PREDICTING THE LIKELIHOOD OF CREDIT CARD CHURN

BY PHONG PHAM

THE PROBLEM

- ABC Bank management is concerned with 16% of customers leaving their credit card services in 2020
- The bank wants to find data-driven solutions to encourage customers to continue using their cards.

OBJECTIVES

- To increase service level to customers who are most likely to churn on credit cards.
- To encourage credit card utilization to keep churn rate below 10% by the end of 2021.

STRATEGIES

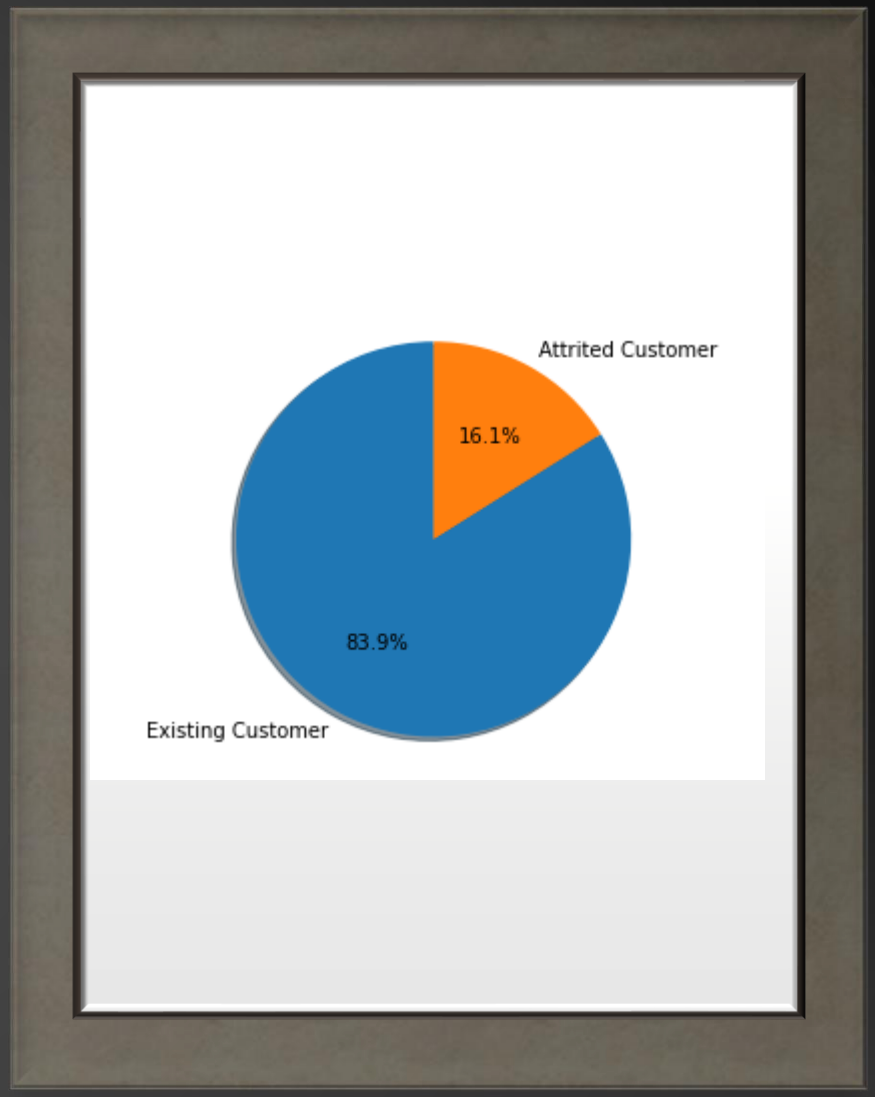
- Analyze demographics, banking and credit card usage details of more than 10,000 customers who have used at least 1 credit card issued by ABC Bank.
- Find out the main reasons that lead to credit card churning and come up with possible solutions to reduce churning rates.

