- A Data-Driven Approach to Understand and Prevent Customer Loss

Predicting Customer Churn in Banking

- Vamshi Krishna Perabathula

Why Are Our Customers Leaving?

- Business Understanding - Problem Definition

- Imagine you're running a bank, and you notice some customers are leaving. It's like they're breaking up with you, and that hurts your business. Our goal is to create a smart system that can predict which customers might leave so we can give them some extra love and keep them around.

Understanding Our Customer Data

- Data Understanding

- We've got a treasure trove of information about our customers - over 10,000 records! It tells a detailed story about each person, including their credit score, demographics, and balance. Notably, about 20% of customers have left the bank, and we want to understand why.
- **Key Points**:
- 10,000 records with customer info.
- Attributes: Credit score, age, location, balance.
- 20% churn rate.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
None
```

```
         RowNumber    CustomerId   CreditScore           Age        Tenure  \
count  10000.00000  1.000000e+04  10000.000000  10000.000000  10000.000000
mean    5000.50000  1.569094e+07    650.528800     38.921800      5.012800
std     2886.89568  7.193619e+04     96.653299     10.487806      2.892174
min        1.00000  1.556570e+07    350.000000     18.000000      0.000000
25%     2500.75000  1.562853e+07    584.000000     32.000000      3.000000
50%     5000.50000  1.569074e+07    652.000000     37.000000      5.000000
75%     7500.25000  1.575323e+07    718.000000     44.000000      7.000000
max    10000.00000  1.581569e+07    850.000000     92.000000     10.000000

             Balance  NumOfProducts    HasCrCard  IsActiveMember  \
count   10000.000000   10000.000000  10000.00000    10000.000000
mean    76485.889288       1.530200      0.70550        0.515100
std     62397.405202       0.581654      0.45584        0.499797
min         0.000000       1.000000      0.00000        0.000000
25%         0.000000       1.000000      0.00000        0.000000
50%     97198.540000       1.000000      1.00000        1.000000
75%    127644.240000       2.000000      1.00000        1.000000
max    250898.090000       4.000000      1.00000        1.000000

