CAPSTONE PROJECT

VALIDATION, MONITORING AND GOVERNANCE

PREDICTING DEFAULT IN CREDIT CARD PAYMENT FOR THE NEXT MONTH

Ibiyengha Tobin
301256083

1.0 INTRODUCTION

Predicting credit card defaults is now a crucial part of risk management for banks and other financial organisations in the current financial environment. These institutions are able to minimise financial losses and ensure appropriate lending practises by being able to predict prospective credit card defaults with accuracy. Using Python modelling, I explored the world of Classification analysis in this project and developed a reliable model for predicting credit card defaults.

Predictive modelling has become a useful tool for decision-making across many industries as the world gets more data-driven. This project attempts to create a predictive model that can identify those who are more likely to default in a credit card payment the following month by utilising prior credit card data. The project's main goal is to give financial organisations a trustworthy tool for evaluating credit risk and more efficiently allocating resources. This predictive ability improves the entire costumers experience while also assisting in the management of credit card portfolios.

I explored several stages of the predictive modelling procedure during the course of this project. Involved in this are the gathering and preprocessing of data, choosing a model, developing it, and testing it. Due to its adaptability, large library of tools, and strong ecosystem for machine learning and data analysis, Python served as our programming language of choice. Any predictive modelling project must include both validation and governance to guarantee the accuracy, dependability, and moral application of the produced models.

1.1 VARIABLE LEVEL MONITORING

Variable Level Monitoring focuses on precisely observing certain metrics within a system or process. This method assures precision and early detection of problems, reduces risks, maintains quality, and allows for compliance with industry norms and laws. These financial institutions can make educated decisions, drive continuous development, allocate resources efficiently, and give transparent reporting to stakeholders by analysing selected factors.

1.2 MODEL BUILD VARIABLE LEVEL STATISTICS

It is important to know the variables in order to ensure the correctness and integrity of data used in model building. This process reveals data quality issues, improves comprehension of data distributions and linkages by analysing variables individually. These insights allow for more informed judgements, increase model interpretability, and aid in evaluating performance across several datasets. Furthermore, tracking these statistics over time detects data drift and assures regulatory compliance, lowering the risk of deploying incorrect or biased models and contributing to the model's overall success during deployment and continuing management.

The numerical variables in this dataset for this project are:

1. ID: ID of each client
2. LIMIT_BAL: Amount of given credit in New Taiwan(NT) dollars
3. AGE: Age in years
4. REPAY_SEPT : Repayment status in September, 2005
5. REPAY_AUG: Repayment status in August, 2005
6. REPAY_JULY: Repayment status in July, 2005
7. REPAY_JUNE: Repayment status in June, 2005
8. REPAY_MAY: Repayment status in May, 2005
9. REPAY_APR Repayment status in April, 2005
10. BILL_AMT_SEPT: Amount of bill statement in September, 2005 (NT dollar)
11. BILL_AMT_AUG: Amount of bill statement in August, 2005 (NT dollar)
12. BILL_AMT_JULY: Amount of bill statement in July, 2005 (NT dollar)
13. BILL_AMT_JUNE: Amount of bill statement in June, 2005 (NT dollar)
14. BILL_AMT_MAY: Amount of bill statement in May, 2005 (NT dollar)
15. BILL_AMT_APR: Amount of bill statement in April, 2005 (NT dollar)
16. PREV_AMT_SEPT: Amount of previous payment in September, 2005 (NT dollar)
17. PREV_AMT_AUG: Amount of previous payment in August, 2005 (NT dollar)
18. PREV_AMT_JULY: Amount of previous payment in July, 2005 (NT dollar)
19. PREV_AMT_JUNE: Amount of previous payment in June, 2005 (NT dollar)
20. PREV_AMT_MAY: Amount of previous payment in May, 2005 (NT dollar)
21. PREV_AMT_APR: Amount of previous payment in April, 2005 (NT dollar)
22. DEFAULT_PAY_NEXTMNTH: (1=yes, 0=no)

The Categorical Variables are:

1. GENDER: (1 = Male, 2 = Female)
2. EDUCATION (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
3. MARRIAGE: Marital status (1=married, 2=single, 3=others)