DATA INFORMATION

- Dataset downloaded from: Credit Card customers | Kaggle
- File format: CSV
- 10,120 records and 21 columns, each records represents a customer
- 9 data columns are numerical columns, the rest are categorical.
- Target variable is Attrition Flag, which is a categorical variable

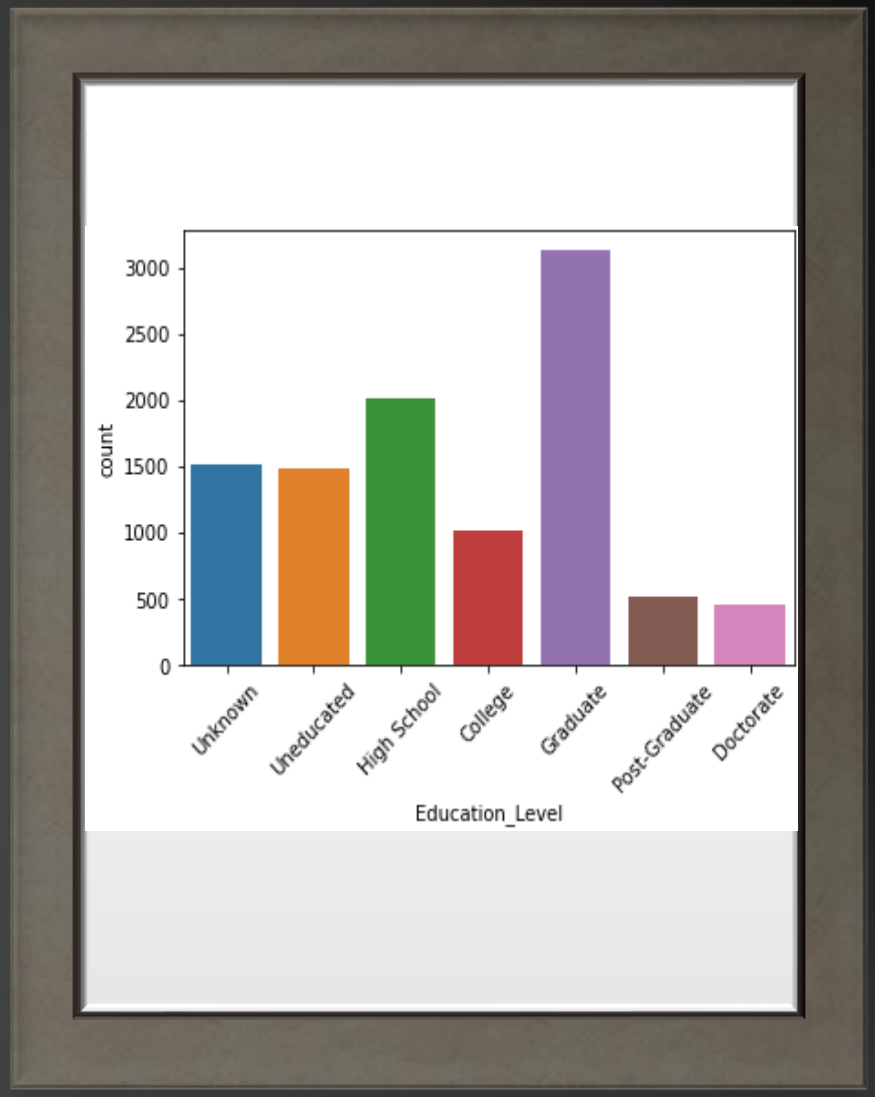
DATA EXPLORATION: CLASS IMBALANCE

- Only 16.1% of all customer have churned on their credit cards.
- Might lead to prediction inaccuracies with too many false negatives (i.e. actual churners but labeled as "not churn")



