Supervised Learning with feature Engineering Report

Shubham Idekar
College of Engineering
Northeastern University
Boston, MA
idekar.s@northeastern.edu

Abstract

This study embarks on a journey into the realm of supervised learning with a specific focus on churn prediction in the banking sector. The dataset under scrutiny is a comprehensive collection of customer information obtained from Kaggle, with a primary goal of understanding customer churn and identifying the factors that influence it.

1 Dataset

About dataset -

- RowNumber —corresponds to the record (row) number and has no effect on the output.
- CustomerId —contains random values and has no effect on customer leaving the bank.
- Surname —the surname of a customer has no impact on their decision to leave the bank.
- CreditScore —can have an effect on customer churn, since a customer with a higher credit score is less likely to leave the bank.
- Geography —a customer's location can affect their decision to leave the bank.
- Gender —it's interesting to explore whether gender plays a role in a customer leaving the bank.
- Age—this is certainly relevant, since older customers are less likely to leave their bank than younger ones.
- Tenure—refers to the number of years that the customer has been a client of the bank. Normally, older clients are more loyal and less likely to leave a bank.
- Balance—also a very good indicator of customer churn, as people with a higher balance in their accounts are less likely to leave the bank compared to those with lower balances.
- NumOfProducts—refers to the number of products that a customer has purchased through the bank.
- HasCrCard—denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
- IsActiveMember—active customers are less likely to leave the bank.
- EstimatedSalary—as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
- Exited—whether or not the customer left the bank.
- Complain—customer has complaint or not.
- Satisfaction Score—Score provided by the customer for their complaint resolution.
- Card Type—type of card hold by the customer.
- Points Earned—the points earned by the customer for using credit card.

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9995	9996	15606229	Obijiaku	771	France	Male	39	5	0.00	2	1	0	96270.64	0	0	1	DIAMOND	300
9996	9997	15569892	Johnstone	516	France	Male	35	10	57369.61	1	1	1	101699.77	0	0	5	PLATINUM	771
9997	9998	15584532	Liu	709	France	Female	36	7	0.00	1	0	1	42085.58	1	1	3	SILVER	564
9998	9999	15682355	Sabbatini	772	Germany	Male	42	3	75075.31	2	1	0	92888.52	1	1	2	GOLD	339
9999	10000	15628319	Walker	792	France	Female	28	4	130142.79	1	1	0	38190.78	0	0	3	DIAMOND	911

Fig1:How our dataframe looks like.

2. Exploratory Data Analysis

2.1 Data Profiling

```
Dataset Shape: (10000, 18)
```

	Name	dtypes	Missing	Uniques	Sample Value	Entropy
0	RowNumber	int64	0	10000	1	4.00
1	CustomerId	int64	0	10000	15634602	4.00
2	Surname	object	0	2932	Hargrave	3.23
3	CreditScore	int64	0	460	619	2.55
4	Geography	object	0	3	France	0.45
5	Gender	object	0	2	Female	0.30
6	Age	int64	0	70	42	1.60
7	Tenure	int64	0	11	2	1.03
8	Balance	float64	0	6382	0.0	2.71
9	NumOfProducts	int64	0	4	1	0.36
10	HasCrCard	int64	0	2	1	0.26
11	IsActiveMember	int64	0	2	1	0.30
12	EstimatedSalary	float64	0	9999	101348.88	4.00
13	Complain	int64	0	2	1	0.22
14	Satisfaction Score	int64	0	5	2	0.70
15	Card Type	object	0	4	DIAMOND	0.60
16	Point Earned	int64	0	785	464	2.88
17	Churn	int64	0	2	1	0.22

Fig2.1- This is the output for our data profile

Entropy is defined as the randomness or measuring the disorder of the information being processed.

Note: We should always handle missing values carefully, this dataset does not contain any so we are good to go for the next step.

Action:
a) Remove RowNumber, CustomerID and Surname as it is unique for every row and wont be used in prediction.
b) Use label encoder for columns having 2 unique values, except for Gender every other such column is already encoded.

2.2 Churn Rate

```
Customer Churn Rate(%):
0    79.62
1    20.38
Name: Churn, dtype: float64
```

Fig2.2- Churn Rate where 0 is not exited and 1 is exited

This tells us that 20% of the Customers are likely to churn which is really a big loss for a Bank to lose so much of their customers. We can further check what kind of customers are exiting and analyze.

