1. Objective

The objective of the model is to predict customer default , using 23 predictors and a binary response variable namely default payment (valued '1'), and no default payment (valued '0').

Thereafter, a report to evaluate the models and to study its risk factors shall be produced.

2. Findings on Data Exploration and Preparation of Data for Mining

The response variable on default payment class is 22%, and on no-default payment class is 78% as shown in the below pie chart. The class distribution is significantly unbalanced.

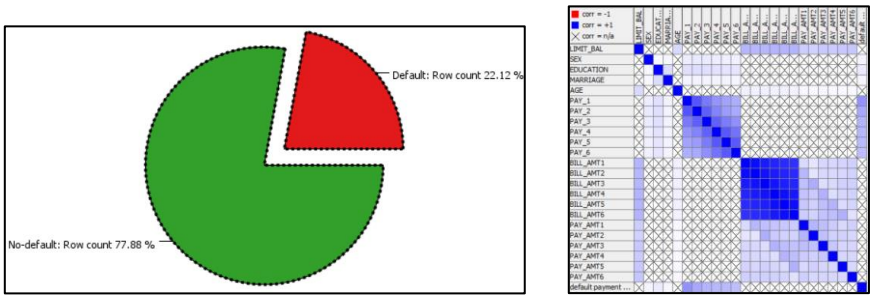

Diagram 1: Pie Chart on distribution of default and no-default classes (left); Correlation Chart on predictors to response variable (right)

There are no missing values. All the billing amounts, payment amounts, limit balance are positively skewed, with kurtosis > 10. For PAY_AMT2, the kurtosis is highest at 1,642, implying the amounts are heavy-tailed with outliers. The outliers are not imputed nor removed to avoid information loss, as these outliers might be genuine reflection of the data.

Using the linear correlation node under KNIME, it shows there is little correlation between the billing and payment amounts with the response variable on defaulting payment. The repayment status correlates better with the response variable positively but still weakly (in the range of +0.25 to +0.42), as depicted in diagram 1.

Data cleaning is performed on Education (to recode '5', '6', and '0' to '4'), as well as Marriage (to recode '0' to '3'). All the predictors are also being transformed into the correct format. ID column which is not meaningful for modelling, was dropped. Finally, the response variable is formatted to 'Default' and 'No-default' for classification prediction.

3. Selection of Decision Tree and Logistic Regression Models, and implementing with KNIME software

As 'labelled' dataset is to be mined, supervised learning algorithm is adopted to allow the machine to learn from the labelled data, and then used the learned model to predict the classification ('default' or 'No-default' payment) on new data.

Decision Tree Model and Logistic Regress Model, which are both popular supervised classifier are chosen for predicting this categorical binary outcome or response.

Below table summarised the algorithm and characterics of the 2 types of model:

	Decision Tree	Logistic Regression
Algorithms	Applied node splitting to iteratively dividing node into multiple sub-nodes to create relatively pure nodes. A pure node has only one class which implies it splits the class perfectly.	Applied logit transform on regression to predict and classify probalities.
Error Minimization Technique	Maximise Information gain. An attribute with highest information gain will be tested and splitted first. Gini and Entropy are criterions for calculating Information Gain.	Maximum likelihood estimation to arrive at a best possible fit.
Advantages	• Capable of handling both continuous and categorical variables • Provide indication of which fields are most important for prediction that is easily understood by human	• Capable of providing probabilities and classifying new data with continuous and discrete datasets • Provide indication on importance of predictors via coefficient values, and its positive/ negative association.

Table 1: Characteristics of Decision Tree and Logistic Regression Models

Considering appropriateness of the dataset to be mined and predicted on, as well as the above described advantages, and since Decision Tree and Logistic Regression models are easy to be trained and implementated, they are selected for this homework.

4. Evaluation of Decision Tree Model and Logistic Regression Model

As the class for the response is imbalanced, under-sampling shall be performed prior to training the models. Imbalance class refers to the problem when an outfome occurs less frequently than the other in this binary response dataset.

In this scenario, the minority class is default payment that make up only 22% of the total records. The majority class is no-default payment which is more than 3 times higher than the minority. To prevent bias in prediction due to the uneven distribution, it is necessary to apply 'balancing' on the class.

