

GERMAN CREDIT CARD

Analysis and Recommendations

CONTENT



Business Problem



Data Preparation



**Predictive
Modeling/Classification**

Decision Tree
Naïve Bayes
Logistic Regression



**Post-predictive
Analysis**

K-Means
Apriori Algorithm



**Conclusion and
Recommendation**

Business Problem



The bank needs to make a decision regarding loan applications based on each customer classified as good or bad credit



Solution is to create a data analytics-based strategy for the bank managers to approve or deny loan applications



DATA PREPARATION

GERMAN CREDIT CARD DATA

- ▶ 1,000 instances with 21 attributes
 - ▶ 700 good credit cases
 - ▶ 300 bad credit cases
- ▶ Attribute Type
 - ▶ 6 Numeric, 10 Nominal, 5 Ordinal
 - ▶ E.g. of attributes: Type of apartment, Credit Amount, Value savings/stocks etc.
- ▶ Class attribute describes instances as good or bad

- **Duration of Credit**
 - Average approximately 21 months
- **Credit Amount**
 - Mean > Median
 - Right skewed distribution
 - More people with smaller amounts of credit comparatively maximum amount for the loan is \$18,424
- **Instalment per cent**
 - Mean < Median
 - Slightly left skewed distribution
 - More people who have 3-4% instalment rate
- **Age**
 - Mean > Median
 - Right Skewed distribution
 - Younger people are more likely to apply for loan or credits (early 30s)
- **No. of Credits at this Bank**
 - Mean > median
 - Right Skewed distribution
 - More than 50% of the data has only 1 credit at this bank
- **No. of Dependents:**
 - Minimum, 1st Quartile, Median and 3rdQuartile all are 1
 - 75% or more people have only 1 dependent in their credit application

NUMERICAL DATA SUMMARY

Attribute	Maximum	Minimum	Mean	Median	Std. Deviation
Duration of credit (month)	72	4	20.903	18	12.059
Credit Amount	18424	250	3271.25	2320	2822.752
Installment per cent	4	1	2.973	3	1.119
Age (years)	75	19	35.542	33	11.353
No. of Credits at this Bank	4	1	1.407	1	0.578
No. of dependents	2	1	1.155	1	0.362

DISTRIBUTION OF ALL ATTRIBUTES

► Outlier & Missing Value

- Outliers were not removed
 - 975 No (if it was not an outlier) and only 25 Yes (if it was an outlier)
- There were no missing value for any of the 21 attributes

► Right-Skewed Distribution

- **Age:** People who are below the age of 40 tend to apply for loan. Early part of the histogram has higher frequency
- **Credit Amount:** People tend to seek a credit amount less than 8,000



FEATURE SELECTION

► Elimination or Inclusion of attributes based on the different feature selection methods

- Correlation
- Information Gain
- Learner Based - J48
- Chi-Square
- Symmetrical Uncertainty

► Attribute Identified

- Account Balance
- Value Savings/Stocks
- Duration of Credit
- Purpose
- Payment Status of Previous Credit

