PROJECT

BANK CUSTOMER CHURN PREDICTION

VUONG HUY HOANG

Presentation

WHY IS CUSTOMER CHURN PREDICTION IMPORTANT?

Churn is expensive. The cost of any new customer acquisition is always higher than the cost of retaining existing customers.

Customer churn prediction identifies which customers are at a high risk of canceling their subscription or abandoning your product. Churn prediction, therefore, tells you whether a customer will leave and why.

ABOUT THE DATASET

The dataset contains:

- 10,000 rows – each row is a unique customer of the bank
- 13 columns:

Customer ID: A unique identifier for each customer
Surname: The customer's surname or last name
Credit Score: A numerical value representing the customer's credit score
Geography: The country where the customer resides
Gender: The customer's gender
Age: The customer's age.
Tenure: The number of years the customer has been with the bank
Balance: The customer's account balance
NumOfProducts: The number of bank products the customer uses (e.g., savings account, credit card)
HasCrCard: Whether the customer has a credit card
IsActiveMember: Whether the customer is an active member
EstimatedSalary: The estimated salary of the customer
Exited: Whether the customer has churned (Target Variable)
Dataset link: Bank Customer Churn Prediction

BUSINESS UNDERSTANDING

The data will be used to predict whether a customer of the bank will churn. If a customer churns, it means they left the bank and took their business elsewhere. If you can predict which customers are likely to churn, you can take measures to retain them before they do. These measures could be promotions, discounts, or other incentives to boost customer satisfaction and, therefore, retention.

TABLE OF CONTENTS

DATA PREPARATION

DATA PREPARATION

```
[64] df.isnull().values.any()
```

```
False
```

```
[65] df.dtypes
```

```
CustomerId           int64
Surname             object
CreditScore          int64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

```
[66] ## removing the columns we do not need - row number , customerid and surname
     df.drop(columns = ['CustomerId', 'Surname'], inplace = True)
```

```
[68] df.Geography.unique()
```

```
array(['France', 'Spain', 'Germany'], dtype=object)
```

```
[69] df.Gender.unique()
```

```
array(['Female', 'Male'], dtype=object)
```

```
[70] df['Gender'] = df['Gender'].replace('Male', 1)
     df['Gender'] = df['Gender'].replace('Female', 0)
```

```
[71] df['Geography'] = df['Geography'].replace('France', 1)
     df['Geography'] = df['Geography'].replace('Spain', 2)
     df['Geography'] = df['Geography'].replace('Germany', 3)
```

Checking for null values, data types, encoding categorical features using for prediction model.

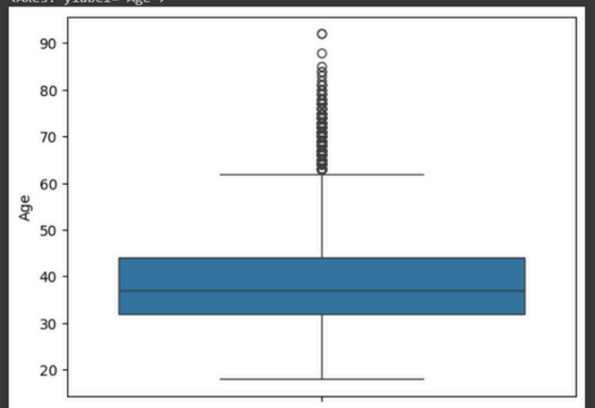
DATA PREPARATION

```
[73] ##check the outliers. Logistic regression algorithm does not like outliers
     df.describe()
     ## looking at the 75th percentile and the max values, we can see that the data has outliers for Salary, balance, age
```

	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
count	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.000000	10000.00000	10000.000000	10000.000000	10000.000000
mean	650.528800	1.749500	0.545700	38.921800	5.012800	76485.889288	1.530200	0.70550	0.515100	100090.239881	0.203700
std	96.653299	0.830433	0.497932	10.487806	2.892174	62397.405202	0.581654	0.45584	0.499797	57510.492818	0.402769
min	350.000000	1.000000	0.000000	18.000000	0.000000	0.000000	1.000000	0.00000	0.000000	11.580000	0.000000
25%	584.000000	1.000000	0.000000	32.000000	3.000000	0.000000	1.000000	0.00000	0.000000	51002.110000	0.000000
50%	652.000000	1.000000	1.000000	37.000000	5.000000	97198.540000	1.000000	1.00000	1.000000	100193.915000	0.000000
75%	718.000000	3.000000	1.000000	44.000000	7.000000	127644.240000	2.000000	1.00000	1.000000	149388.247500	0.000000
max	850.000000	3.000000	1.000000	92.000000	10.000000	250898.090000	4.000000	1.00000	1.000000	199992.480000	1.000000

