Digital Skill Fair 32 - Data Science

Bank Churn Prediction Using Machine Learning

Latifatuzikra Suhairi

Project Overview

Background

In the highly competitive banking industry, the challenge of customer retention has become increasingly critical, as losing customers directly impacts sustainable growth and profitability.

Goals

To develop a predictive model to identify customers likely to churn.

Methodology

Utilized machine learning technique: Random Forest Classifier vs Gradient Boosting Classifier.

Dataset

Bank Customer Churn from Kaggle

Project Phase

01.

Data PreProcessing

Preparing the data for analysis (if necessery)

02.

EDA

Perform in depth analysis to extract valuable insight

03.

Feature Engineering

Create or modify the feature of data to improve model performance

04.

Model & Evaluation

Applied machine learning algorithms to build predictive model and evaluate using metrics evaluation

Data Overview

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Card Type	Point Earned
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1	1	2	DIAMOND	464
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	1	3	DIAMOND	456
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1	1	3	DIAMOND	377
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0	0	5	GOLD	350
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	0	5	GOLD	425
5	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1	1	5	DIAMOND	484
6	7	15592531	Bartlett	822	France	Male	50	7	0.00	2	1	1	10062.80	0	0	2	SILVER	206
7	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1	1	2	DIAMOND	282
8	9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.50	0	0	3	GOLD	251
9	10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0	0	3	GOLD	342

10.000 rows

18 features

Data Pre-processing

Feature Selection

#	Column
0	CreditScore
1	Geography
2	Gender
3	Age
4	Tenure
5	Balance
6	NumOfProducts
7	HasCrCard
8	IsActiveMember
9	EstimatedSalary
10	Exited
11	Complain
12	Satisfaction Score
13	Card Type
14	Point Earned

3 features were removed as they were not needed for the analysis: Surname, Customer ID, and RowNumber

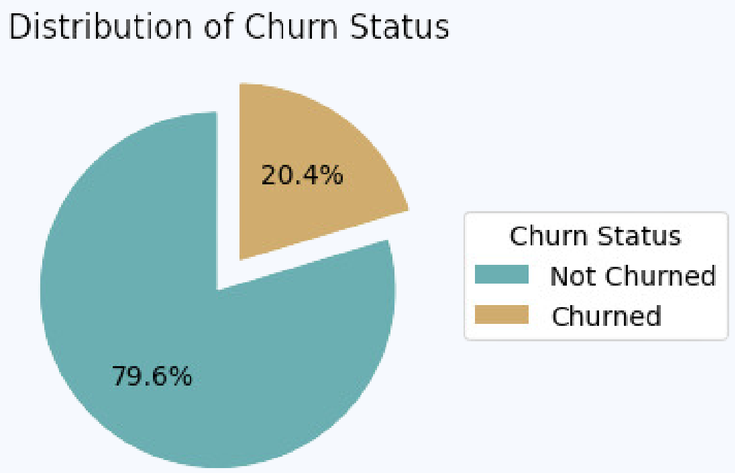
Check Missing Value

	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0
Complain	0
Satisfaction Score	0
Card Type	0
Point Earned	0