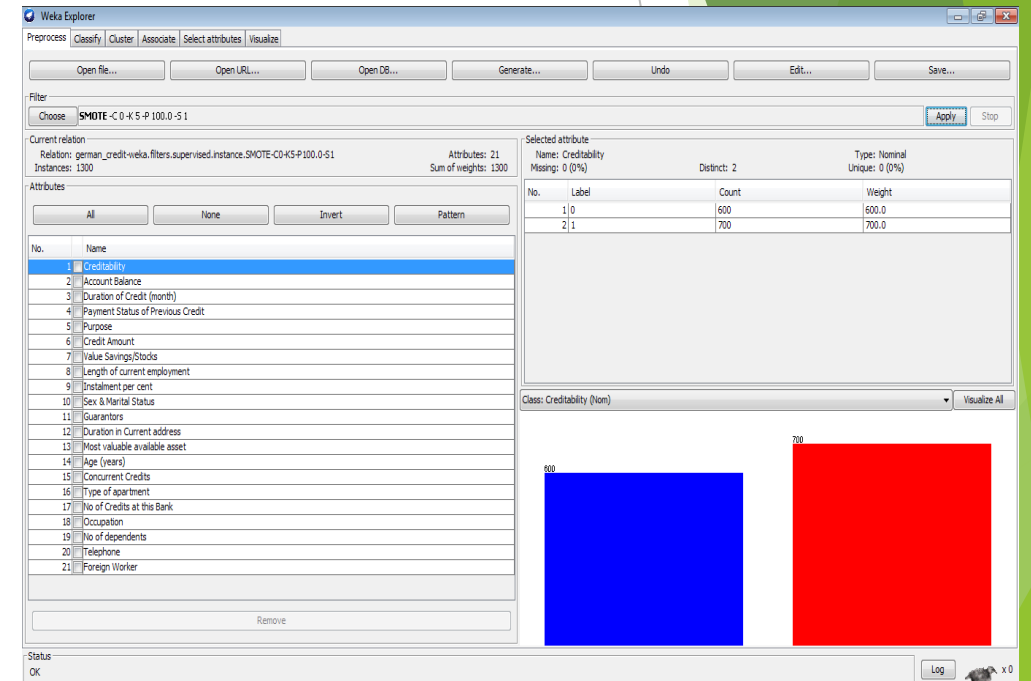
Attribute	Correlation	Information Gain	Learner based - J48	Chi-Square	Symmetrical Uncertainty
Account Balance	0.23276	0.094739	1	123.7209	0.070613
Duration of Credit (month)	0.21493	0.0329	2	46.8311	0.030272
Value Savings/Stocks	0.13162	0.028115	4	36.0989	0.021887
Payment Status of Previous Credit	0.08988	0.043618	3	61.6914	0.033641
Credit Amount	0.15474	0.018333		26.3992	0.021448
Type of apartment	0.12283	0.013077		18.674	0.012963
Purpose	0.07494	0.024894		33.3564	0.014033
Length of current employment	0.0527	0.013102		18.3683	0.00863
Instalment per cent	0.0724	0		0	0
Sex & Marital Status	0.07192	0.006811		9.6052	0.005644
Most valuable available asset	0.05838	0.016985		23.7196	0.012008
Age (years)	0.09127	0.011278		16.3681	0.014251
Concurrent Credits	0.108	0.008875		12.8392	0.010284
Occupation	0.01904	0.001337	5	1.8852	0.001166
Foreign Worker	0.08208	0.005823	6	6.737	0.010495
No of dependents	0.00301	0		0	0
No of credit at this Bank	0.04573	0		0	0
Guarantors	0.00612	0.00479		6.6454	0.006758
Duration in current Address	0.01096	0.000543		0.7493	0.000398
Telephone	0.03647	0.000964		1.3298	0.001039



PREDICTIVE MODELING

Classification & smote

- ▶ **Classification** is a two-step process
 - ▶ Learning step
 - ▶ Model is developed based on given training data
 - ▶ Prediction step
 - ▶ Model is used to predict the response for given data
- ▶ **Three** classification algorithms were used
 - ▶ Naïve Bayes
 - ▶ Decision Tree
 - ▶ Logistic Regression
- ▶ **SMOTE OVERSAMPLING**
 - ▶ 300 additional records (synthetic data) were added to '0' class attribute
 - ▶ Increased the total number of '0' to 600 and total records to 1300



Naïve Bayes

- ▶ Probability-based machine learning model
- ▶ The classifier is based on the Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- ▶ Fast & easy classifier to predict a class of data set
- ▶ Performs well in case of categorical variable compared to numerical variables

Naïve Bayes results

Balanced Data with 20 Attributes (with SMOTE)

Algorithm (Naive Bayes)	Cross Validation, K=10,	Percentage Split, 60%	Percentage Split, 90%
Correctly Classified instances (%)	80%	80.9%	83.8%
Incorrectly Classified instances (%)	20%	19.8%	16.1%
F-Score	0.800	0.803	0.839
Precision rate	0.800	0.804	0.840
Recall rate	0.800	0.802	0.838

- ▶ Tests were performed using all 20 attributes and later selecting only the top 6 attributes. No significant improvement in accuracy was observed when the number of parameters were decreased
- ▶ Dataset was split using strategies such as 10-Fold and percentage split techniques (60-40). When using an unbalanced dataset, a 10-fold split strategy performed better
- ▶ When using a balance dataset (SMOTE). The Maximum Accuracy was achieved by using a 60- 40% split

Confusion matrix

	Classified	
Test Data	Good application	Bad Application
Good application	47	9
Bad application	12	62
Accuracy = $(47+62)/130 = 83.8\%$		

DECISION TREE

- ▶ Supervised learning algorithm
- ▶ Used for solving Regression and Classification problems
- ▶ Data is split into Nodes and Leaves
 - ▶ Account Balance was the root node and had the highest gain among all the parameters
- ▶ **J48 classifier was used in Weka for Decision Tree**

DECISION TREE results

- ▶ Account Balance was the root node in all the test cases as it had the highest gain
- ▶ A total of 16 tests were performed and Pruning was set to True
- ▶ All 20 independent attributes used to construct a decision tree and two types of testing techniques were used
 - ▶ 10-fold test: accuracy of 72.8%
 - ▶ Percentage Split (ratio of 60:40): accuracy of 71.8%
- ▶ SMOTE improved the accuracy of the model
 - ▶ 78.4 % accuracy using J48 algorithm on a total of 520 test records
 - ▶ Bagging technique used to check if overfitting was caused
 - ▶ Highest accuracy of 79.4% with a combination of Bagging + SMOTE + Percentage Split Test
- ▶ Same tests conducted on 6 attributes
 - ▶ Highest accuracy of 78.4% achieved by combination of Bagging + SMOTE + Percentage Split Test
- ▶ Decreasing the number of attributes to 6 did not make any improvement in the classification of the model