```
##visualize the outliers for age => limit the age to 85
sns.boxplot(df, y= 'Age')
```

```
<Axes: ylabel='Age'>
```




```
[75] ## limiting the data set to have only records of age less than 85
     df = df[df['Age']<85]
```

```
[76] scaler = MinMaxScaler()
     df[['CreditScore','EstimatedSalary','Balance','Age']] = scaler.fit_transform(df[['CreditScore','EstimatedSalary','Balance','Age']])
```

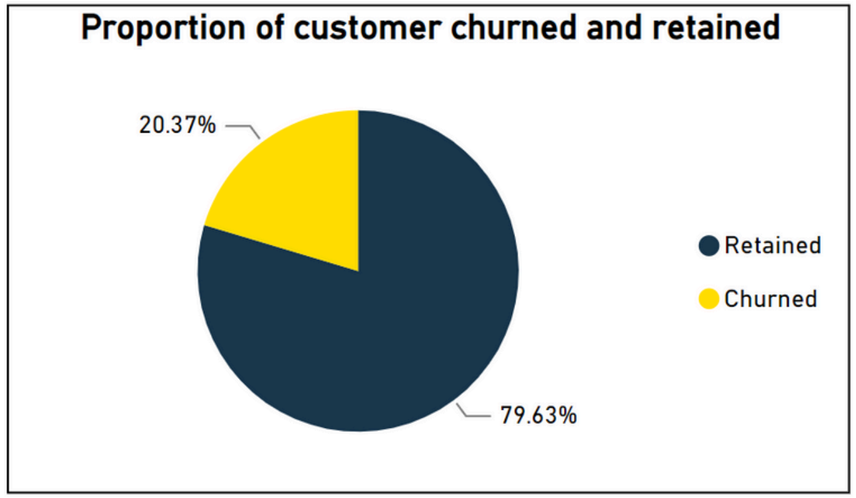
```
<ipython-input-76-3c411de6fcd2>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df[['CreditScore','EstimatedSalary','Balance','Age']] = scaler.fit_transform(df[['CreditScore','EstimatedSalary','Balance','Age']])
```