Under-stampling is one of the techniques that could be applied to balance the class, by keeping all the minority class records and then decreasing the size of the majority class records.

The Decision Tree and Logistic Regression model were also constructed with cross validations and tuned hyper parameter (on Minimum number of records per node and prior variance). Their scored metrics were tabulated in the below table:

	Decision Tree, with under-sampling		Logistic Regression, with under-sampling	
	Train Set	**Test Set**	**Train Set**	**Test set**
Confusion Matrix	Default | No-default 3439 | 1870 1358 | 3917	Row ID | Default | No-default Default | 838 | 489 No-default | 356 | 963	Row ID | Default | No-default Default | 3681 | 1628 No-default | 2340 | 2935	Row ID | Default | No-default Default | 889 | 438 No-default | 411 | 908
Accuracy	0.695	0.681	0.625	0.679
Sensitivity	0.648	0.631	0.693	0.670
Specificity	0.743	0.730	0.556	0.688

Table 2: Scored metrics from Decision Tree and Logistic Regression Model

In the confusion matrix for the test set, the Decision Tree correctly classified 1,801 as default and no-default (out of 2,646 test records). 489 records were incorrectly classified as No-default, while 356 records were incorrectly classified as Default.

For the same number of test records, the Logistics Regression correctly classified 1,797 records. 438 reords were incorrectly classified as No-default, while 908 records were incorrectly classified as Default.

The accuracy for both models on this dataset is similar and about 68%. Accuracy is the proportion of the summation of true positives as well as the true negatives that are predicted correctly

The Logistic Regression model has a slightly better Sensitivity score at 67%, compared to that of 63% from the Decision Tree model. Sensitivity is the proportion of true positives that are predicted correctly.

However, the Decision Tree model has a slightly better Specificity score at 73%, compared to that of 69% from the Logistic Regression model. Specificity is the proportion of true negatives that are correctly predicted by the model.

Depending on objective(s) of the problem statement and data to be studied, the appropriate or a combination of the above metrics shall be adopted to evaluate the desired model.

For example, if the objective of the model is to preduct the positive or default payment bahviour, model with higher sensitivity is preferred. If the objective of the model is to predict the no-default beviour for loyalty membership, model with high specificity is preferred, etc.

4.1 Decision Tree Model and its interpretation

Below is the model generated from the learner node in KNIME:

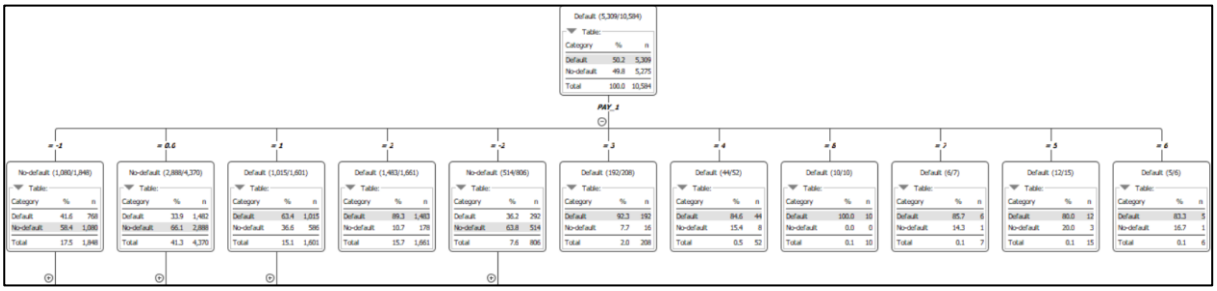

Diagram 2: Ouput of Decision Tree generated by KNIME

As justified in section 3, Decision Tree Model provides clear indication of which field is the most important for prediction, and based on the output in diagram 2, repayment status (PAY_1) is the most important predictor.

If the PAY_1 is -1, 0 or -2, it is more likely for No-default at likelihood 58%, 66% and 64% respectively. However, if the PAY_1 is 1, 2 or 3, it is more likely to Default payment at likelihood 63%, 89% and 92% respectively.

4.2 Logistic Regression Model and its interpretation

In section 3, it was explained that Logistic Regression Model provides indication on importance of predictors and its association via the coefficient values.

