PREDICT
CREDIT CARD CHURN
CASE STUDY

JAMEEL KHAN

Objective

- To predict and intervene Credit Card customers before they renounce Credit card usage.
- Credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged on every user irrespective of usage, while others are charged under specified circumstances.
- Customers' leaving credit cards services would lead bank to loss, so the bank wants to analyze the data of customers' and identify the customers who will leave their credit card services and reason for same
- What are the different factors which affect this? What business recommendations can we give based on the analysis?
- How can we improve model performance using hyperparameter tuning and prevent data leakage using pipelines while building a model to predict the response of a customer?

Data Dictionary

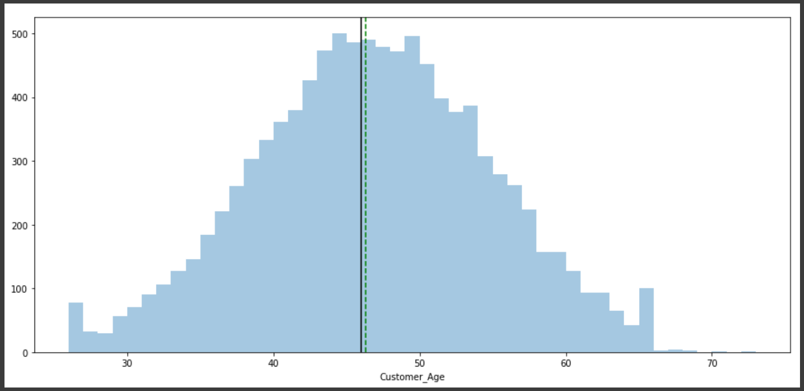
Dataset Column	Description
CLIENTNUM	Client number. Unique identifier for the customer holding the account
Attrition_Flag	Internal event (customer activity) variable - if the account is closed then 1 else 0
Customer_Age	Age in Years
Gender	Gender of the account holder
Dependent_count	Number of dependents
Education_Level	Educational Qualification of the account holder
Marital_Status	Marital Status of the account holder
Income_Category	Annual Income Category of the account holder
Card_Category	Type of Card
Months_on_book	Period of relationship with the bank
Total_Relationship_Count	Total no. of products held by the customer
Months_Inactive_12_mon	No. of months inactive in the last 12 months
Contacts_Count_12_mon	No. of Contacts in the last 12 months
Credit_Limit	Credit Limit on the Credit Card
Total_Revolving_Bal	Total Revolving Balance on the Credit Card
Avg_Open_To_Buy	Open to Buy Credit Line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in Transaction Amount (Q4 over Q1)
Total_Trans_Amt	Total Transaction Amount (Last 12 months)
Total_Trans_Ct	Total Transaction Count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in Transaction Count (Q4 over Q1)
Avg_Utilization_Ratio	Average Card Utilization Ratio

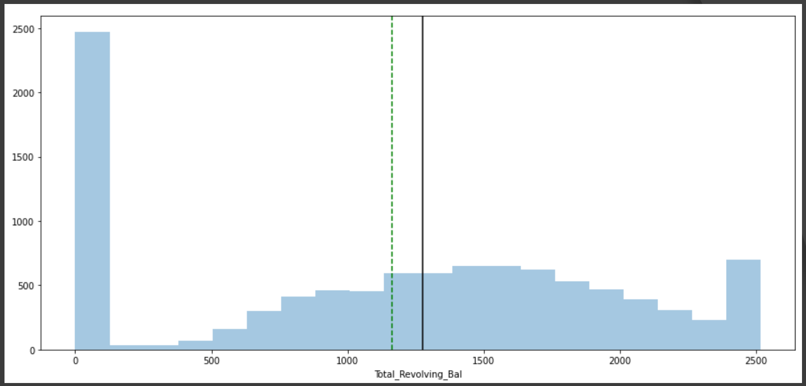
The dataset contains Customer centric and other metrics of customers