3 Analyzing and Preprocessing

3.1 Identifying correlation with Churn

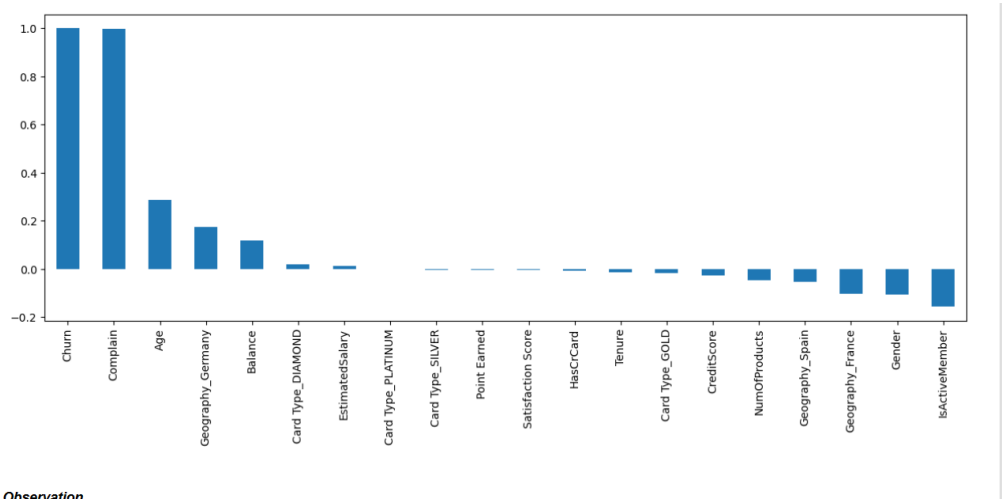

Fig 3.1- Correlation between all variables and Churn, using dummy variables for categorical data.

Observation:

- **Age** and **Estimated Salary** are **positively** correlated to Churn. Obviously, Complain has high positive correlation with Churn.
- Surprisingly **Balance** is also **positively** correlated to churn, which means even customers with higher balance are likely to exit bank.
- **CreditScore**, **NumberOfProducts** and **ActiveMember** are **negatively** related to Churn.

3.2 Analyzing Numerical Variables

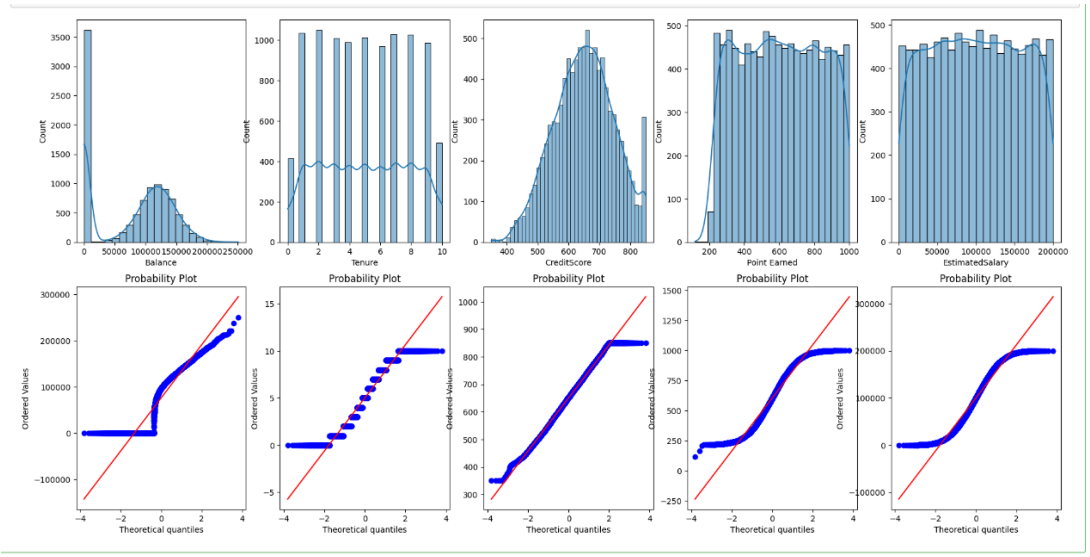

Fig3.2.1 – Probably plot and Histogram to understand the distribution of numerical variables.

