Customer Churn Prediction Using Machine Learning

Individual Assignment #1 | Kaylyn Nguyen

Introduction

- **Goal:** Build predictive models to identify customers likely to churn.
- **Tools:** Python, Google Colab, scikit-learn, seaborn, matplotlib
- **Dataset:** Bank Churn Modeling (Kaggle)

Churn_Modelling.csv (684.86 kB)

Detail Compact Column 14 of 14 columns

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age
1 10000	15.6m 15.8m	2932 unique values	350 850		Male 55% Female 45%	18
1	15634602	Hargrave	619	France	Female	42
2	15647311	Hill	608	Spain	Female	41
3	15619304	Onio	502	France	Female	42
4	15701354	Boni	699	France	Female	39
5	15737888	Mitchell	850	Spain	Female	43
6	15574012	Chu	645	Spain	Male	44
7	15592531	Bartlett	822	France	Male	50
8	15656148	Obinna	376	Germany	Female	29
9	15792365	He	501	France	Male	44
10	15592389	H?	684	France	Male	27
11	15767821	Bearce	528	France	Male	31
12	15737173	Andrews	497	Spain	Male	24
13	15632264	Kay	476	France	Female	34
14	15691483	Chin	549	France	Female	25

Business Questions

01

How can the bank predict customer churn before it happens?

02

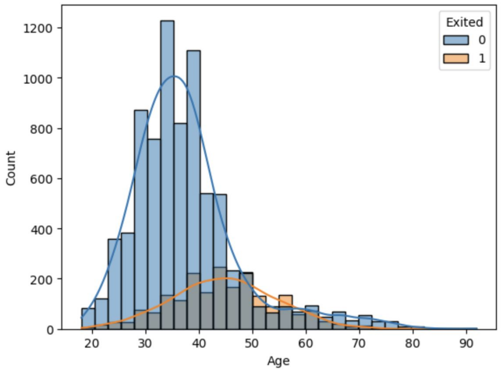
Which customers are most likely to leave the bank?

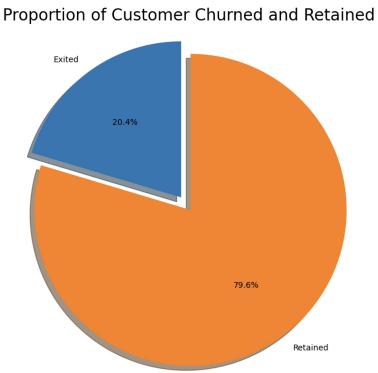
03

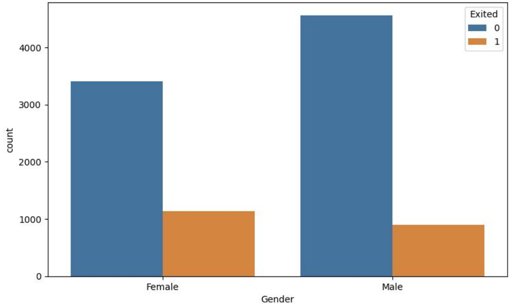
How can we use ML models to help target retention efforts?

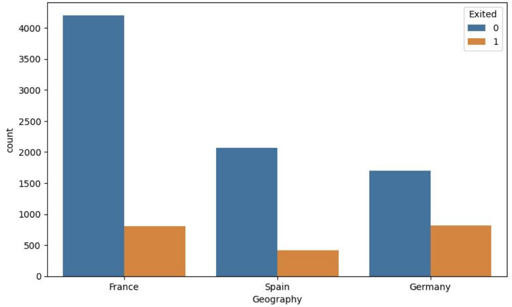
Data & EDA

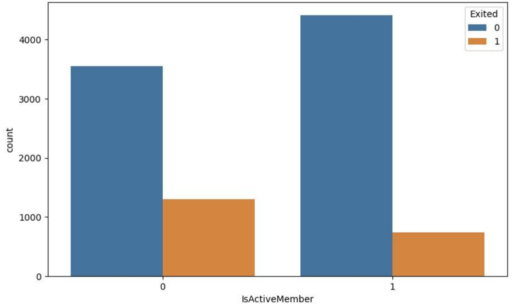
- 10,000 customer records with features like Age, Gender, HasCrCard, IsActiveMember, etc.
- **Target variable:**
  - Exited (1 = churned, 0 = stayed)
- **Key findings:**
  - Churn rate ~20%
  - Higher churn in Germany
  - Churn increases with age
  - Less active members more likely to churn
  - Females have a higher churn rate