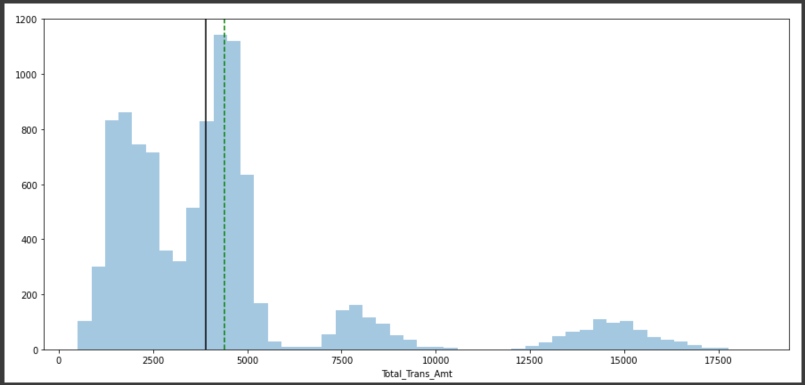
Observations on Data Set:

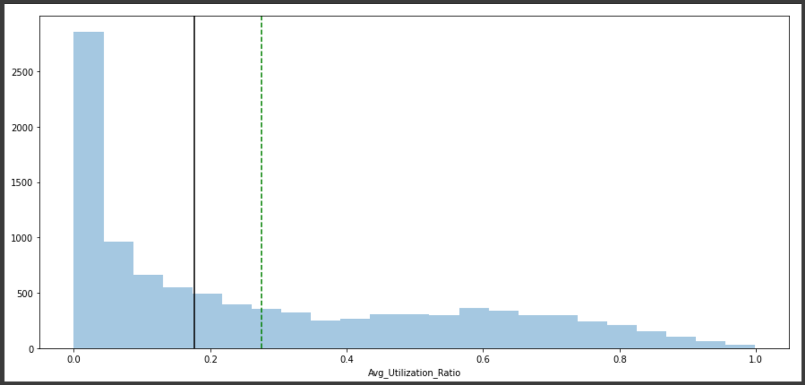
1. There are 21 columns of data for each customer, with a total of 5000 rows of data
2. The dataset has some missing unknown data that will get KNN imputation