The descriptive statistics of the numerical variables excluding the target variables are:

	count	Mean	std	**min**	**25%**	50%	75%	**max**
ID	8789.0	14762.0	8734.0	4.0	7225.0	14548.0	22329.0	30000.0
LIMIT BAL	8789.0	138302.0	112129.0	10000.0	50000.0	100000.0	200000.0	510000.0
AGE	8789.0	35.0	9.0	21.0	28.0	34.0	42.0	60.0
REPAY_SEPT	8789.0	-0.0	1.0	-2.0	-1.0	0.0	0.0	1.0
REPAY AUG	8789.0	-1.0	1.0	-2.0	-1.0	0.0	0.0	1.0
REPAY JULY	8789.0	-1.0	1.0	-2.0	-2.0	0.0	0.0	0.0
REPAY JUNE	8789.0	-1.0	1.0	-2.0	-2.0	0.0	0.0	0.0
REPAY MAY	8789.0	-1.0	1.0	-2.0	-2.0	-1.0	0.0	0.0
REPAY APR	8789.0	-1.0	1.0	-2.0	-2.0	-1.0	0.0	0.0
BILL AMT SEPT	8789.0	19886.0	24916.0	-4894.0	390.0	8710.0	33634.0	156786.0
BILL AMT AUG	8789.0	17387.0	22228.0	-5174.0	181.0	5889.0	29122.0	120651.0
BILL AMT JULY	8789.0	15166.0	19648.0	-6877.0	0.0	4160.0	26602.0	96940.0
BILL AMT JUNE	8789.0	12563.0	16295.0	-7905.0	0.0	2799.0	21090.0	74329.0
BILL AMT MAY	8789.0	10763.0	14222.0	-10213.0	0.0	1848.0	19324.0	59984.0
BILL AMT APR	8789.0	9900.0	13711.0	-11610.0	0.0	1261.0	18694.0	49756.0
PREV AMT SEPT	8789.0	1498.0	1370.0	0.0	208.0	1443.0	2007.0	7662.0
PREV AMT AUG	8789.0	1338.0	1229.0	0.0	0.0	1300.0	2000.0	6336.0
PREV AMT JULY	8789.0	1007.0	1023.0	0.0	0.0	933.0	1583.0	5000.0
PREV AMT JUNE	8789.0	807.0	888.0	0.0	0.0	595.0	1210.0	4186.0
PREV AMT MAY	8789.0	798.0	887.0	0.0	0.0	557.0	1254.0	4000.0
PREV AMT APR	8789.0	752.0	873.0	0.0	0.0	477.0	1179.0	3746.0

1.3 ACCEPTABLE RANGES

Maintaining the integrity of the dataset in this credit card default prediction project is essential for precise model training and precise predictions. Dealing with outliers, or extreme data values that can skew the analyses and impair the effectiveness of the models, is a frequent problem. I used the Quantile method, a strong statistical technique that focuses on particular percentiles of the data distribution to identify and mitigate outliers, to address this.

Calculating the lower and upper quantiles—specifically, the lower quantile, or 25th percentile, and the upper quantile, or 75th percentile—allows to establish acceptable ranges for the data. These quantiles mark the boundaries at which data points are no longer regarded as outliers. Potential outliers include any data points that are below Q1 or above Q3:

- Upper Bound: Any data point over the upper quantile is removed at the upper quantile value. This makes sure that extremely high results, which are frequently linked to errors or abnormalities, do not skew the findings. There would be a more consistent and representative dataset by deleting these outliers.
- Lower Bound: In a similar manner, data points below the lower quantile are eliminated. This method protects against extremely low values that might not adequately reflect the underlying trends in the data. There would be better consistency and dependability in our dataset by deleting these outliers.

This increased the stability of the credit card default prediction model by using the Quantile approach. The algorithm is able to concentrate on patterns and trends that truly aid in the prediction of credit card defaults since outliers that may otherwise generate noise or bias were successfully handled.

1.4 MISSING VALUES

There are no missing values in this dataset. It made model construction easier, decreased biases, prevented imputation errors, and made sure that the learnt patterns were consistent throughout the dataset.

1.5 VARIABLE DRIFT MONITORING

In machine learning, the phenomenon known as "model drift" occurs when a machine learning model’s training assumption are no longer true in its application context.

The significance of conducting variable drift modelling becomes more apparent in the context of credit card default prediction due to its wide-ranging ramifications. Financial data are subject to dynamic variations brought on by shifting economic environments, changing consumer preferences, and altered regulatory frameworks. One can develop a thorough grasp of how particular predictors change over time by methodically studying variable drift. This exercise protects the correctness and dependability of the model in addition to maintaining data integrity. The key is in the model's adaptability, which shows solid performance in both training and testing contexts as it detects and accounts for fluctuations in feature distributions. In the financial industry, where rapid insights into shifting dynamics are crucial for accurate risk assessment and strategic decision-making, such adaptation is unavoidable.