```
[77] X = df.drop('Exited',axis = 1)
     y = df['Exited']
```

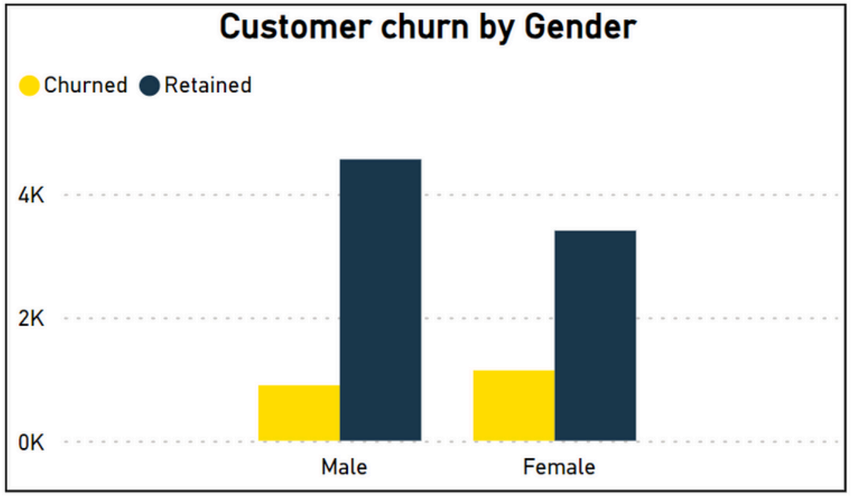
Checking for outliers, scaling features and split data into input and target set (X and y)

DESCRIPTIVE ANALYSIS

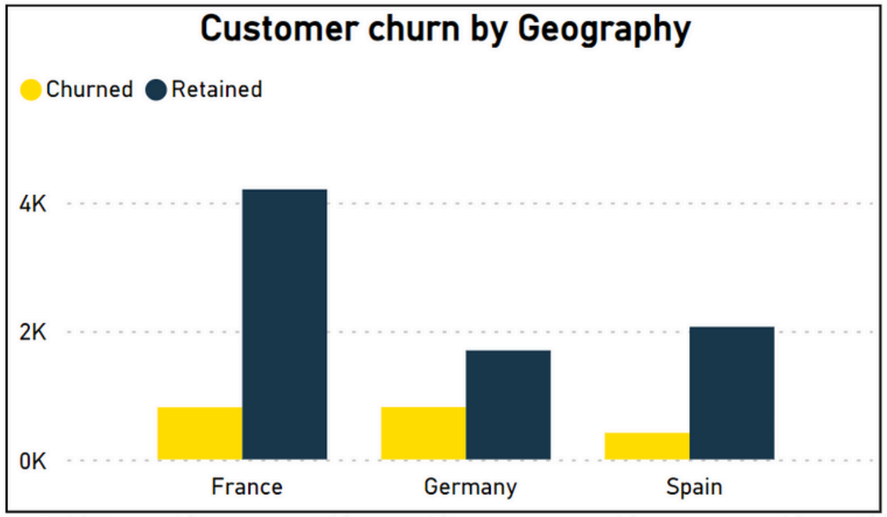
DESCRIPTIVE ANALYSIS



- Around 1 in 5 customers are leaving, which is a significant proportion that warrants investigation.



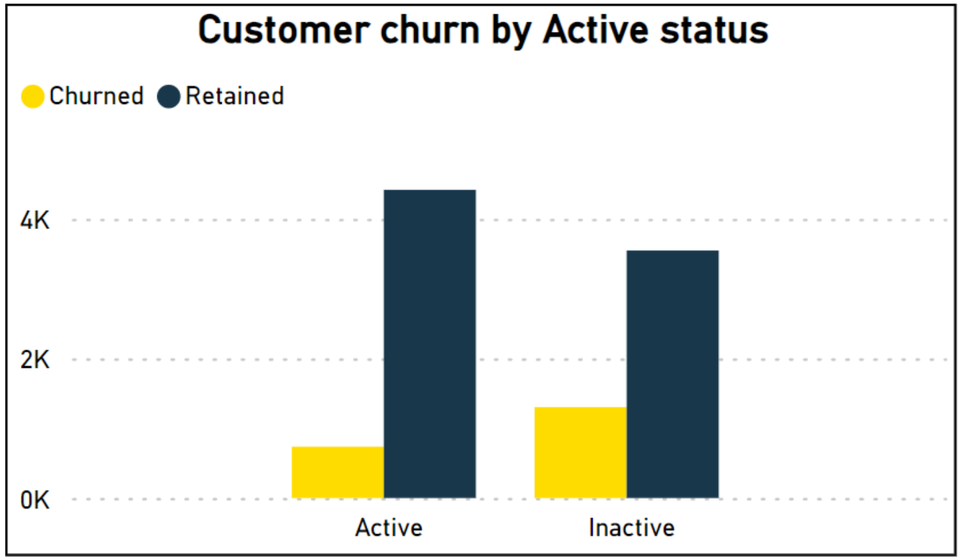
- The proportion of female customers churning is greater than that of male customers.



- France: Higher number of churned customers compared to Germany and Spain, indicating a possible issue in this market.
- Germany: Moderate churn rate.