Exploratory Data Analyses – Distribution





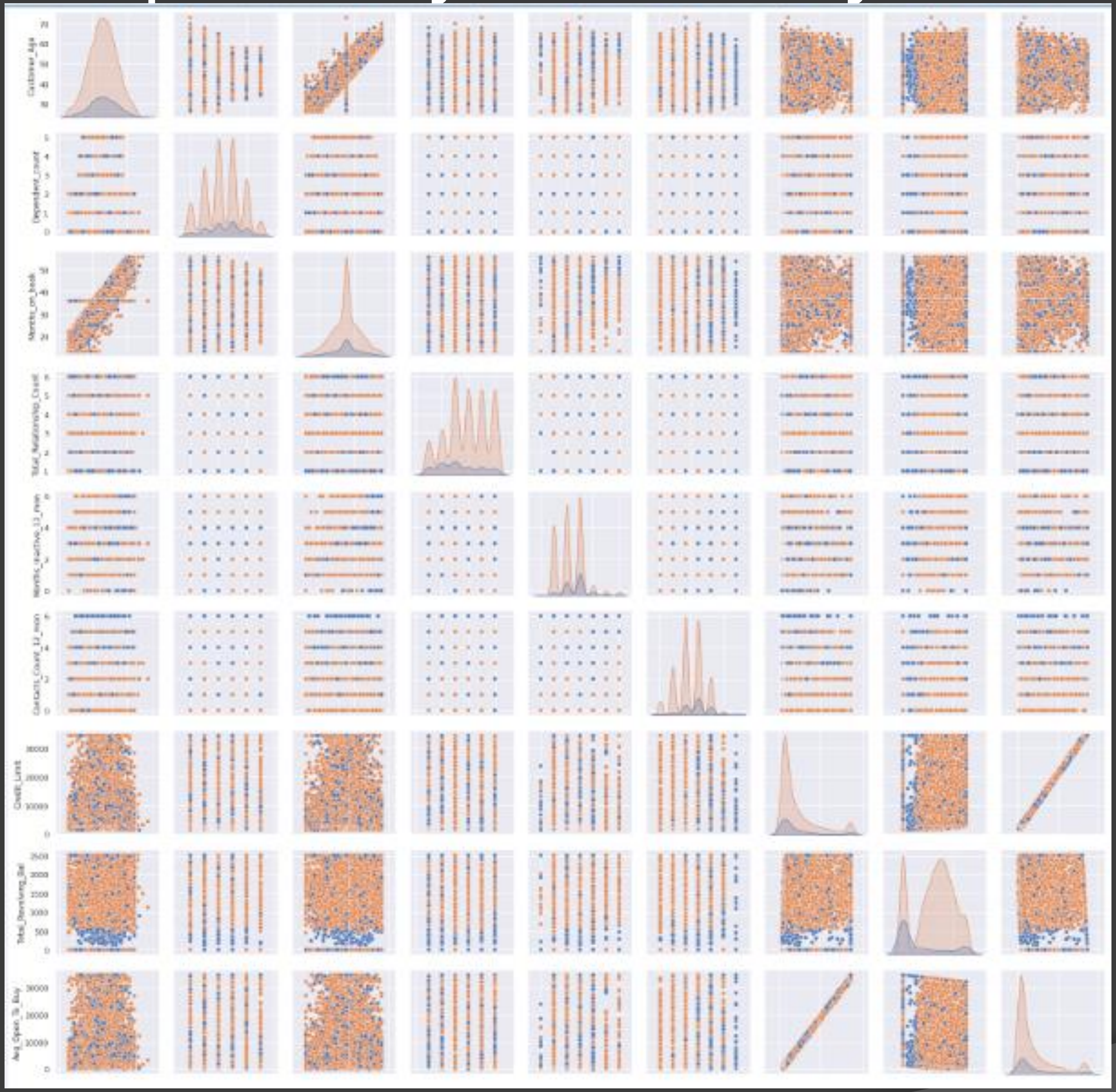




Observations on Continuous Data Distribution in the dataset:

1. Age is normally distributed. Underserved age category of customers belong to the age group of 20 to 30 year olds and 65 and over.
2. Customers carry low balance. Mean is around $1250. Some customers carry higher $2500 balances too.
3. Transaction Amount looks to be a bimodal distribution. Customers with high transaction amounts could be offered more cards to improve relationship
4. Significant proportion of users are underutilizing the card. Mean and median is around 25% only.