Additionally, variable drift modelling mirrors regulatory vigilance by satisfying the need for routine model validation and assuring conformity with industry standards. In the end, this practise equips stakeholders, regulators, corporate executives with a tool that gives reliable, consistent predictions, cultivating a solid foundation for successful risk management strategies and maximising financial outcomes.

1.5.1 Tolerance for Drift of Each Variable

The more significant variables' drift tolerance is set at 6945.1, or 5% of the mean of the features utilised in model training. The drift tolerance is set at 1208.4 for the less significant features, which is 12% of the mean of the features utilised in model training. Instead of immediately rebuilding the models when a feature's drift exceeds its corresponding threshold, the accuracy score, confusion matrix, K-Fold Cross Validation, and Stratified K-Fold Cross Validation will be analysed in order to determine the health and stability of the models. Based on the risk tiering, the proper steps will then be taken.

2.0 MODEL MONITORING, HEALTH AND STABILITY

The stability, health, and monitoring of the model play a crucial role in ensuring its dependability and usefulness as a predictor of credit card default. There is a continuously evaluation of the model's real-world performance through diligent model monitoring by contrasting its predictions with actual results. This continuing examination not only finds any potential disparities but also sheds light on how well the model can adjust to shifting market conditions and consumer trends. In order to monitor the model's health, it is necessary to check that important metrics like accuracy, precision, and recall fall within set limits. The model's behaviour over time, notably its resistance to idea drift and variable shifts, must also be carefully examined in order to assess the model's stability. We ensure that the model maintains its capacity to produce precise forecasts, maintain regulatory compliance, and assist in making informed decisions by routinely assessing key variables.

2.1 INITIAL MODEL FIT STATISTICS

The accuracy score, confusion matrix, K-Fold Cross Validation and Stratified K-Fold Cross Validation were used to evaluate the effectiveness of this prediction model on credit card default. The accuracy score is a summary of how successfully the model, on average, predicts default and non-default instances. This accuracy is broken down in the confusion matrix as true positives, true negatives, false positives, and false negatives. Together, these measures help comprehend the model's capacity for correctly classifying default and non-default cases as well as the likelihood that it will misclassify some instances. An understanding of the model's efficacy is established through the initial assessment using these indicators.

K-fold cross-validation offers a more in-depth evaluation of the model's generalisation skills. This method offers a more thorough picture of the model's predictive capability and robustness, assisting in making defensible conclusions about its applicability for credit card default prediction. It does this by taking into account multiple training and validation subsets.

An improvement on K-fold that takes into account the class distribution within the dataset is called stratified K-fold cross-validation. Each fold in a stratified K-fold preserves a distribution of classes that is identical to the distribution in the original dataset since the class proportions are retained in each fold.

Accuracy Score

The selected model in this project is the Random Forest with an Accuracy Score of 79.64.

Confusion Matrix

The Random Forest Model correctly predicted 77% No Credit Card default (True Negatives) and correctly predicted 2.5% there would be a default (True Positive). The model predicted 6% that there would be a default where there was actually no default (False Positive), and predicted 14% there would not be a default but there was a default (False Negative).

K-Fold Cross Validation

The selected model in this project is the Random Forest with a K-Fold of 0.825.

Stratified K-Fold Cross Validation

The selected model in this project is the Random Forest with a Stratified K-Fold of 0.8238

3.0 RISK TIERING

Risk tiering includes evaluating the seriousness of prospective problems or hazards that may develop as a result of factors like variable drift in this credit card default prediction model. Distinct levels of risk are categorised or classified according to various components, such as characteristics, variables, or scenarios.

Action Steps for Each Risk Tier:

◊ Low-risk Tier: No urgent action is required if drift is less than 6945.1 for more significant features and 1208.4 for less significant features (both < 2%). The models continue to be consistent and match the initial training set of data.

◊ Moderate-Risk Tier: Drift between 2% and 5% suggests potential changes in the data distribution, falling into the moderate risk category.

◊ High- Risk Tier (6%-12% drift): When drift reaches this level, model refitting is carried out using fresh data samples. This proactive technique takes into account the effects of changing data patterns, ensuring that the model is reliable and accurate throughout time.

◊ Unacceptable- Risk Tier (more than 12% drift): Data that has drifted more than 10% may have undergone severe structural changes. To maintain accurate predictions in this scenario, model rebuilding becomes crucial.

4.0 Conclusion

Validation, monitoring and governance are crucial steps in the creation of any data-driven project or model, but they are especially important in complex and important fields like finance. The correctness, dependability, and moral application of models and data are ensured by the structured framework they offer.