- Spain: Lowest churn rate but also the smallest customer base.

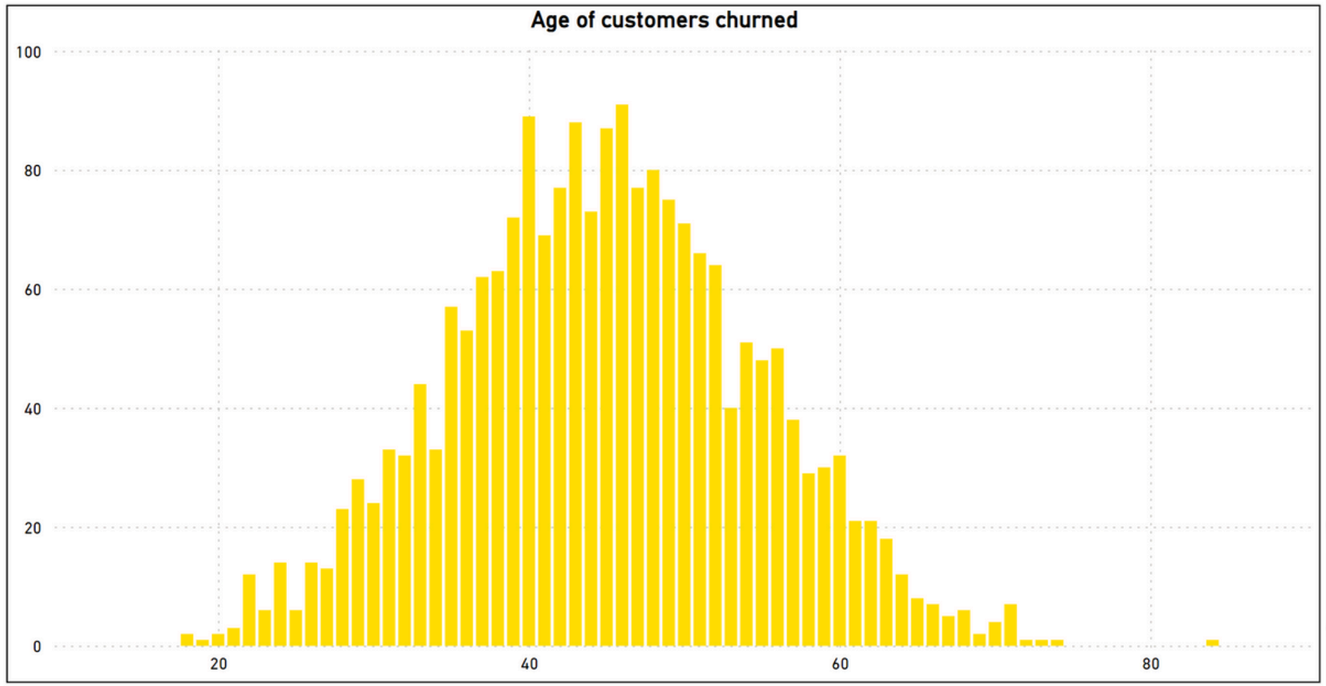
=> ***The higher churn in France could be due to market saturation, stronger competition, or specific customer dissatisfaction issues.***

DESCRIPTIVE ANALYSIS



- Inactivity appears to be a strong predictor of churn, suggesting that engagement levels are crucial for retention.

- Peak churn ages are around 30 to 50.
- Younger (around 20) and older customers (above 60) have lower churn rates. This might be due to life stage factors, where middle-aged customers have changing priorities, financial constraints, or less time.



PREDICTION

PREDICTION

