Kaggle

**Churn modeling by using dataset provided by Kaggle**

1. What is the Credit Score in this dataset and the standard how to measure it?
2. What is the balance in this dataset? And what is the meaning of this word?
3. **What is the Credit Score in this dataset and the standard how to evaluate it?**

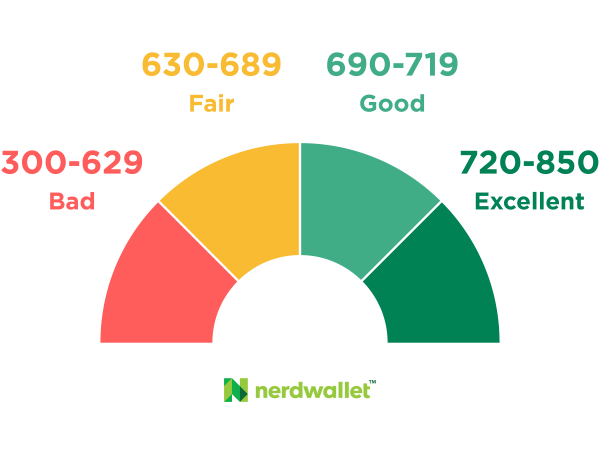


Figure 1 Credit Score Ranges

Credit Score is a three-digit number that estimates how likely you are to repay borrowed money.

A low credit score may not keep you from being approved for credit, but you may have to pay a higher interest rate or put money on deposit. And a higher credit score can give you access to more credit products – and at lower interest rates.

The information on your credit accounts is stored by credit-reporting agencies, also called credit bureaus. The three largest are Equifax, Experian and TransUnion. If you use credit, they probably have a record of it Credit-scoring companies use the information to produce credit scores, and creditors buy reports and score to evaluate applicants.

# 인용 자료

(O'SheaBev, June 23, 2020) “What is Credit Score, and what are the Credit Score Ranges?”

1. What is the balance in this dataset? And what is the meaning of this word?

In banking and accounting, the balance is the amount of money owed that remains in a deposit account. It is more likely people who have lots of money on their accounts keep paying for services than who have not. So it definitely matters in some ways.

* Simple INSIGHT from EDA that I’ve worked so far

1. 고객 데이터의 나라 점유율은 프랑스가 스페인과 독일보다 약 2배가량 높음
2. 고객의 성별의 비율은 남성이 여성보다 약 10%가량 높음
3. 주요 고객층은 20~40대로 이 층들의 비율이 전체의 약 85%를 차지하고, 그 중 30대가 45%를 차지하며 제일 높음
4. 30대의 고객 이탈률은 12%로 매우 낮고, 주요 20~40대의 이탈률 또한 40,50대의 이탈률보다 현저히 낮음
5. 프랑스, 독일, 스페인 모두에서 나이별 고객 수 대비 이탈률이 각각 52%, 70%, 46%로 50대가 제일 높음

* Derived Variables that I’ve created

1. AgeGroup : Age변수를 이용해 고객의 나이를 10의 단위로 나눠 분류한 변수
2. AgeGroup\_title : 피벗테이블 생성을 하기 위해 AgeGroup값(숫자)에 나이를 설명해주는 한글 “대”(문자)를 추가한 변수
3. Is\_balance : Balance변수를 이용해 통장잔고가 0인지 아닌지 여부에 따라 0과 1로 부여한 변수
4. Class\_CrdtScore : CreditScore변수를 이용해 신용점수 기준표를 바탕으로 등급을 Bad, Fair, Good, Excellent로 나눠 부여한 변수

* The Variables that I’ve dropped

1. CustomerId
2. Surname
3. RowNumber

* Logistic Regression

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Geography[T.Germany] | B = 0.7747 | (+)값, 독일은 프랑스에 비해 churn 위험도 증가 |
| P = 0.000 | <0.05, 통계적으로 유의 |
| Exp(B) = 2.169 | 독일은 프랑스에 비해 churn 위험도 2.1배 |
| Geography[T.Spain] | B = 0.0352 | (+)값, 스페인은 프랑스에 비해 churn 위험도 증가 |
| P = 0.618 | >0.05, 통계적으로 유의하지 않음 |
| Exp(B) = 1.035 | 스페인은 프랑스에 비해 churn 위험도 1.035배 증가 |
| Gender[T.Male] | B = -0.5285 | (-)값, 남성은 여성에 비해 churn 위험도 감소 |
| P = 0.000 | <0.05, 통계적으로 유의 |
| Exp(B) = 0.5894 | 남성은 여성에 비해 churn 위험도가 0.5배 |
| HasCrCard[T.1] | B = -0.0447 | (-)값, 카드 보유자는 아닌 자보다 churn 위험도 감소 |
| P = 0.452 | >0.05, 통계적으로 유의하지 않음 |
| Exp(B) = 0.9562 | 카드 보유자는 아닌 자보다 churn 위험도 0.9배 |
| IsActiveMemeber[T.1] | B = -1.0754 | (-)값, 활동유저가 아닌 유저보다 churn 위험도 감소 |
| P = 0.000 | <0.05, 통계적으로 유의 |
| Exp(B) = 0.3411 | 활동유저가 아닌 유저보다 churn 위험도 0.34배 |
| CreditScore | B = -0.0007 | (-)값, 신용점수가 증가할수록 churn 위험도 감소 |
| P = 0.017 | <0.05, 통계적으로 유의 |
| Exp(B) = 0.9993 | 신용점수가 1증가할 때 churn 위험도 0.9배, 감소 |
| Age | B = 0.0727 | (+)값, 나이가 증가할수록 churn 위험도 증가 |
| P = 0.000 | <0.05, 통계적으로 유희 |
| Exp(B) = 1.075 | 나이가 1증가할수록 churn위험도 1.07배, 증가 |
| Tenure | B = -0.0159 | (-)값, 기간이 증가할수록 churn 위험도 감소 |
| P = 0.088 | >0.05, 통계적으로 유의하지 않음 |
| Exp(B) = 0.9842 | 기간이 1년 증가할수록 churn 위험도 0.98배, 감소 |
| Balance | B = 2.637e-06 | (+)값, 통장잔고가 증가할수록 churn 위험도 증가 |
| P = 0.000 | <0.05, 통계적으로 유의 |
| Exp(B) = 0.000 | 통장잔고가 1원 증가할 때 churn 위험도 0배, 증가 |
| NumOfProducts | B = -0.1015 | (-)값, 상품 수가 증가할수록 churn 위험도 감소 |
| P = 0.031 | <0.05, 통계적으로 유의 |
| Exp(B) = 0.9034 | 상품이 1개 증가할 때 churn 위험도 0.9배, 감소 |
| EstimatedSalary | B = 4.807e-07 | (+)값, 연봉이 증가할수록 churn 위험도 증가 |
| P = 0.310 | >0.05, 통계적으로 유희하지 않음 |
| Exp(B) = 0.000 | 연봉이 1증가할 때 churn 위험도 0배, 증가 |