Observation:

- Balance is following edge peak distribution with peak at 0.
- Tenure, Point Earned and EstimatedSalary have a uniform distribution.
- CreditScore shows left skewed distribution with tail on the right.
- These numerical variables are not following a normal distribution. These distributions indicate there are different data distributions present in population data with separate and independent peaks.

Action :

Check if the target variable represents the different distribution present in the data due to presence of different classes.
Data scaling, As most of the algorithms assume the data to be normally (Gaussian) distributed we can do one of the following:

- Normalization : Carried for features whose data does not display normal distribution.
- Standardization : Carried out for features that are normally distributed where their values are huge or very small as compared to other features.

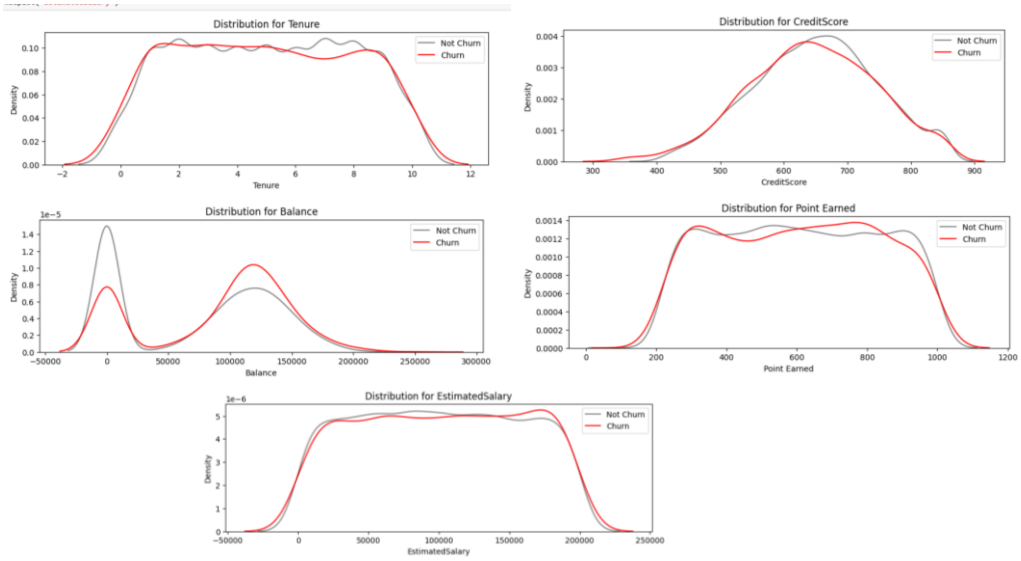

Fig3.2.2- This is Kernel Destiny Estimation plot used to understand shape of data distributions and the probability density at different points along the range of the variable.

- Churn by Tenure : Clients with 8-10 years are less likely to churn
- Churn by Balance : Surprisingly clients with balance in range 100000-150000 are more likely to exit
- Churn by CreditScore : It seems that CreditScore doesnt have much affect on churn
- Churn by Point Earned : Clients with high Points Earned are more likely to churn
- Churn by EstimatedSalary : Customers with higher Estimated Salary are more likely to churn

3.3 Analyzing Categorical Values

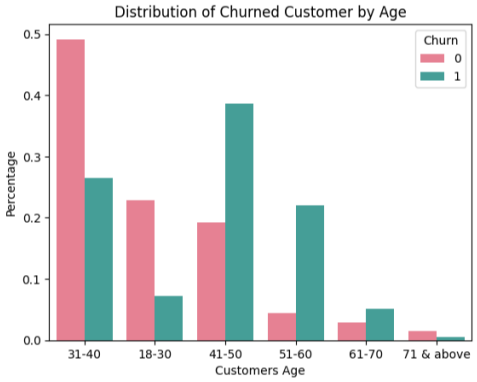

Fig3.3.1- Stacked Bar Plot to understand age with churn.

- Churn by Age : Older customers are more likely to churn rather than younger customers.