```
[ ] model = LogisticRegression(solver='liblinear', C=0.05, multi_class='ovr',
                               random_state=0)
    model.fit(X,y)
```

```
LogisticRegression(C=0.05, multi_class='ovr', random_state=0,
                   solver='liblinear')
```

```
[ ] y_pred =  model.predict(X)
```

```
[ ] y_pred= np.where(y_pred>0.5,1,0)
```

```
[ ] print('MSE', mean_squared_error(y, y_pred))
```

```
MSE 0.1874749899959984
```

- To identify high-risk customers before they churn, in this case, we use logistic regression algorithm.
- MSE Value: The MSE value here is 0.1875, which suggests that, on average, the squared differences between the actual and predicted values are relatively low.

```
[ ] data = pd.DataFrame({'Actual':y, 'Predicted':y_pred})
```

```
[ ] report = classification_report(y, y_pred)
    print(report)
```

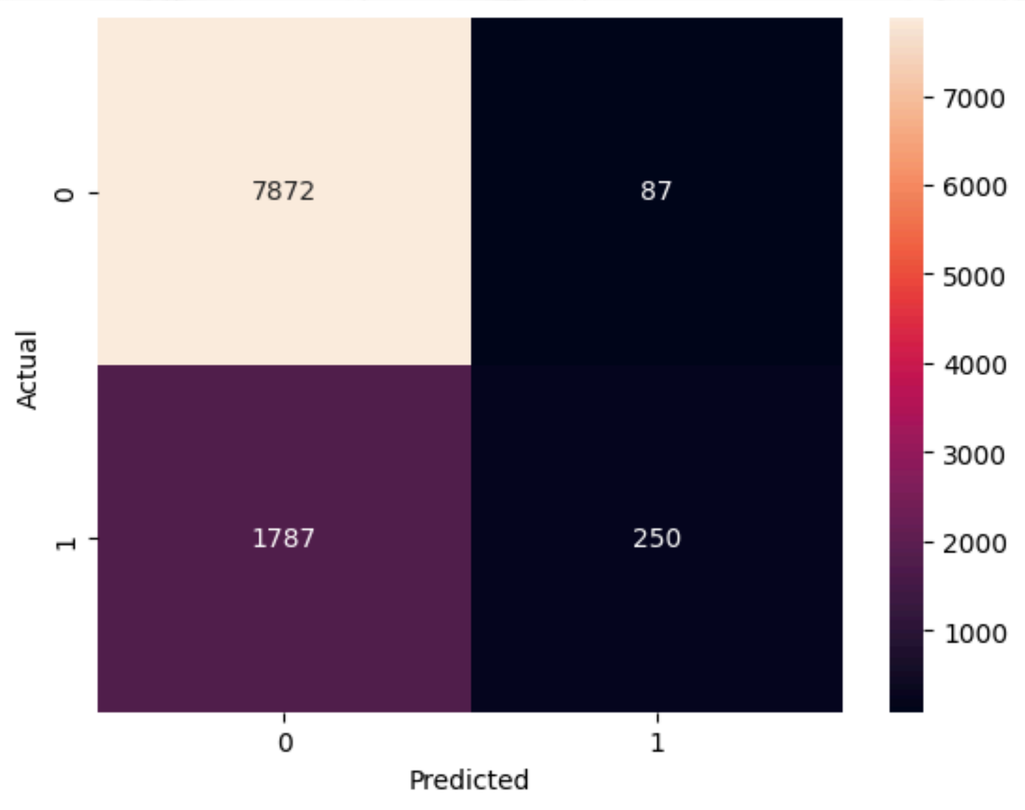
```
              precision    recall  f1-score   support

           0       0.81      0.99      0.89      7959
           1       0.74      0.12      0.21      2037

    accuracy                           0.81      9996
   macro avg       0.78      0.56      0.55      9996
weighted avg       0.80      0.81      0.75      9996
```

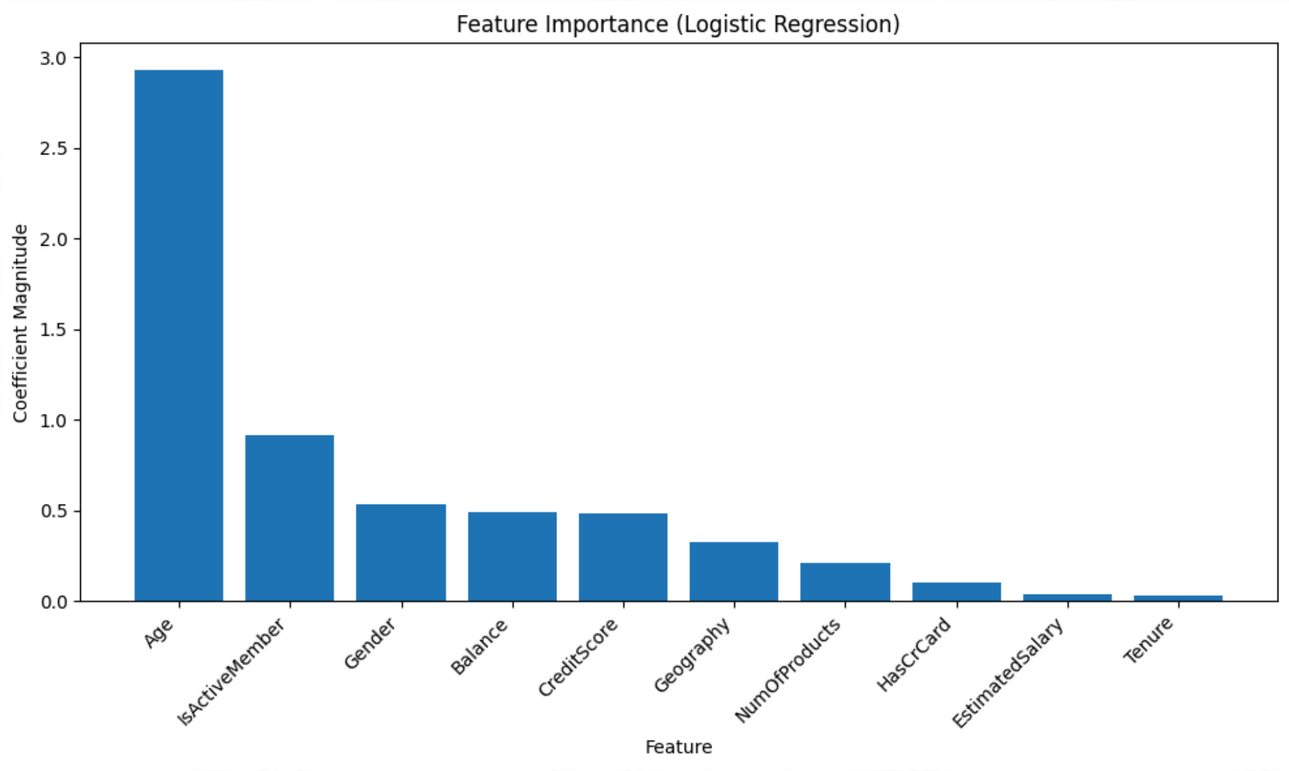
```
[ ] print("Accuracy of model:",accuracy_score(y,y_pred)*100,'%')
```