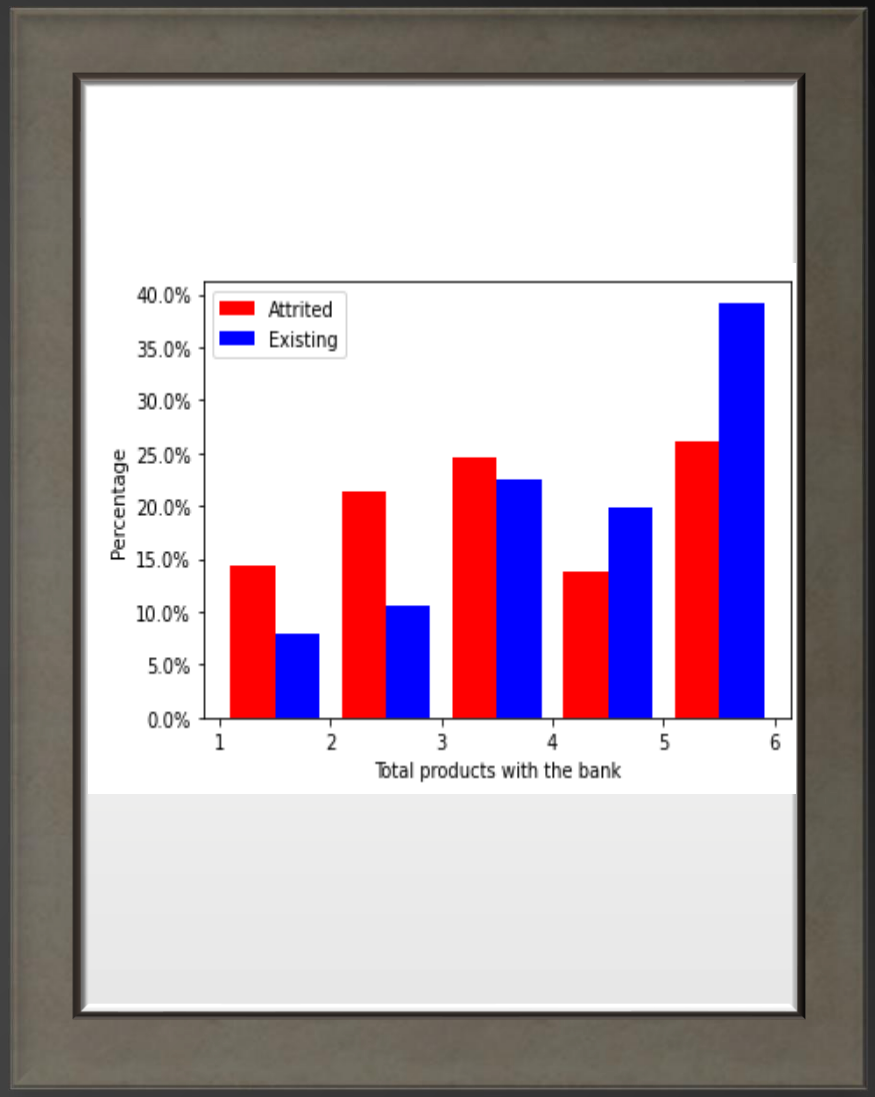
DATA EXPLORATION: EDUCATION LEVEL

- The number of people who have graduate degrees is the highest, followed by high school degree holders
- A considerable number of customers did not report their education level