* Logistic Regression evaluation

Test accuracy : 0.812

* Feature Importance

Geography\_Germany (the highest, +)

즉, 지역이 독일인 고객의 이탈률이 더 높을 확률이 크다

IsActiveMember(the lowest, -) , Gender\_Male(the second lowest, -)

즉, 활동유저가 아닌 유저보다 이탈률이 더 낮을 확률이 크다

그리고 2번째로 값이 작은 남성이 여성보다 이탈을 할 확률이 더 낮을 것이다

But, 로지스틱 회귀모형을 시행할 때 Category로 분류해 모델을 돌리면서 그것에 해당하는 변수들에 의미를 두지 않아도 되는 것일 수도 있다 ? <Googling>

* Scaled Logistic Regression

Test accuracy : 0.799

* Feature Importance

The same as Logistic regression result

But, weird thing is a result of confusion matrix, which is FP and TP values are zero !!

How can I explain this ?

That is not a good sign to explain evaluate the model because precision and recall values are zero so that means low predicted positive rate.

* Dummied Logistic Regression

Test accuracy : 0.800

Scaled Logistic Regression과 같이 precision, recall값이 0이 되고 FP, TP값이 0이므로 불균형한 데이터. 예측하기에 좋은 분류가 아님 (모델이든 데이터의 불균형 존재 가능성 큼)

* 분류 – Decision Tree
* Final model (max\_depth =5)

Train Accuracy : 0.864

Test Accuracy : 0.846

It seems like the confusion matrix is right.

* Main Feature Importance

Age (0.424) > NumOfProducts (0.321) > IsActiveMember\_1 (0.135) > Geography\_Germany (0.048)

* 분류 – RandomForest
* Final model1 (min\_samples\_leaf = 30, min\_samples\_split = 100, max\_depth = 10 , n\_estimators = 50, random\_state -=1234)

Train Accuracy : 0.855

Test Accuracy : 0.842

* Main Feature Importance

Age (0.427) > NumOfProducts (0.194) > Balance (0.072) > IsActiveMember\_1 ( 0.067)

* 분류 – RandomForest (GridSearch)
* Final Model2 (max\_depth =10, min\_samples\_leaf = 10, min\_samples\_split =2, n\_estimators = 50, random\_state =1234)

Train Accuracy : 0.880

Test Accuracy : 0.855

* Main Feature Importance

Age (0.368) > NumOfProducts (0.203) > Balance (0.090) > CreditScore (0.055)

* GBM (Gradient Boosting Machine)
* Final Model 1 (random\_sstate =1234)

Train\_Accuracy : ?????

Test\_Accuracy : 0.8597

* GBM (Gradient Boosting Machine) ( Grid Search )
* Final Model 2 (learning\_rate = 0.05, max\_depth = 1, n\_estimators = 500)

Train Accuracy : 0.8589

Test Accuracy : ????

하이퍼파라미터를 직접 구하고 높은 예측 성능인 최종 모델 선정

* Main Feature Importance

Age (0.4) > NumOfProducts (0.4) >>> Geography\_Germany (0.06) > Is\_ActiveMember\_0 (0.05)

* 분류 - XGBoost
* Final model (max\_depth =12, eta = 0.05, objective : binary:logistic, eval\_metric : logloss, early\_stoppings =200 , num\_rounds =500)

Train logloss : 0.155

Eval logloss : 0.369

Train Accuracy : ????

Test Accuracy : 0.8510

하이퍼파라미터 조정해서 다른 최종 모델 찾기 !!!

* Main Feature Importance

EstimatedSalary(10556) > CreditScore(9833) > Balance(8034) > Age(5837) > Tenure(4440)

피처중요도가 다른 알고리즘과 상이한 값이 나옴 <Googling>

* LGBM(LightGBM)
* Final model (n\_estimators = 48 , n\_stopping\_rounds = 100, eval\_metric = ‘logloss’)

Early stopping, best iteration is :

Valid’s binary\_logloss : 0.3339

Test\_accuracy : 0.8610

* Main Feature Importance

Balance(295) > CreditScore(245) > EstimatedSalrary(241) > Age(226) > NumOfProduct(153)

* SMOTE 적용 후 LGBM(LightGBM)
* Final model(n\_estimators = 48 , n\_stopping\_rounds = 100, eval\_metric = ‘logloss’)

Early stopping best iteration is :

Valid’s binary\_logloss : 0.3394

Test\_accuracy : 0.8623

* Main Feature Importance