       EstimatedSalary        Exited
count     10000.000000  10000.000000
mean     100090.239881      0.203700
std       57510.492818      0.402769
min          11.580000      0.000000
25%       51002.110000      0.000000
50%      100193.915000      0.000000
75%      149388.247500      0.000000
max      199992.480000      1.000000
```

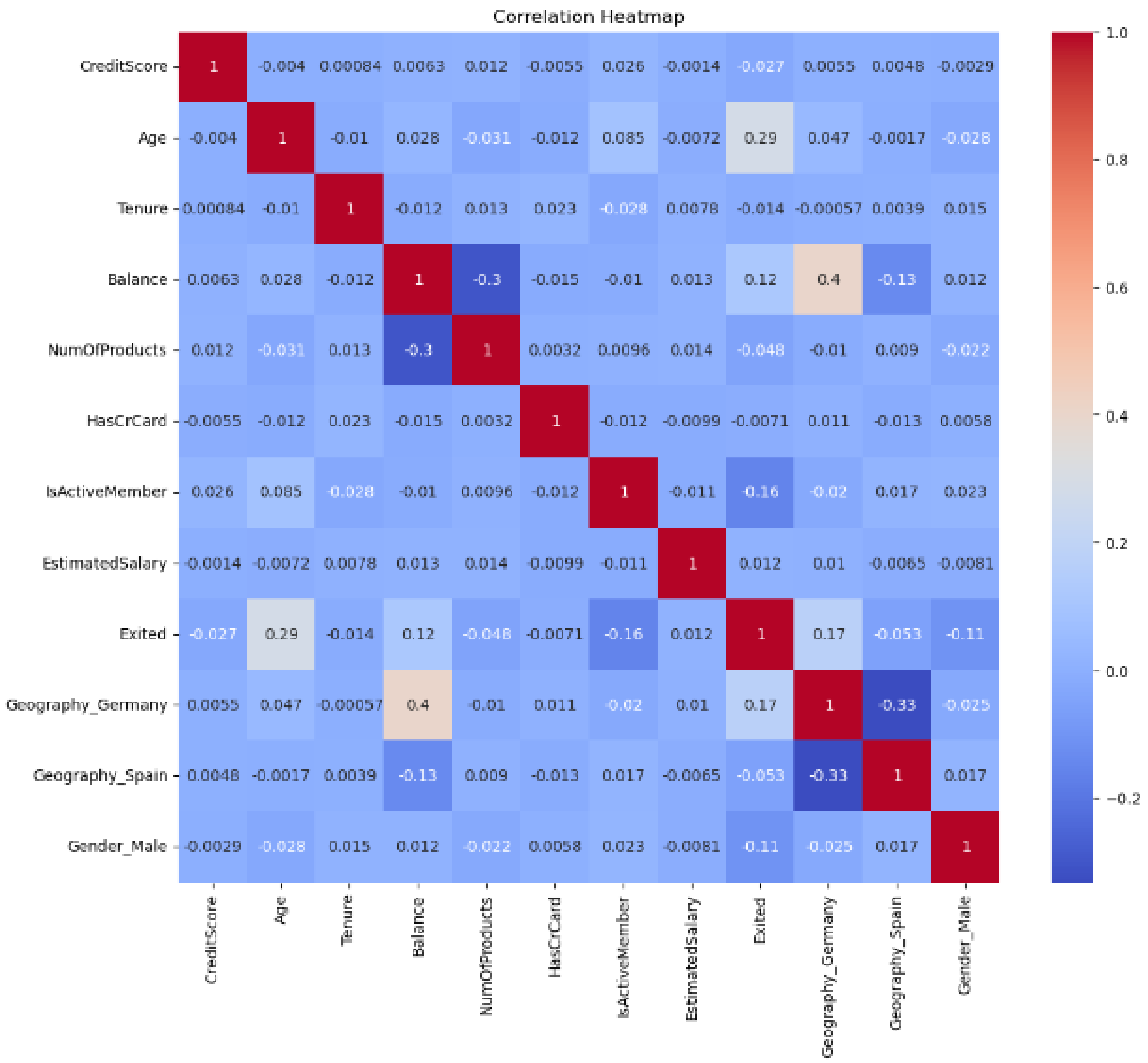
```
RowNumber          0
CustomerId         0
Surname            0
CreditScore        0
Geography          0
Gender             0
Age                0
Tenure             0
Balance            0
NumOfProducts      0
HasCrCard          0
IsActiveMember     0
EstimatedSalary    0
Exited             0
dtype: int64
```

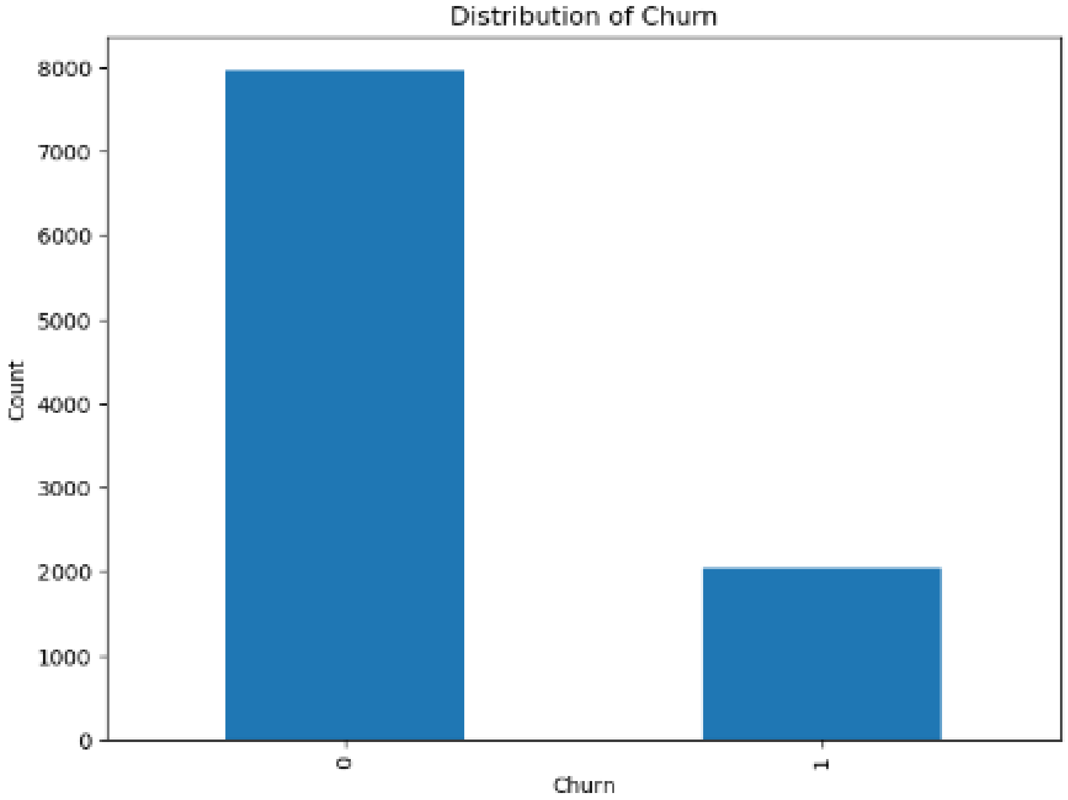
```
Available columns:
Index(['CreditScore', 'Geography', 'Gender', 'Age', 'Tenure', 'Balance',
       'NumOfProducts', 'HasCrCard', 'IsActiveMember', 'EstimatedSalary',
       'Exited'],
      dtype='object')
```

Cleaning and Digging Deep into Data

- Data Preparation & Exploratory Data Analysis

- **Data Cleaning:** Removed unnecessary information like customer IDs.
- **Exploratory Analysis**: Patterns discovered in the data.
- **Key Findings**:
    - Older customers with larger balances are more likely to leave.
    - Customers from Germany leave more often.
    - Customers with multiple products are less likely to leave


Correlation Heatmap
CreditScore