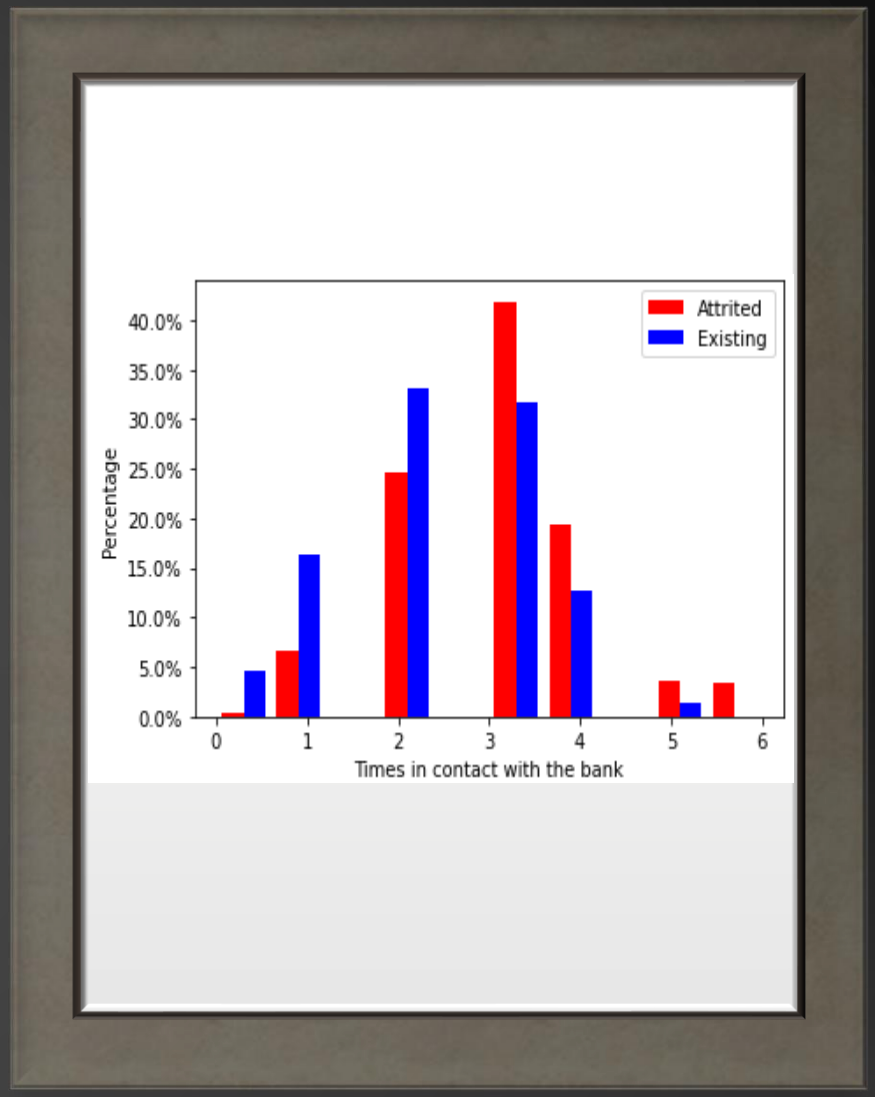
DATA EXPLORATION: NUMBER OF BANK PRODUCTS

- Customers who have 4 with the bank are least likely to churn on their credit cards.
- People with 5 products account for the highest percentages of both "churned" and "not churn"