LOGISTIC REGRESSION

- ▶ Logistic Regression is a linear method
 - ▶ Set threshold at 0.5
 - ▶ Output greater than 0.5 will be classified as good credit individual
- ▶ Cost sensitive learning applied to reduce the imbalance feature of our dataset
 - ▶ Cost of prediction errors when training the model
 - ▶ As a bank, the cost of lending out a loan to a bad customer (false positive) can result in much greater losses than denying a loan to a good customer (false negative)
 - ▶ The aim of using a cost sensitive model is to minimize the overall **error cost** and not optimize the correctness of classifications
- ▶ Metaclassifier **CostSensitiveClassifier** in Weka was chosen

LOGISTIC REGRESSION results

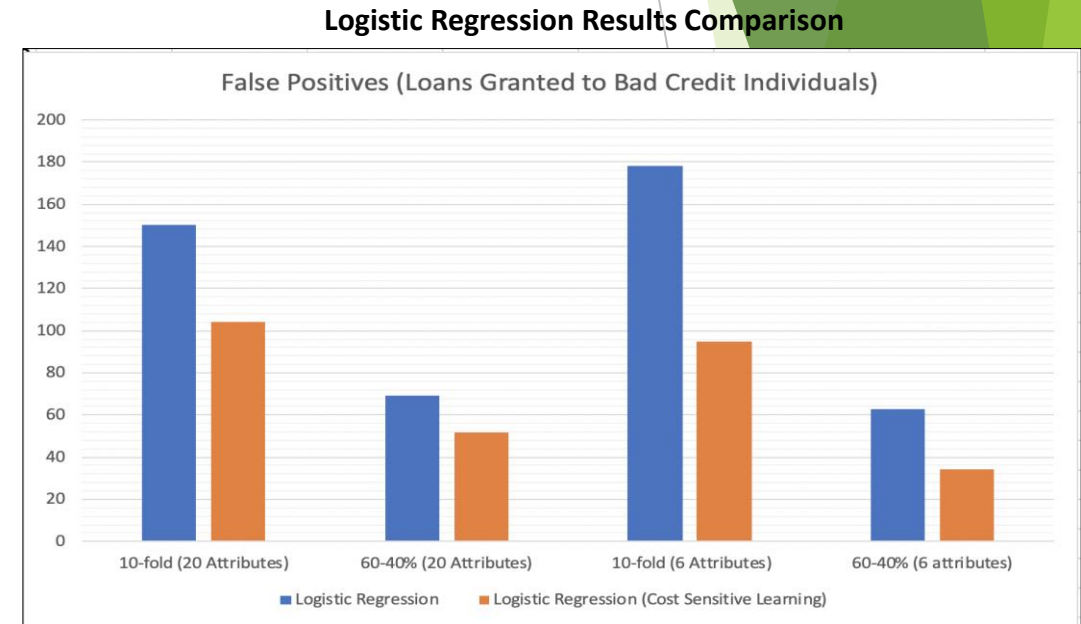
- Accuracy varied between 70 and 75% in all the test cases
- Best strategy to split the dataset is 10-fold cross validation rather than a simple percentage split
 - All instances have a higher accuracy using the cross-validation training strategy
 - The f-score also confirms the same
 - Best results in terms of accuracy was found to be 75.2% using all 20 attributes and a 10- fold strategy

Results of Logistic Regression on 20 Attributes

S. No	Parameters Used	No of Attributes	Cost Matrix	Test Data Method	Test Data Size	Accuracy (Correctly Classified Instances)	Weighted Precision	Weighted Recall	Weighted F Measure	Confusion Matrix
1	Logistic regression	20	NA	10-fold	1000	75.20%	0.742	0.752	0.745	a b <-- classified as 150 150 a = 0 98 602 b = 1
2	Logistic Regression	20	NA	60-40%	400	73.50%	0.725	0.735	0.725	a b <-- classified as 66 69 a = 0 37 228 b = 1
3	Logistic regression with cost sensitive learning	20	Cost Matrix 0 2 1 0	10-fold	1000	72.40%	0.745	0.724	0.731	a b <-- classified as 196 104 a = 0 172 528 b = 1
4	Logistic regression with cost sensitive learning	20	Cost Matrix 0 2 1 0	60-40%	400	70.50%	0.713	0.705	0.708	a b <-- classified as 83 52 a = 0 66 199 b = 1

LOGISTIC REGRESSION (Cost SENSITIVE STRATEGY)