Age
Tenure
Balance
NumOfProducts
HasCrCard
IsActiveMember
EstimatedSalary
Exited
Geography_Germany
Geography_Spain
Gender_Male
1 -0.004 0.00084 0.0063 0.012 -0.0055 0.026 -0.0014 -0.027 0.0055 0.0048 -0.0029
-0.004 1 -0.01 0.028 -0.031 -0.012 0.085 -0.0072 0.29 0.047 -0.0017 -0.028
0.00084 -0.01 1 -0.012 0.013 0.023 -0.028 0.0078 -0.014 -0.00057 0.0039 0.015
0.0063 0.028 -0.012 1 -0.3 -0.015 -0.01 0.013 0.12 0.4 -0.13 0.012
0.012 -0.031 0.013 -0.3 1 0.0032 0.0096 0.014 -0.048 -0.01 0.009 -0.022
-0.0055 -0.012 0.023 -0.015 0.0032 1 -0.012 -0.0099 -0.0071 0.011 -0.013 0.0058
0.026 0.085 -0.028 -0.01 0.0096 -0.012 1 -0.011 -0.16 -0.02 0.017 0.023
-0.0014 -0.0072 0.0078 0.013 0.014 -0.0099 -0.011 1 0.012 0.01 -0.0065 -0.0081
-0.027 0.29 -0.014 0.12 -0.048 -0.0071 -0.16 0.012 1 0.17 -0.053 -0.11
0.0055 0.047 -0.00057 0.4 -0.01 0.011 -0.02 0.01 0.17 1 -0.33 -0.025
0.0048 -0.0017 0.0039 -0.13 0.009 -0.013 0.017 -0.0065 -0.053 -0.33 1 0.017
-0.0029 -0.028 0.015 0.012 -0.022 0.0058 0.023 -0.0081 -0.11 -0.025 0.017 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Distribution of Churn
Count
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
Churn

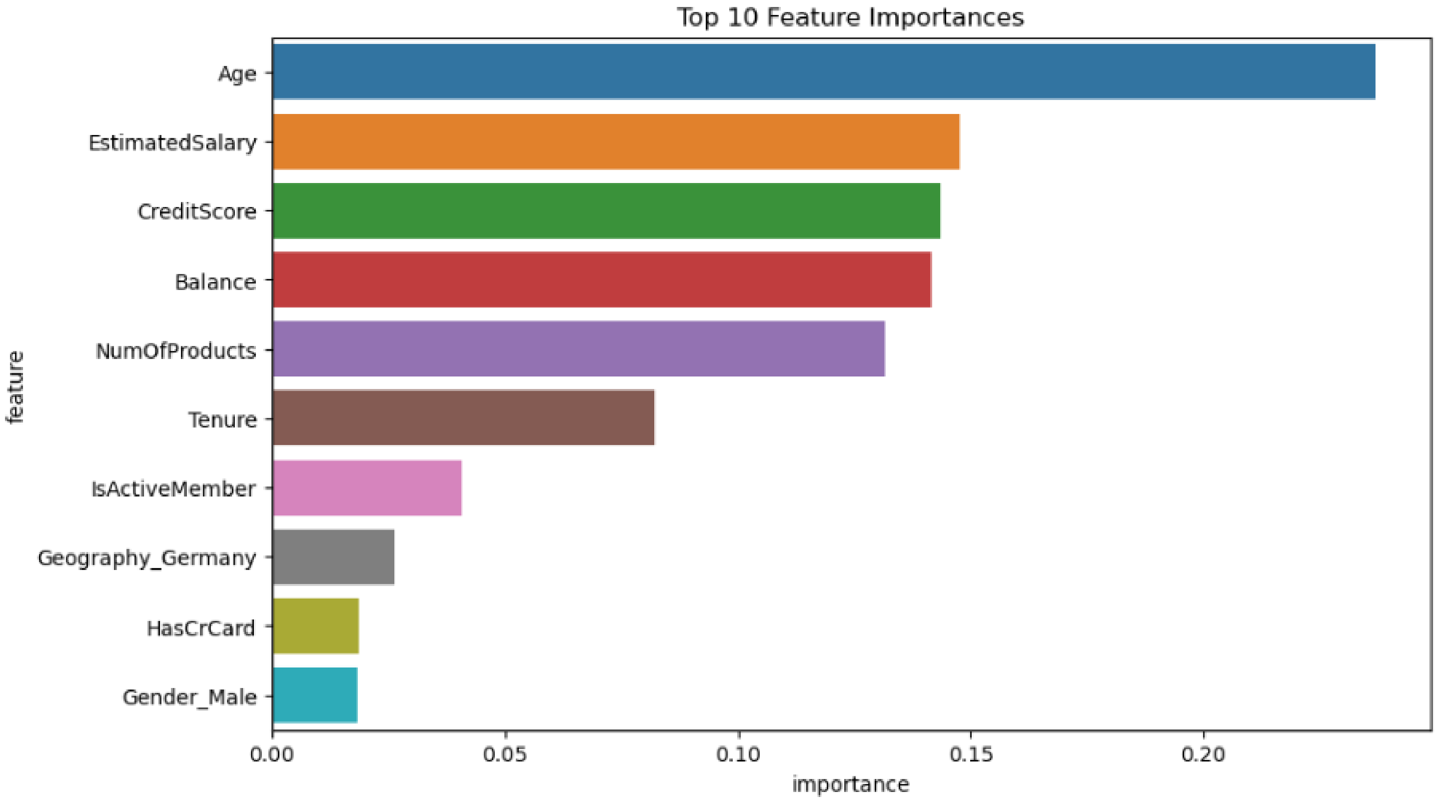
Building the Predictive Systems

- Modeling

- Developed two models:
  - **Logistic Regression**: A straightforward yes/no prediction.
  - **Random Forest**: A more complex decision-making model.
- Goal: Predict if a customer will stay or leave.

```
Logistic Regression Results:
Accuracy: 0.811

Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.96      0.89      1607
           1       0.55      0.20      0.29       393

    accuracy                           0.81      2000
   macro avg       0.69      0.58      0.59      2000
weighted avg       0.78      0.81      0.77      2000
```

```
Random Forest Results:
Accuracy: 0.8665

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.96      0.92      1607
           1       0.76      0.47      0.58       393

    accuracy                           0.87      2000
   macro avg       0.82      0.72      0.75      2000
weighted avg       0.86      0.87      0.85      2000
```



How Well Do Our Models Perform?

- Evaluation

- Accuracy: Both models predicted correctly about 86% of the time.
- Random Forest performed slightly better at identifying potential leavers.
- Key Predictors: Customer age, balance, and location.

What Did We Learn & What Can We Do?

- Summary and Recommendations

- We can predict customer churn accurately.
- **Focus Areas**:
  - Older customers and those with high balances need extra perks.
  - Improve customer experience for German customers.
  - Encourage multi-product usage to boost retention.
- **Next Steps**: Can Continue improving the model and apply targeted retention strategies.

Thank You