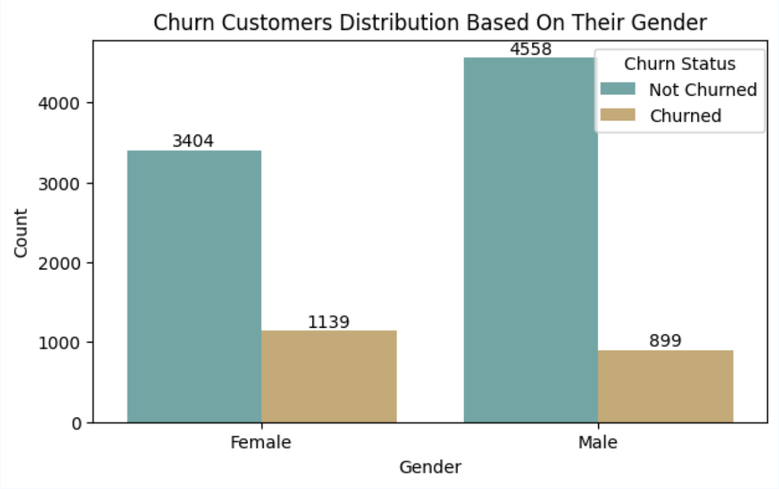
dtype: int64

The data was determined to be clean, no missing values

Exploratory Data Analysis

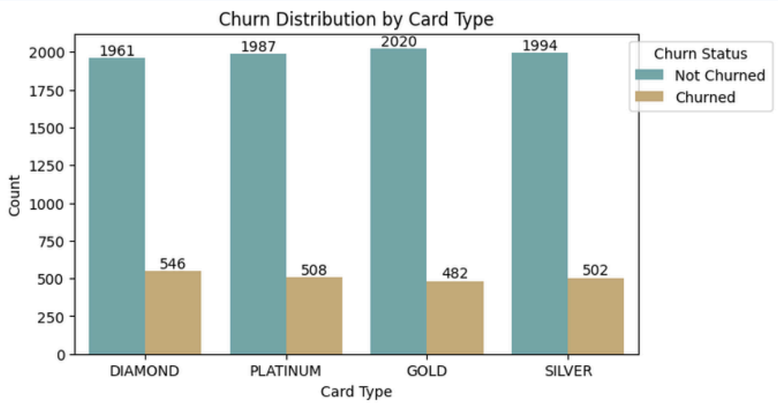


The pie chart indicates that 20.4% of customers have churned from the bank, while a significant 79.6% remain loyal. This highlights a notable retention rate, suggesting that the majority of customers are satisfied with their banking experience.

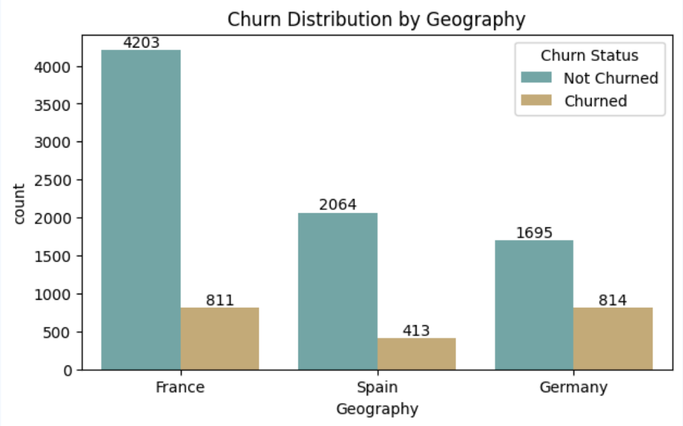


The chart shows that while both genders have a higher number of "Not Churned" customers, female customers exhibit a relatively higher churn rate compared to male customers, indicating a larger proportion of female customers are leaving the bank.

Exploratory Data Analysis

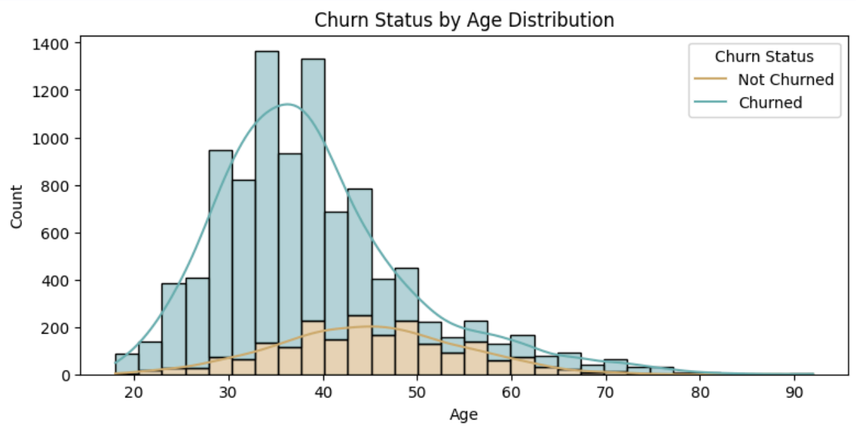


Overall, the churn rate across card types is relatively similar, ranging from 19-22%. This minimal difference suggests that a comprehensive retention strategy across all card types may be more effective than focusing on any single card type.