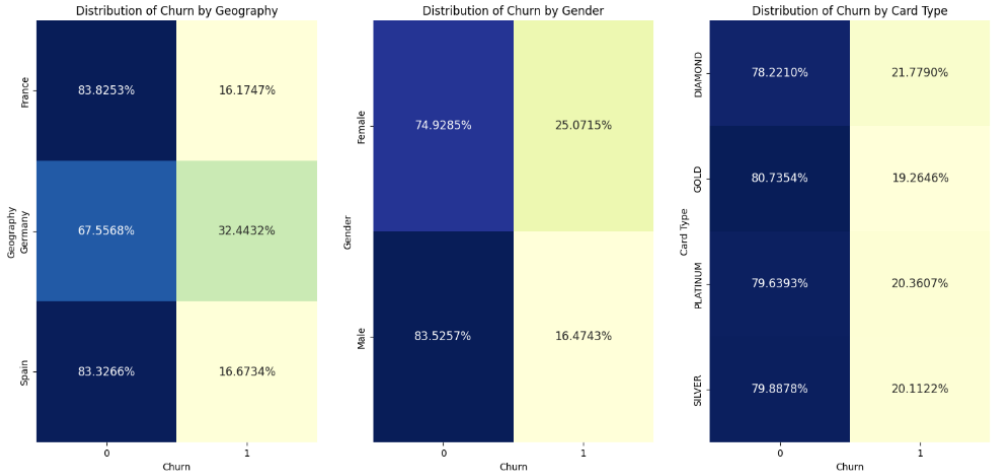

Fig3.3.2- Seaborn Heatmaps to analyze Categorical variables by churn.

- Churn by Geography : Customers in Germany have the higher churn rate compared to Spain & France
- Churn by Gender : Female customers are more likely to exit
- Churn by Card Type : Card type individually doesn't have much effect on Churn Rate

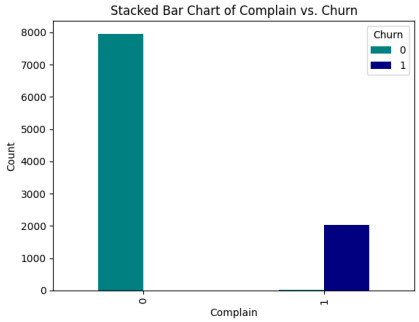

Fig3.3.3- Stacked bar chart for Complain against churn

- Churn by Complain : We can see that almost all the customers with complains have exited

Fig3.3.4- Exploding Pie Chart for Satisfaction Score

- Churn by Satisfaction Score: All these values are biased and show no specific distribution

Observation :

No direct relation observed between the following features and the target variable:

Satisfaction Scorea
Card Type
CreditScore
EstimatedSalary

We can drop these columns

3.4 Multicolinearity among predictors

Fig3.4.1- Correlation Heatmap for analyzing correlation of predictors with other predictors

Observation :

- Geography_France, Geography_Spain and Geography_Germany are highly corelated to each other.
- HasCrCard has very less correlation with any of the variables.

Action: Dropping some of the highly correlated categorical variables.

3.5 OLS Regression Model

Dep. Variable:	Churn	**R-squared (uncentered):**	0.993
Model:	OLS	**Adj. R-squared (uncentered):**	0.993
Method:	Least Squares	**F-statistic:**	1.454e+05
Date:	Sun, 15 Oct 2023	**Prob (F-statistic):**	0.00
Time:	21:21:51	**Log-Likelihood:**	18700.
No. Observations:	10000	**AIC:**	-3.738e+04
Df Residuals:	9990	**BIC:**	-3.731e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Age	0.0001	2.71e-05	4.281	0.000	6.29e-05	0.000
Tenure	-0.0015	0.001	-1.226	0.220	-0.004	0.001
Balance	-8.838e-05	0.001	-0.059	0.953	-0.003	0.003
IsActiveMember	-0.0024	0.001	-3.204	0.001	-0.004	-0.001
Complain	0.9925	0.001	959.002	0.000	0.991	0.995
Point Earned	-0.0034	0.001	-2.524	0.012	-0.006	-0.001
Geography_France	0.0002	0.001	0.251	0.802	-0.001	0.002
Gender_Female	0.0002	0.001	0.294	0.769	-0.001	0.002
NumOfProducts_four	0.0056	0.005	1.153	0.249	-0.004	0.015
NumOfProducts_three	0.0046	0.002	1.914	0.056	-0.000	0.009

Omnibus:	17391.198	**Durbin-Watson:**	1.576
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	205524068.590
Skew:	-11.197	**Prob(JB):**	0.00
Kurtosis:	704.966	**Cond. No.**	530.