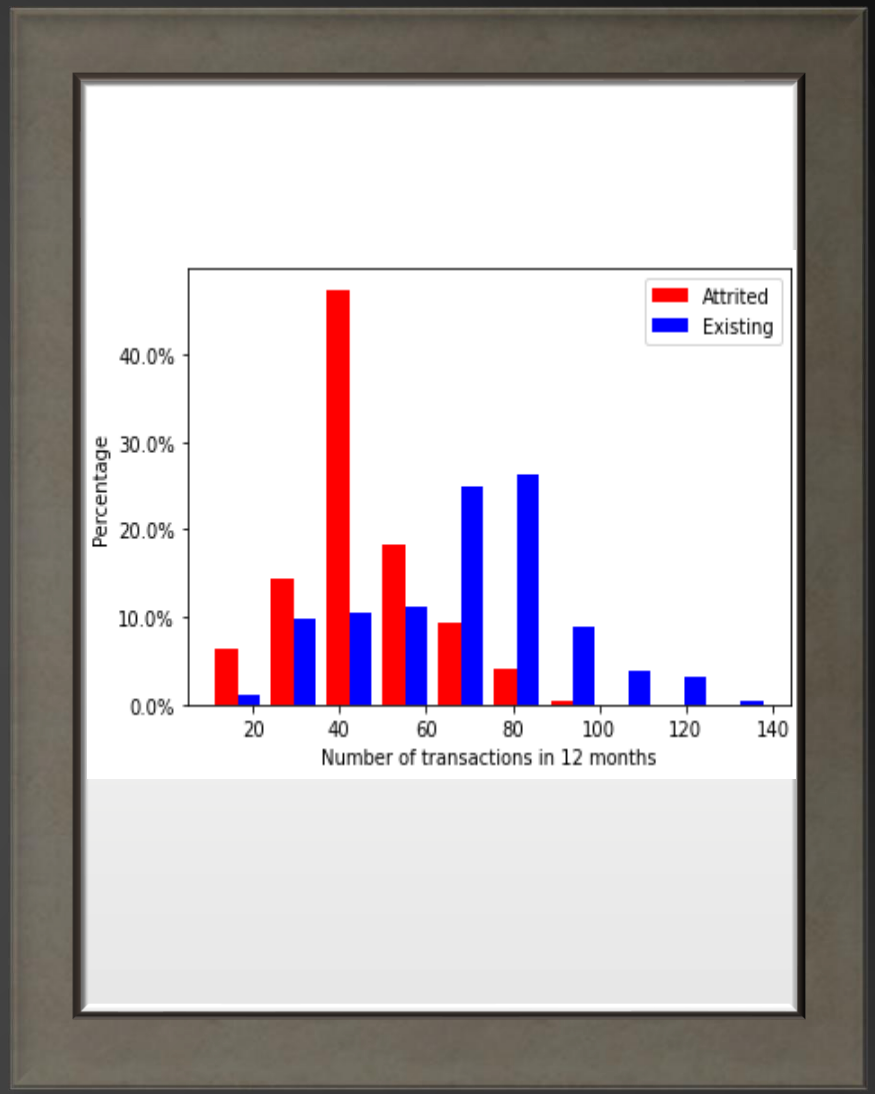
DATA EXPLORATION: CONTACT FREQUENCY

- Customers contacting the bank 3 or more times will be more likely to churn on their cards.
- They might not happy about regarding a product offered by the bank.



DATA EXPLORATION: TRANSACTIONS

- Attritted customers are more likely to use their cards less often (60 times or fewer in 12 months).



MACHINE LEARNING: PREPROCESSING STEPS

1. One-hot encoding for categorical variables
2. Scaling numerical variables to a range between 0 and 1
3. Oversampling the minority class (Attritted Customers) using SMOTE to have a balanced dataset of 50% of each class
4. Split data into training and testing sets (with a 75:25 ratio)

MACHINE LEARNING: MODELING

- The following classification algorithms were used
  - Logistic Regression
  - K-Nearest Neighbor (KNN)
  - Support vector machine (SVM)
  - Random Forest
  - Gradient Boosting
- Accuracy, recall, cross validation scores were used to evaluate each model

MACHINE LEARNING: MODELING

Algorithm	Accuracy	Recall	CV test score
Logistic Regression	0.8545	0.8538	0.9272
K-Nearest Neighbors	0.8437	0.9678	0.8747
SVM	0.9009	0.9101	0.9449
Random Forest	0.9602	0.9546	0.9865
Gradient Boosting	0.9569	0.9484	0.9864

- Random Forest and Gradient Boosting yielded the best accuracy and CV test scores
- Hyperparameter tuning done on these 2 methods using randomized cross validation

MACHINE LEARNING: MODELING

- Results from using optimal parameters for Random Forest and Gradient Boosting

Algorithm	Accuracy	Recall	CV test score
Random Forest	0.9642	0.9612	0.9892
Gradient Boosting	0.9653	0.9574	0. 9918

- The fewer false negatives the better the prediction model

  ➔ Random Forest should be used for future predictions

CONCLUSIONS

- 21 original features were engineered into 29, all of which were used in predicting the attrition flag
- Random Forest provided the best results out of 5 supervised classification models

SOLUTIONS

- Running the Random Forest model every month to find out if any existing customers are flagged as potential churners
- Contact these customers to get feedback on the credit cards they own and the experience with the bank
- Customize promotions for each group of customers

LIMITATIONS AND FUTURE ACTIONS

- Data was taken only in the last 12 months

  ➔ Useful to incorporate data from at least another year for training the model

- Many records lack demographic details (Unknown education level, income, marital status)

  ➔ Reach out to customers to update these details