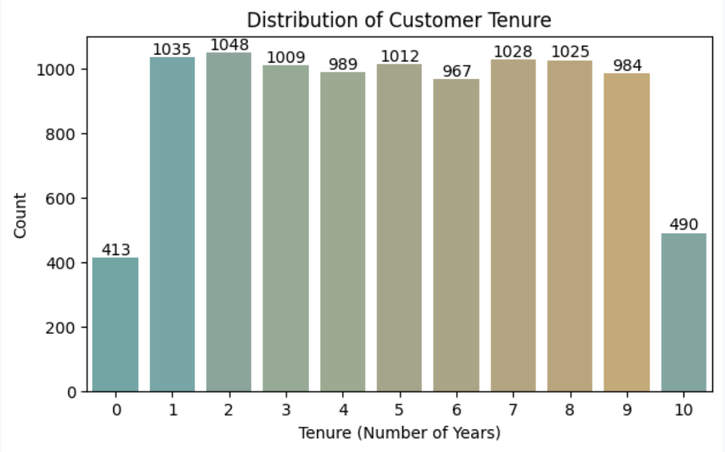


The chart shows that customers in France have the highest churn and retention numbers, while Germany and Spain show lower customer counts with similarly reduced churn rates.

Exploratory Data Analysis

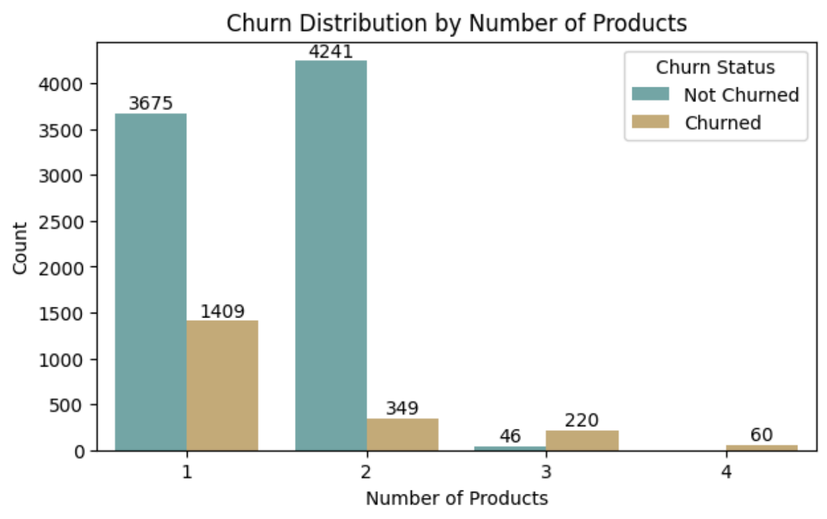


The distribution shows that churn rates are highest among customers in their 30s and 40s, while non-churned customers are concentrated in the >60 age range, suggesting that younger to middle-aged customers are more likely to leave, whereas older customers are more likely to stay

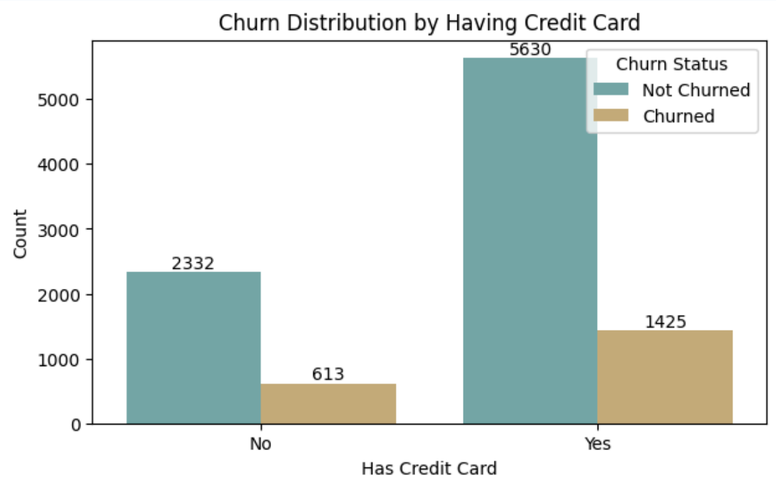


The distribution shows that customer tenure is relatively stable between 1 to 8 years, with a noticeable drop at 0 and 10 years, indicating fewer new and very long-term customers.

