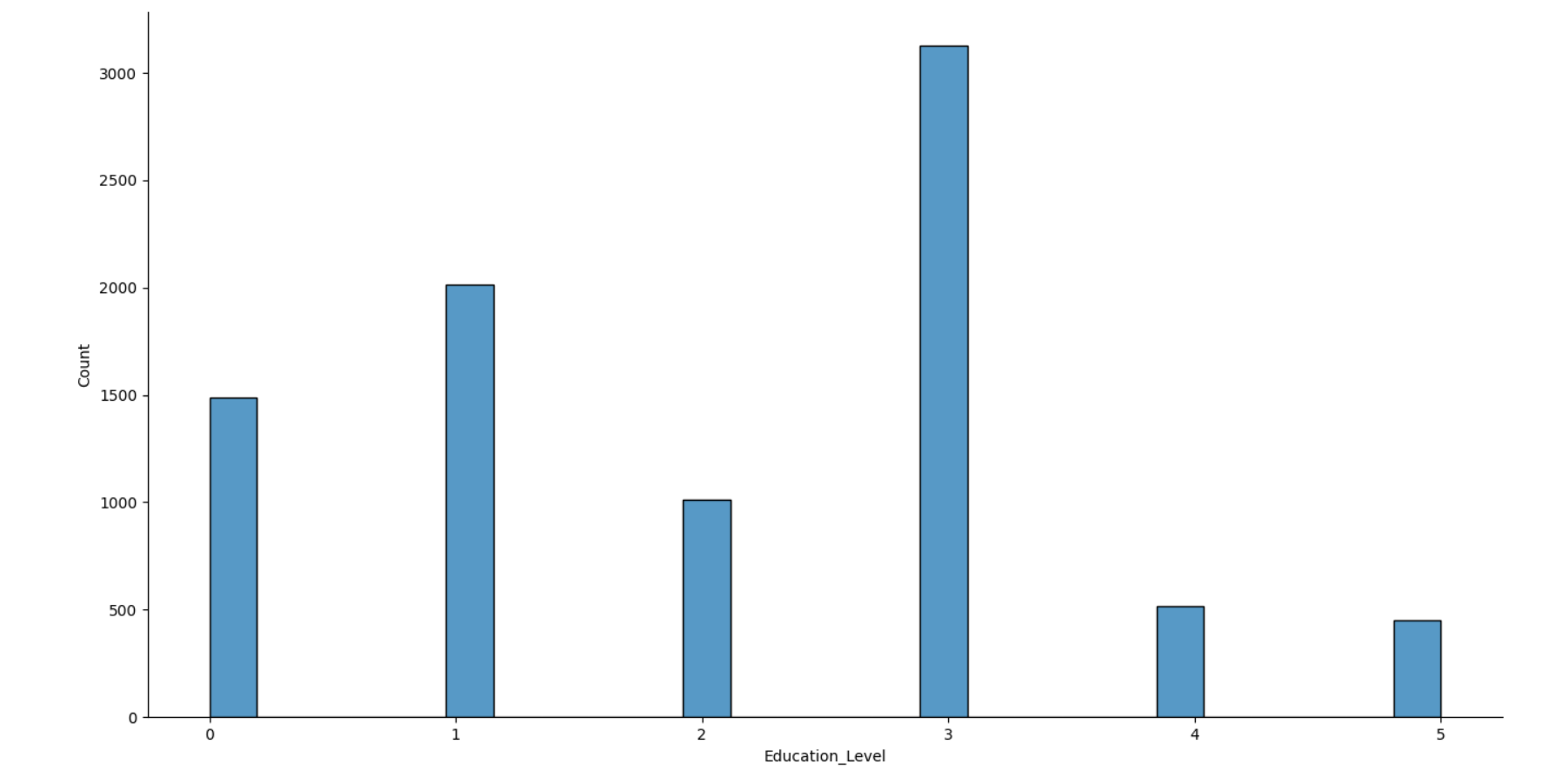
Columns % of missing values

Income\_Category ~ 11%

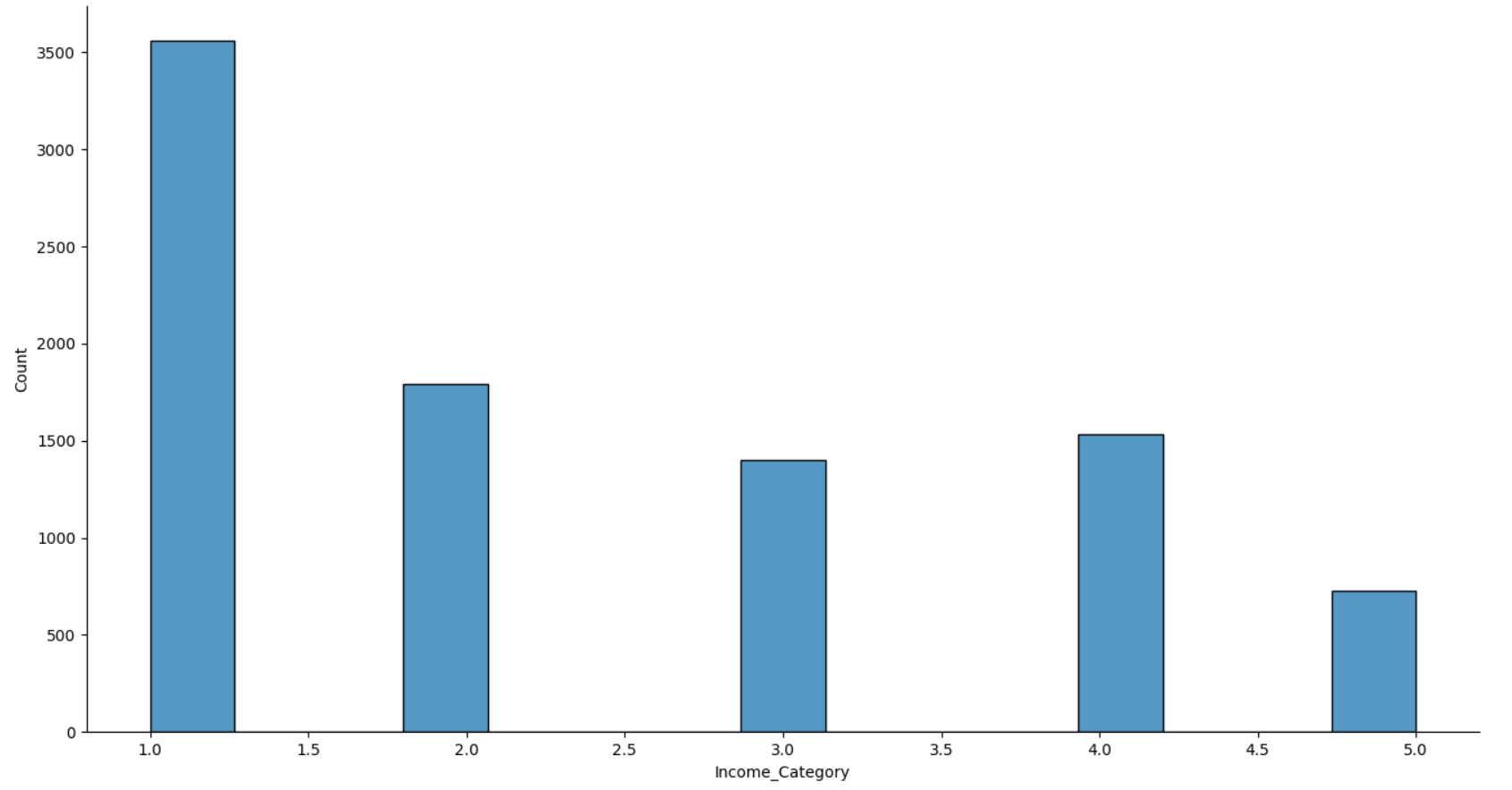
Marital\_Status ~ 7.4%

Education\_Level ~ 15%

* Since, we are hot encoding marital\_status, let’s just treat ‘Unknown’ (missing values) as another level of the category.
* Education Level has skewness values of approx. 0.1 which is quite less, thus we can say that it’s normally distributed and let’s assign its mean to the missing values