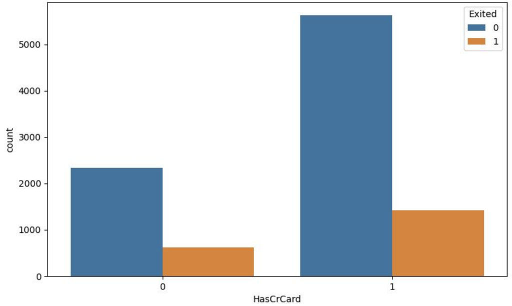






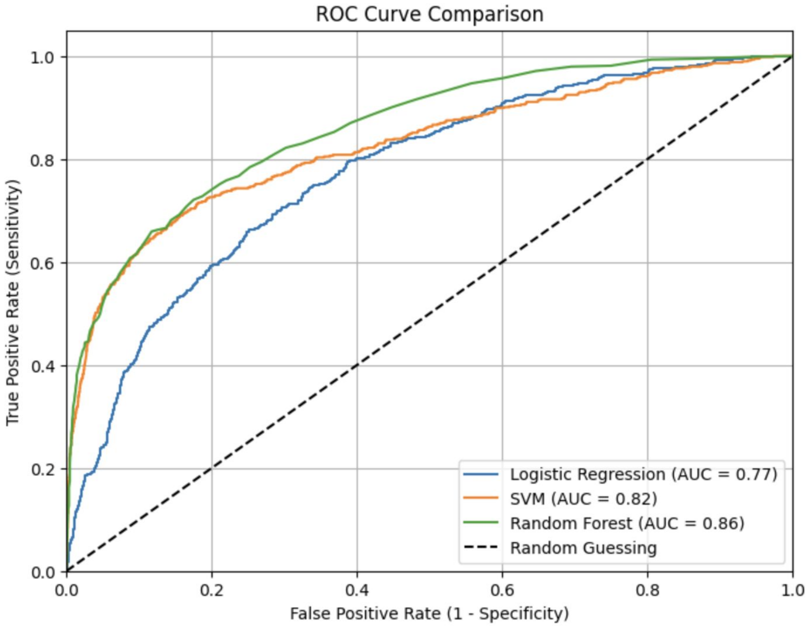






Insights (Predictive Analytics)

- **Applied 3 models:** Logistic Regression, SVM, Random Forest
- Evaluated using Confusion Matrix, Precision, Recall, ROC AUC



Model Performance Overview:

- Random Forest:
  - Highest accuracy: 87%
  - ROC AUC: 0.86
  - Recall: 48% (captures nearly half of churners)
- Logistic Regression:
  - Accuracy: 81%
  - Recall: 20%
  - Missed churners: 468 (high risk of revenue loss due to false negatives)
- Support Vector Machine (SVM):
  - Accuracy: 86%
  - Recall: 38% (misses 62% of churners)
  - Lower false positives compared to Random Forest

Takeaway:

- Prioritizing detection of churners is crucial for customer retention strategies.
- Random Forest is the optimal choice due to its ability to identify more at-risk customers, despite having 81 false positives.

Business Action

01

Built machine learning models (especially Random Forest) to predict which customers are at high risk of churning, based on their behavior and attributes (age, geography, activity, etc.).

02

Used feature importance + EDA to identify high-risk groups (e.g., older customers, less active members, those from Germany).

03

Used model predictions to flag high-risk customers, segment customers for retention campaigns, reach out via personalized offers or loyalty incentives

Business Outcome

01

The Random Forest model correctly identified ~48% of churners, meaning the bank now has a tool to catch nearly half of the potential leavers before it happens.

02

By acting on model predictions, the bank can focus marketing efforts, reduce churn-related revenue loss, and improve customer lifetime value—instead of spending resources on customers unlikely to churn.

Key Takeaways & Next Steps

- **Key Takeaways:**
  - Predictive modeling is effective for churn detection
  - Random Forest = best model for balancing accuracy and recall
- **Next:** Deploy model for real-time churn alerts

Thank You