Exploratory Data Analysis

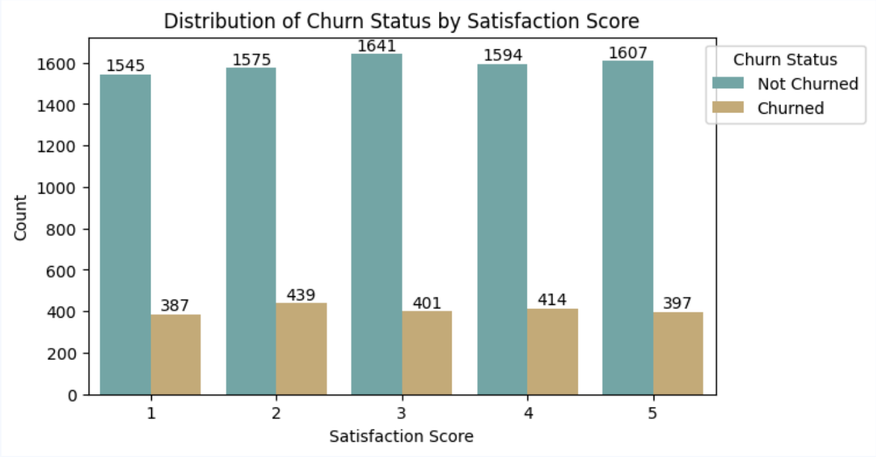


The churn rate is notably higher for customers with three or more products, especially reaching 100% for four products, indicating a need for focused retention strategies for customers with multiple products.

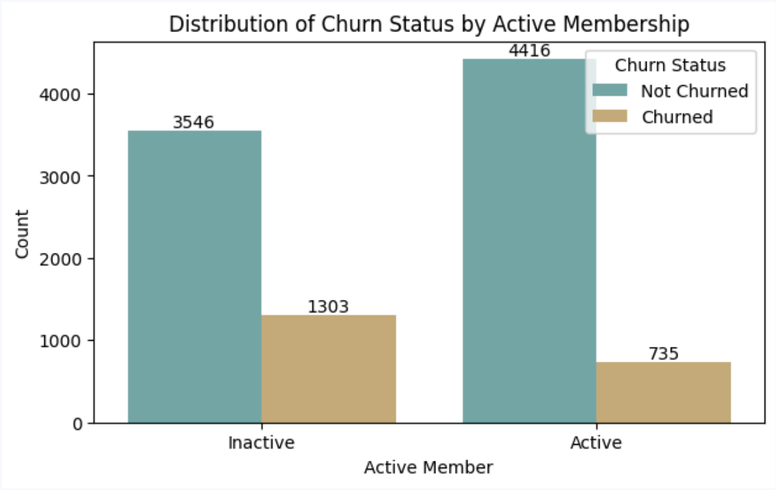


The churn rate for customers without a credit card is 20.8% (613 out of 2945), while it's slightly lower at 20.2% (1425 out of 7055) for those with a credit card. This minimal difference suggests that having a credit card does not significantly impact churn, indicating that other factors may be more influential in customer retention.