Diagram 3 contains the coefficient statistics generated from the learner node in KNIME.

S Variable	D Coeff.	D Std. Err.	D z-score	D P>|z|
LIMIT_BAL	-9.944	0.279	-35.607	0
SEX=1	9.392	0.28	33.519	0
EDUCATION...	-6.457	0.266	-24.236	0
EDUCATION...	-2.85	0.23	-12.402	0
EDUCATION...	-27.32	1.196	-22.839	0
MARRIAGE=2	-4.643	0.216	-21.493	0
MARRIAGE=3	-1.207	0.585	-2.063	0.039
AGE	0.579	0.018	32.899	0
PAY_1=-1	-51.826	0.925	-56.045	0
PAY_1=0	-86.075	1.192	-72.216	0
PAY_1=-2	-55.273	0.947	-58.39	0
PAY_1=1	-33.123	1.061	-31.216	0
PAY_1=3	12.604	1.732	7.277	0
PAY_1=4	-2.176	1.732	-1.256	0.209
PAY_1=8	2.5	1.732	1.444	0.149
PAY_1=7	-0.616	1.732	-0.356	0.722
PAY_1=5	-1.283	1.732	-0.741	0.459
PAY_1=6	-0.761	1.732	-0.439	0.66
PAY_2=0	-3.933	0.821	-4.793	0

BILL_AMT1	-0.193	0.334	-0.578	0.564
BILL_AMT2	6.099	0.476	12.827	0
BILL_AMT3	4.795	0.464	10.336	0
BILL_AMT4	2.02	0.409	4.944	0
BILL_AMT5	3.337	0.57	5.85	0
BILL_AMT6	-2.993	0.509	-5.884	0
PAY_AMT1	-2.127	0.209	-10.175	0
PAY_AMT2	-4.197	0.282	-14.892	0
PAY_AMT3	-4.338	0.422	-10.27	0
PAY_AMT4	2.573	0.172	14.943	0

Diagram 3: Coefficent Statistics generated by KNIME

The following would be interpreted from the coefficient and p-values:

- Keeping the other predictor constant, for a 1 unit change in Age, the odds for defaulting payment increases by 0.78 ($e^{0.579} - 1$).
- The odds of defaulting payment for male (Sex=1) is 11,992 (or $e^{9.392}$) times the odds pf defaulting payment for female, keeping the other predictor constant.
- PAY_1=4, 5, 6, 7, 8 and BILL_AMT1 might not be statistically significant in classifying default payment, as its p-values are > 0.05.

5. Risk Factors and its Mitigation

In general, for most predicting classifiers to perform well, it is important to have:

- Predictors that are correlated well to the response variable
- Predictor variables to be independent of each other, with little or no multicollinearity

As shown in the correlation matrix in diagram 1, the correlation between the response variable and most predictors are not strong (with lighter shade).

In addition, the matrix also discovered that some predictors are related, such as between the repayment statues, and between the billing amounts (which are in darker shades).

Thus, the 23 predictors in the dataset might not be ideal and exhaustive in predicting the defaulting behaviour. Inclusion of additional predictors such as income levels, occupational sectors (such as in enforcement, education, entertainment industries), etc might improve the performance of the prediction.

Below table summarised other risks factors for the 2 models, and the mitigations applicable or put in place:

Model	Risks	Mitigations
Decision Tree	(i) Result in bias model If a class label dominates (ii) Tend to overfit	(i) The followings were done to prevent bias in the model: • Under-sampling using KNIME's equal sampling node is implemented • Stratified sampling by response/outcome variable is specified (ii) The following mitigation measure to avoid over-fitting: • Incorporated Cross-validation. • Specified Post-pruning using Minimum Description Length (MDL)
Logistic Regression	(i) Works better when attributes that are unrelated to the response variable are removed. (ii) Normalisation is needed generally if the numeric predictors have different ranges and heavily tailed.	(i) Insignificant or unrelated predictors (with p-value < 0.05) can be removed from the model. However, this removal is not done, as more information on the metadata needed in order to carry out the removal well. (ii) The goal of normalisation is to change the values of different numeric predictors into a common scale, without distorting the values. Normalizer is applied to the balanced data in KNIME.

Table 3: Risk Factors and Mitigations

<< End of Report >>