Fig3.5- OLS Regression model summary

Observation:

- Most of our variables have values below 0.05, and others have patterns when visualized

3.6 Data balancing

In order to cope with unbalanced data, we will be performing Oversampling, i.e. increase the minority samples of the target variable to the majority samples. For

data balancing, we will use imblearn.

The simplest approach involves duplicating examples in the minority class, although these examples don't add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short.

```
: over = SMOTE(sampling_strategy = 1)

  X = pred.values
  y = Churn.values

  X, y = over.fit_resample(X, y)
  Counter(y)

: Counter({1: 7962, 0: 7962})
```

Fig3.6- We had 20% churn data, now we have balanced data

4 Modelling

4.1 Hyperparameter tuning using Optuna

Optuna is an automatic hyperparameter optimization software framework, particularly designed for machine learning. It features an imperative, define-by-run style user API. Thanks to our define-by-run API, the code written with Optuna enjoys high modularity, and the user of Optuna can dynamically construct the search spaces for the hyperparameters.

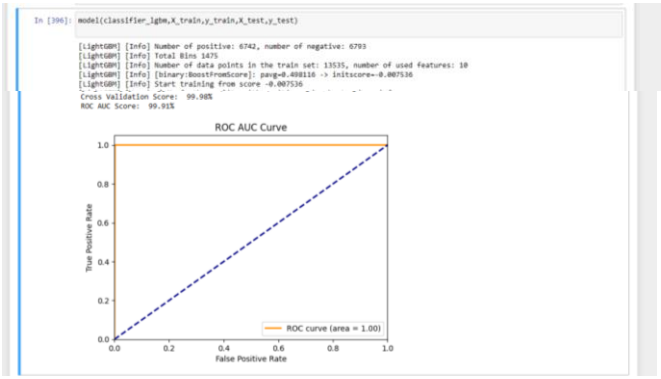

Fig4.1- ROC AUC Curve

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1169
1	1.00	1.00	1.00	1220
accuracy			1.00	2389
macro avg	1.00	1.00	1.00	2389
weighted avg	1.00	1.00	1.00	2389

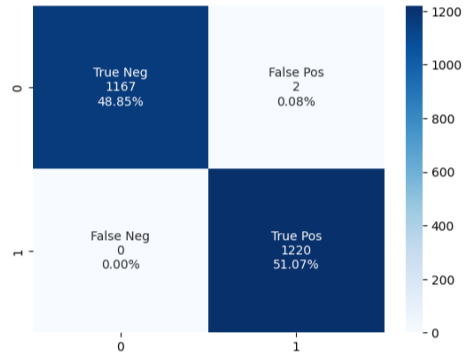

Fig3.4.2- Model Evaluation metrics

Model Evaluation

Let's understand what is meant by Precision and Recall.

- Precision measures the percentage of customers predicted by us to Churn that were correctly classified.
- Recall measures the percentage of actual customers that churned were correctly classified.

When we predict that the customers will churn, 51% of customers actually churned. Whereas, out of all the customers that churned we were able to capture 50% of all them correctly.

We can further tune our models to be more precise or have a better recall based on the business requirement.

5 Conclusions

In this report, I have performed Churn Analysis and Prediction on Bank Customer data using supervised learning with feature selection.

Important insights for business:

- Old customers must be given some benefits to stop them from churning out.
- Customers who aren't active anymore must be attended so that we bank could understand why they are going inactive which leads to exiting at the end.
- Customers with complains and after complain resolution with low Satisfaction Score tend to churn, which means their complains aren't being taken care of.
- This model might show some signs of overfitting leading to higher pseudo accuracy which can be taken care of by tuning the model further.

References

[1] https://optuna.readthedocs.io/en/stable/index.html

[2] https://seaborn.pydata.org/index.html

[3] https://scikit-learn.org/stable/

[4] https://lightgbm.readthedocs.io/en/latest/index.html