Exploratory Data Analysis

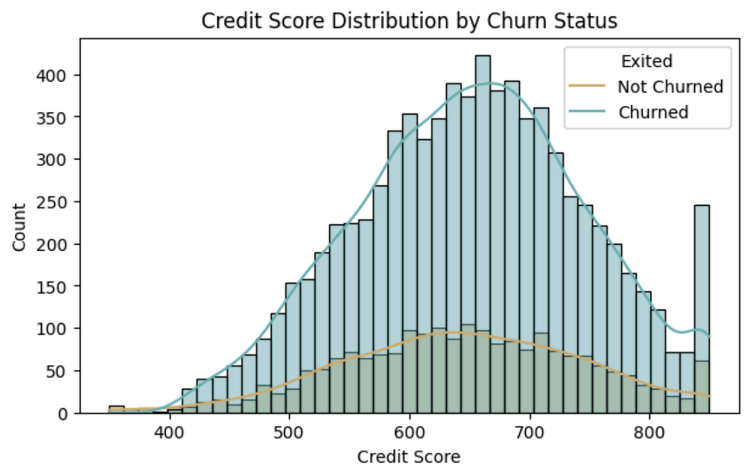


The churn rate remains relatively stable across all satisfaction scores, indicating that satisfaction score alone may not be a strong predictor of churn, as a similar proportion of customers churn regardless of their satisfaction level.

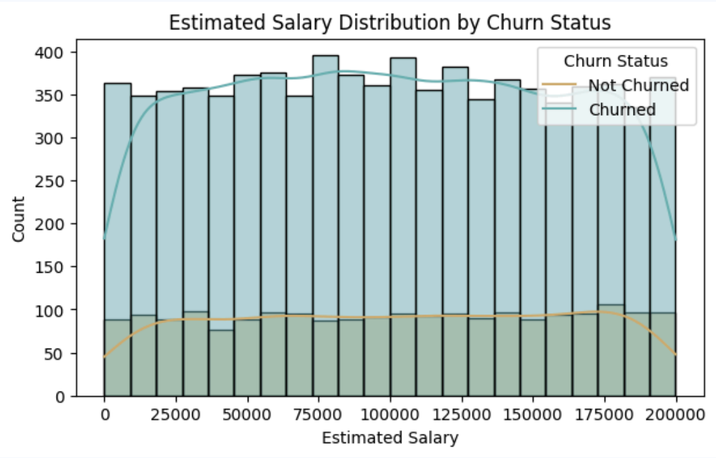


The chart show that an active members tend to have a higher likelihood of staying (not churning) compared to inactive members.

Exploratory Data Analysis

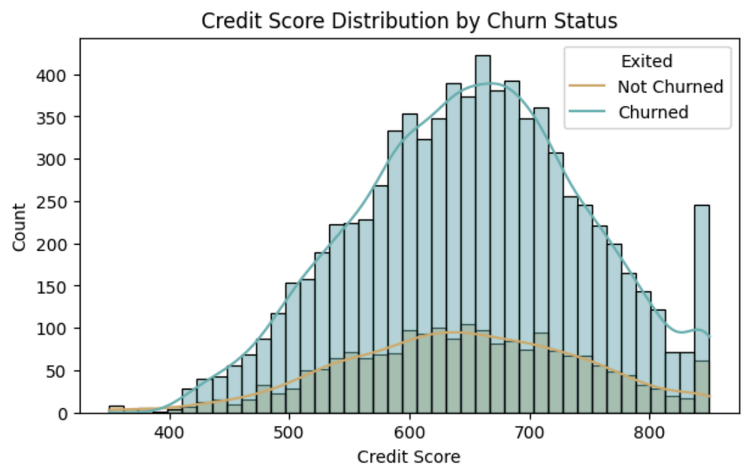


The chart show that customers with lower credit scores (400-500 range) show a higher tendency to churn, indicating a likelihood for churn among this group. Conversely, customers with higher credit scores (700-800 range) exhibit a greater tendency to stay with the service, suggesting higher retention rates among this cohort.

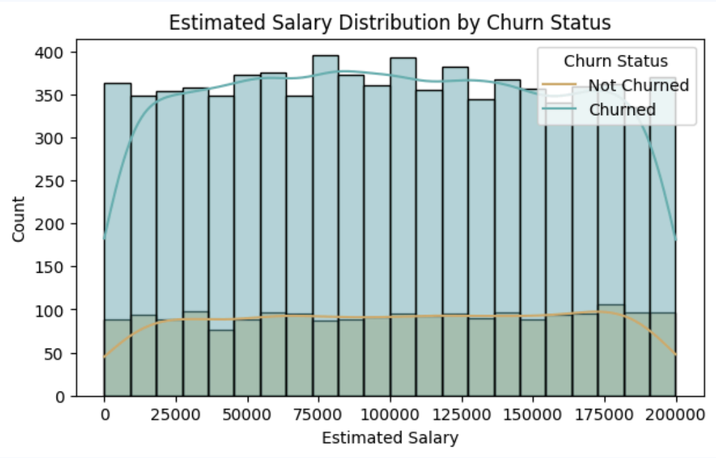


The chart shows that estimated salary does not have a significant impact on whether a customer churns or not. This finding implies that factors other than salary may play a more crucial role in determining customer retention and loyalty.