```
Accuracy of model: 81.25250100040016 %
```

- The accuracy of model after training is 81.25%.

PREDICTION





- The true negatives (7872) are significantly higher compared to the other values in the confusion matrix, this suggests the model is very effective at predicting customers who will be retained (0) correctly. However, it struggles with identifying churned customers accurately, as indicated by the high number of false negatives (1787).

- Age is the most significant predictor of customer churn, with a high positive coefficient. This suggests that as age increases, the likelihood of churn also increases.
- Following by IsActiveMember, Gender, Balance and CreditScore.

PREDICTION

```
# Get feature coefficients
coefficients = model.coef_[0]
intercept = model.intercept_[0]
feature_names = ['CreditScore', 'Geography','Gender', 'Age' ,'Tenure',  'Balance',  'NumOfProducts',  'HasCrCard',  'IsActiveMember',
                 'EstimatedSalary']
```

```
print(coefficients)
```

```
[-0.48419998  0.32425992 -0.53086122  2.93040224 -0.02641487  0.48819082
 -0.21151064 -0.10388174 -0.91244226 -0.03373943]
```

Feature Coefficients

- **Age:** 2.93040224

Positive coefficient: As age increases, the likelihood of churn increases.

- **IsActiveMember:** -0.91244226

Negative coefficient: Being an active member decreases the likelihood of churn.

- **Gender:** -0.53086122

Negative coefficient: Males being less likely to churn.

- **Balance:** 0.48819082

Positive coefficient: Higher balance increases the likelihood of churn.

SOLUTIONS

SOLUTIONS

1. Age

- Current Insight: Older customers are more likely to churn.
- Solutions:
  - Tailored Communication: Develop targeted communication for different age groups. For older customers, personalized and traditional approaches might be more effective, while younger customers might prefer digital and dynamic interactions.
  - Age-specific Promotions: Offer services and promotions that cater to the specific needs of older customers, such as retirement planning, health and wellness programs, or special discounts.
  - Engagement Programs: Create engagement programs specifically designed for older customers, such as community events, exclusive clubs, or personalized service calls.

2. IsActiveMember

- Current Insight: Active members are less likely to churn.
- Solutions:
  - Engagement Campaigns: Develop campaigns to encourage inactive members to become more active. These can include personalized messages, special offers, and reminders about the benefits of using the service.
  - Activity Incentives: Provide incentives for customers to engage more frequently with the platform, such as reward points, discounts, or exclusive access to new features.
  - User Experience Improvements: Ensure the platform is user-friendly and engaging. Regular updates and improvements based on user feedback can keep members active.

3. Gender

- Current Insight: Male customers are less likely to churn.
- Solutions:
  - Gender-specific Marketing: Create marketing campaigns that resonate with female customers to address the higher likelihood of churn. Highlight features and benefits that are particularly appealing to them.
  - Diverse Offerings: Ensure that product offerings and services are inclusive and cater to the preferences of all genders.
  - Feedback Mechanisms: Implement feedback mechanisms to understand the specific needs and pain points of female customers and address them promptly.

4. Balance

- Current Insight: Higher balances increase the likelihood of churn.
- Solutions:
  - Financial Products: Offer financial products that help customers manage and grow their balances effectively, such as investment advice, savings plans, or high-interest accounts.
  - Proactive Support: Provide proactive customer support for those with higher balances, ensuring they feel valued and receive personalized service.
  - Risk Mitigation Strategies: Identify and mitigate risks for high-balance customers by offering insurance products, fraud protection, and financial advice.

THANK YOU!

VUONG HUY HOANG

Presentation