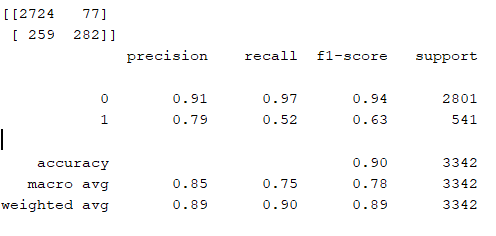
* Income\_category has skewness values of approx. 0.5 which is comparatively on higher side but not that large enough to call it skewed. Let’s assign its median to missing values.



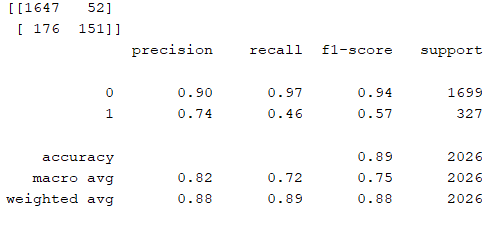
* I’ll also create 2 different columns having value as 1 where education is missing else 0. Same one for income column. Let’s do this after creating a ML model.

**Creating a ML model & evaluation:-**

1. Creating a Logistic Regression Model with 0.33 as sample size. Following is the screenshot of evaluation matrix

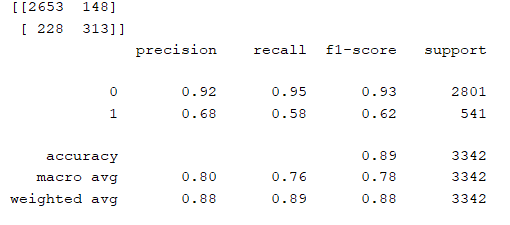


1. Creating a Logistic Regression Model with 0.2 as sample size. Following is the screenshot of evaluation matrix. (Our performance has degraded a little bit, so we’ll stick to sample size as 0.33)



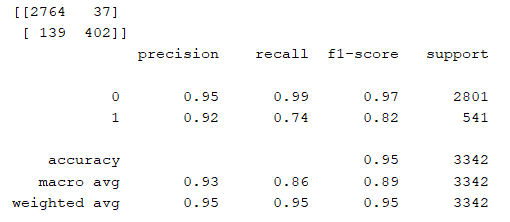
Our objective is to predict customers that can leave our business, so that we can offer them better offers/services to retain them. Our high cost is associated with ‘False Negatives’. That is **we don’t want our model to predict leaving customers as staying customers**. To get a higher recall for output ‘1’.

1. Creating a Naïve Bayes Classification Model with 0.33 as sample size. Following is the screenshot of evaluation matrix



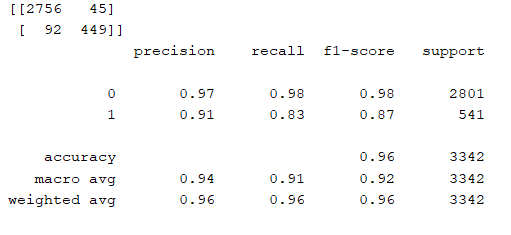
Though precision has fallen but our recall for ‘1’ has improved.

1. Using random forest classifier improves the model performance from a very great margin



Our recall went up to 74%.

1. Let’s do upsampling from our training dataset and then train our model using random forest classifier.



Our recall is now 83%, which is quite as compared to previous models.