Exploratory Data Analysis



The chart show that customers with lower credit scores (400-500 range) show a higher tendency to churn, indicating a likelihood for churn among this group. Conversely, customers with higher credit scores (700-800 range) exhibit a greater tendency to stay with the service, suggesting higher retention rates among this cohort.



The chart shows that estimated salary does not have a significant impact on whether a customer churns or not. This finding implies that factors other than salary may play a more crucial role in determining customer retention and loyalty.

Feature Engineering

01. One Hot Encoding

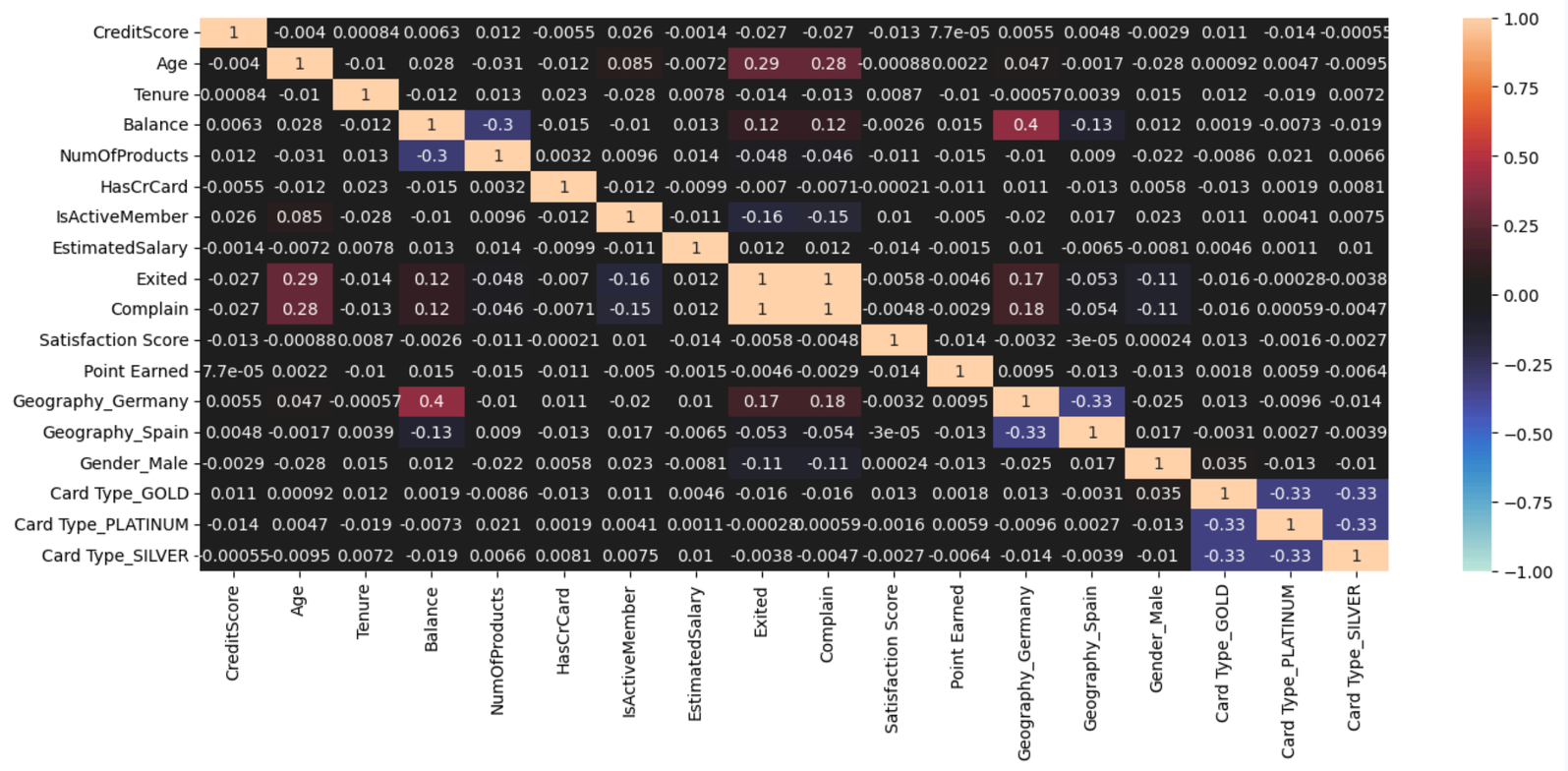
	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Card Type	Point Earned
0	619	France	Female	42	2	0.00	1	1	1	101348.88	1	1	2	DIAMOND	464
1	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0	1	3	DIAMOND	456
2	502	France	Female	42	8	159660.80	3	1	0	113931.57	1	1	3	DIAMOND	377
3	699	France	Female	39	1	0.00	2	0	0	93826.63	0	0	5	GOLD	350
4	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0	0	5	GOLD	425

CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Point Earned	Geography_Germany	Geography_Spain	Gender_Male	Card Type_GOLD	Card Type_PLATINUM	Card Type_SILVER
619	42	2	0.00	1	1	1	101348.88	1	1	2	464	False	False	False	False	False	False
608	41	1	83807.86	1	0	1	112542.58	0	1	3	456	False	True	False	False	False	False
502	42	8	159660.80	3	1	0	113931.57	1	1	3	377	False	False	False	False	False	False
699	39	1	0.00	2	0	0	93826.63	0	0	5	350	False	False	False	True	False	False
850	43	2	125510.82	1	1	1	79084.10	0	0	5	425	False	True	False	True	False	False

The three categorical features—**Geography, Gender, and Card Type**—need to be converted into numerical format for processing by machine learning techniques. This transformation will utilize **one-hot-encoding**, which is well-suited for these nominal, non-ordinal data types.

Feature Engineering

02 Feature Selection By Their Correlation



- This matrix show **correlation between features** in the dataset.
- Result: The factors with the most notable **associations with Exited (churn) are Age, Geography_Germany, and Complain**, while **other** variables show **weak relationships** that may still contribute to understanding churn when combined with other factors.

Feature Engineering

03 Feature Scaling

CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Point Earned	Geography_Germany	Geography_Spain	Gender_Male	Card Type_GOLD	Card Type_PLATINUM	Card Type_SILVER
619	42	2	0.00	1	1	1	101348.88	1	1	2	464	False	False	False	False	False	False
608	41	1	83807.86	1	0	1	112542.58	0	1	3	456	False	True	False	False	False	False
502	42	8	159660.80	3	1	0	113931.57	1	1	3	377	False	False	False	False	False	False
699	39	1	0.00	2	0	0	93826.63	0	0	5	350	False	False	False	True	False	False
850	43	2	125510.82	1	1	1	79084.10	0	0	5	425	False	True	False	True	False	False