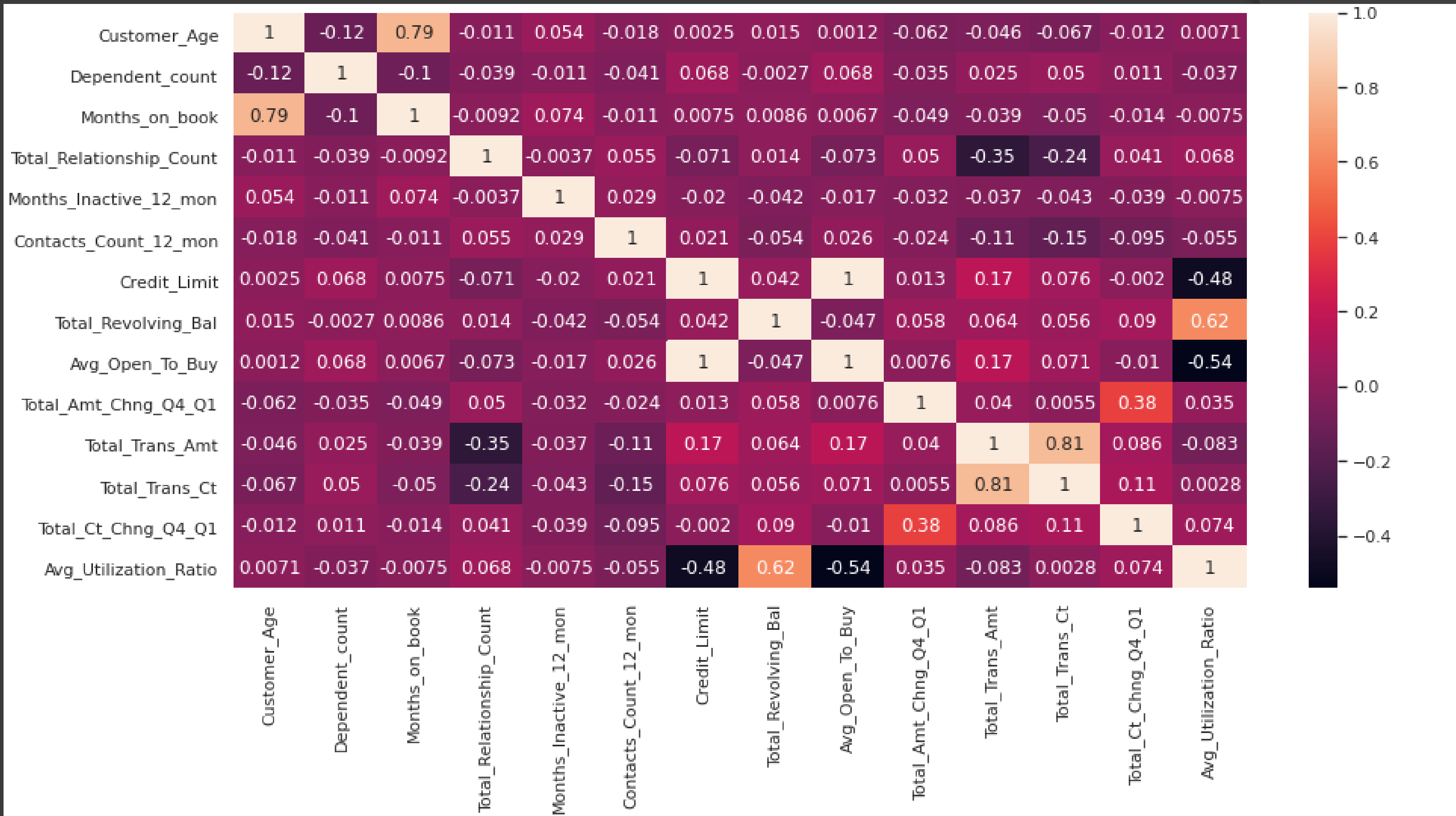
Exploratory Data Analyses – Correlation



Observation on pair plots

1. Using hue of Attrited customer, the data imbalance is clear in each feature. It is about 16%.
2. Some features have high correlation. We will drop some of these features.
3. Customer age and Month on Books are highly correlated
4. Credit Limit and Average Open to buy are highly correlated.
5. Most of the features are categorical - We'll map these features into integers before modelling.

Exploratory Data Analyses – Correlation



Observation on Correlation Matrix of Numeric Features

1. Customer age and Month on Books are highly correlated. We will drop Month on Books before modelling
2. Credit Limit and Average Open to buy are highly correlated. We will drop Average Open to buy.

ML Model Building – Baseline Logistic Regression

Post Data Preprocessing,

1. We selected Attrited_Customers as Dependent(Predicted) Feature and majority of Numerical features as Independent features(Predictors)

2. Next, we mapped the categorical features into integers to map them into hyperplane for modelling. The mapping is reverted after Test Train split

3. Next, we split the dataset into 70% to Train on and 30% to Test the Model on subsequently and used Stratify feature during the Test-Train split to preserve the proportion of target as in the original dataset.

4. Next, for features that had Unknown datapoint, we replaced it with missing values followed by K Nearest Neighbor imputation. This is a better alternative to dropping data.

5. Due to Data Imblance with 16% minority class, we will run SMOTE to upsample the minority class.

6. For baseline model, we fit logistic regression while investigating upsampling and downsampling. We observed that SMOTE didn’t impact the recall metrics. Regularization didn’t improve baseline.

Model	Metric
Baseline Logistic Regression	Test_Recall = 0.96
Regularized Logistic Regression	Test_Recall = 0.85
Logistic Regression with SMOTE undersampling	Test_Recall = 0.83
Logistic Regression with SMOTE oversampling	Test_Recall = 0.83

ML Model Building – Model Pipeline

1. We built a pipeline with multiple ML models. This pipeline will help expedite model retraining when new dataset is available.