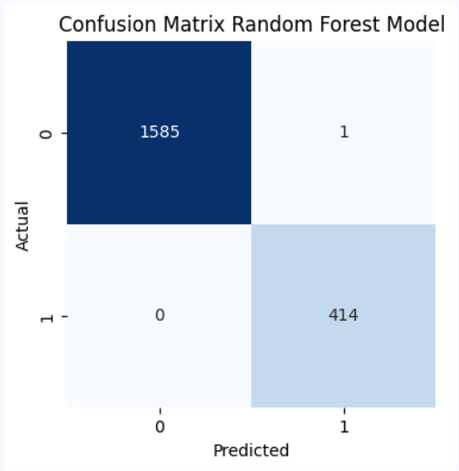
	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Complain	Satisfaction Score	Point Earned	Geography_Germany	Geography_Spain	Gender_Male	Card Type_GOLD	Card Type_PLATINUM	Card Type_SILVER
0	-0.326221	0.293517	-1.041760	-1.225848	-0.911583	0.646092	0.970243	0.021886	1	1.972908	-0.721130	-0.630839	False	False	False	False	False	False
1	-0.440036	0.198164	-1.387538	0.117350	-0.911583	-1.547768	0.970243	0.216534	0	1.972908	-0.009816	-0.666251	False	True	False	False	False	False
2	-1.536794	0.293517	1.032908	1.333053	2.527057	0.646092	-1.030670	0.240687	1	1.972908	-0.009816	-1.015942	False	False	False	False	False	False
3	0.501521	0.007457	-1.387538	-1.225848	0.807737	-1.547768	-1.030670	-0.108918	0	-0.506866	1.412812	-1.135457	False	False	False	True	False	False
4	2.063884	0.388871	-1.041760	0.785728	-0.911583	0.646092	0.970243	-0.365276	0	-0.506866	1.412812	-0.803472	False	True	False	True	False	False

- Feature scaling is use to improeve the performance and convergence of machine learning algorithms, particularly those sensitive to the scale of the data, such as gradient descent-based methods.
- This project employs the **Standard Scaler technique**, applying it exclusively to numerical columns.

Modelling & Evaluation

01 Random Forest Classifier

```
# Train Models
rf_model = RandomForestClassifier(random_state=123)
rf_model.fit(X_train, y_train)
rf_predictions = rf_model.predict(X_test)
```

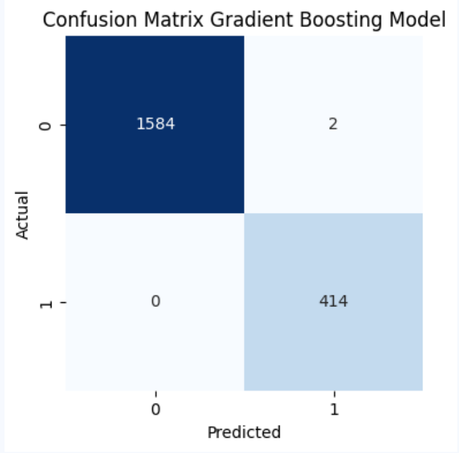



```
Accuracy: 0.9995
F1-Score 0.9987937273823885
```

- The confusion matrix show that there is only one misclassification overall, while all other data points are correctly classified. This indicates a strong performance of the model in accurately predicting the classes.
- **Random Forest Classifier** model **accuracy is 99.95%** and **F1-Score is 0.999** indicates a highly efficient model.

02 Gradient Boosting Classifier

```
# Train GBC Model
gb_model = GradientBoostingClassifier(random_state=123)
gb_model.fit(X_train, y_train)
gb_predictions = gb_model.predict(X_test)
```




```
Accuracy: 0.999
F1-Score 0.9975903614457832
```

- The confusion matrix show that there is two misclassification overall, while all other data points are correctly classified. This indicates a strong performance of the model in accurately predicting the classes.
- **Gradient Boosting Classifier** model **accuracy is 99.90%** and **F1-Score is 0.998** indicates a highly efficient model.

Conclusion

- In this notebook, we analyzed customer churn in a bank, identifying key factors like age, geography, and complaint frequency.
- Our predictive churn models showed that the Random Forest model outperformed Gradient Boosting, achieving 99.95% accuracy and a 0.999 F1 score, compared to 99.97% accuracy and a 0.998 F1 score for Gradient Boosting.

Recommendation:

- **Target high-risk customers by using insights** from the analysis to implement personalized retention strategies based on factors like age, geography, and complaint frequency.
- **Enhance customer support** by training staff to resolve issues effectively and proactively following up with customers who have previously filed complaints.
- **Offer personalized incentives to at-risk customers,** such as discounts or exclusive promotions, to make them feel valued and reduce churn likelihood.

Bank Churn Prediction Using Machine Learning

THANK YOU

Latifatuzikra Suhairi

Connect On:

latifatuzikra.suhairi@gmail.com

www.linkedin.com/in/latifatuzikra-suhairi