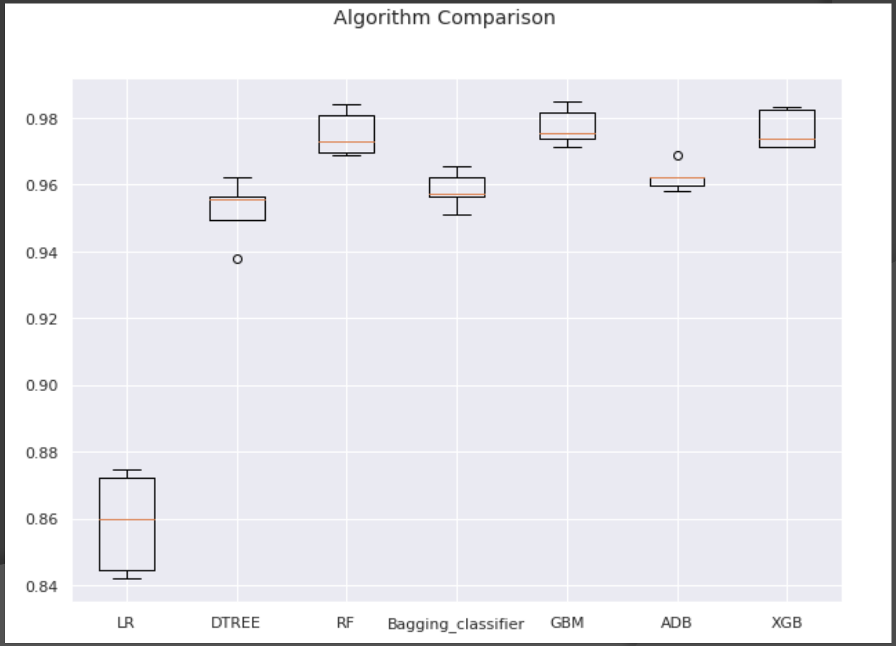
2. Models used for pipeline are Logistic regression, Decision Tree, Bagging classifier, Random Forest and Boosting classifiers like AdaBoost, Gradient Boosting and Xgboost

3. For Model evaluation .

   1. Predicting a customer will stop using the credit card and the customer doesn't - Loss of resources
   2. Predicting a customer will not buy stop using the credit card and the customer does - Loss of opportunity
      Second case is more important with losing on a potential source of credit card revenue. Customer will not targeted by the marketing team when he should be targeted. We will use Recall metric to compare

4. Boxplot shows that RF, GBM and XGB has the highest cross validated recall

5. AdaBoost is also good and has a few outliers, so we will Hypertune this instead of GBM to check its performance. We will also tune RF and XGB.



Hypertuning – GridSearchCV and RandomSearchCV

1. To improve the ML models, we build a parameter grid of most important parameters and then search the best combination of these parameteters – first comprehensively with GridsearchCV and then randomly a few combinations with RandomSearchCV.

2. For RandomForest, pereformance has improved with hypertuned parameters found with GridsearchCV and RandomsearchCV. RandomsearchCV and GridsearchCV took about the same time. If we tune more param_grid paramaters, the GridsearchCV will take longer

3. For AdaBoost, the test recall has increased as compared to cross validated recall, even with outliers. Grid search took significantly longer time(15 min) than random search(5min).

4. For XGBoost, the test recall has increased further after model tuning. Although both Test and Train Recall is high, there is a chance that the model is overfitting the training data. We could retrain the model with more data, if available. RandomsearchCV is much faster than Grid search. This is important for quick deployment and training

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
3	XGBoost with RandomizedSearchCV	0.695243	0.881540	0.997983	0.998824	0.621610	0.877108
1	Decision Tree with RandomizedSearchCV	0.993444	0.987456	0.993444	0.987456	0.989950	0.972963
0	Decision Tree with GridSearchCV	0.997815	0.969398	0.998319	0.984320	0.997313	0.979329
2	XGBoost with GridSearchCV	0.839301	0.920039	0.981846	0.978832	0.764029	0.929635

Business Recommendations

1. Months inactive have a increasing linear trend from 1 to 3 months and then there is a sharp drop, whi ch might imply that the account was closed or customers started reusing accounts.
2. Customers were contacted generally only 2 to 3 times. If this is increased, we could get valuable feed back in the 3 months of inactivity then the attrition could be reduced.
3. Based on feature importance, the most important feature is the number of cards held by the custome This implies that we can offer more cards to select customers to prolong the relationship
4. Next important feature is inactive_months. This would directly indicate churn
5. The next two important features are transaction count and amount, which is also intuitive.
6. Blue cards are the most popuplar at 93%. We can investigate which customers could be offered othe r cards.
7. Interesting to note that ~80% customers have 3 or more accounts