- ▶ Cost sensitive model
 - ▶ To minimize the overall **error cost** and not optimize the correctness of classifications
- ▶ Cost sensitive strategy will not improve the **weighted** accuracy, precision and recall of the model but does improve the false positive rate of the model
- ▶ Out of all the bad credit individuals, the number of misclassifications is lower



CLASSIFICATION ALGORITHM COMPARISON

- ▶ The Highest accuracy of 75.4% was achieved with Naïve Bayes classification method on all 20 attributes and a 10-fold test
- ▶ 10-fold training/testing strategy yielded better results than a simple percentage split
- ▶ Introducing SMOTE improves the accuracy of the model but this may be due to overfitting

Comparison of All Classification Algorithms

No	Algorithm	Testing Method	Number of Instances	Accuracy	Precision	Recall	F-Measure
1	Naive Bayes	10-Fold Test	1000	75.4	0.743	0.754	0.746
2	Naive Bayes	60-40 Split	400	74.5	0.737	0.745	0.726
3	Decision Tree	10-Fold Test	1000	72.8	0.712	0.728	0.716
4	Decision Tree	60-40 Split	400	71.75	0.706	0.718	0.687
5	Logistic Regression	10-Fold Test	1000	75.2	0.742	0.752	0.745
6	Logistic Regression	60-40 Split	400	73.5	0.725	0.735	0.725

POST-PREDICTIVE MODELING

POST PREDICTIVE ANALYSIS

- ▶ To identify the characteristics of customer whose loan was approved as well as the characteristics of those who got disapproved
- ▶ Strategies used
 - ▶ **K-Means**
 - ▶ **Apriori**

K-MEANS CLUSTER results

- ▶ K-Means algorithm was used for cluttering data and to find homogeneous subgroups within the data
- ▶ Partition the dataset into k pre-defined clusters. Points within a cluster are as similar as possible and inter-cluster distances are as far as possible
- ▶ It uses the sum of least squares technique. Elbow method was used to determine the number of clusters
- ▶ Class to cluster evaluation in K-Means was used to segregate the data, based on credibility u
 - ▶ Three different tests that we performed
 - ▶ Class to Cluster evaluation
 - ▶ Cluster population whose credit was approved (700) u Test on only those who were disapproved (300)

K-MEANS Results

Cluster analysis of population whose credit was approved (700)

- ▶ Cluster 0 consists of middle-aged people who seek lower credit amount and seek loans for a short duration of time. Another characteristic of this population is that they have low valuable assets.
- ▶ Cluster 1 consists of an older population who applied for a higher credit amount. people in this cluster also required a longer duration of credit. The credit purpose for these individuals leans towards a mean of 0. This means that these clients used the loan to buy a car.
- ▶ Cluster 2 consisted of middle-aged clients who had applied for higher credit. The people in this cluster also applied for a longer duration of credit. An important trait of these customers consisted that they had 3 or more valuable assets.

Cluster analysis of population who were disapproved (300)

- ▶ Loan applicates in Cluster 0 were denied as they were old people who applied for very high credit amount. Bank regarded these individuals as being too risky.
- ▶ Both clusters consisted of young individuals who had relatively less experience but wanted long duration of credit. It is possible that their loan application was declined due to the high amount requested as well as lower job security.

Class to Cluster evaluation

- ▶ Account balance, payment status of previous credit, credit amount, instalment per cent, duration in current address and most valuable available asset are attributes which differentiates good and bad debtors.
- ▶ Accuracy of this experiment was low a 55%.

Good Credit Clusters				
Attribute	Full Data (700.0)	Cluster#		
		0 (184.0)	1 (223.0)	2 (293.0)
Account Balance	4	4	4	4
Duration of Credit (month)	19.2071	16.5761	19.157	20.8976
Payment Status of Previous Credit	2	2	4	2
Purpose	3	3	0	3
Credit Amount	2985.4429	2460.1196	3146.5695	3192.7065
Value Savings/Stocks	1	1	1	1
Length of current employment	3	2	5	3
Instalment per cent	2.92	2.5978	3.0493	3.0239
Sex & Marital Status	3	2	3	3
Guarantors	1	1	1	1
Duration in Current address	4	4	4	2
Most valuable available asset	3	1	2	3
Age (years)	36.22	34.1576	42.148	33.0034
Concurrent Credits	3	3	3	3
Type of apartment	2	2	2	2
No of Credits at this Bank	1.4243	1.2228	1.722	1.3242
Occupation	3	3	3	3
No of dependents	1.1557	1.0435	1.3004	1.116
Telephone	1	1	1	1
Foreign Worker	1	1	1	1

Bad Credit Clusters			
Attribute	Full Data (300.0)	Cluster#	
		0 (131.0)	1 (169.0)
Account Balance	1	1	1
Duration of Credit (month)	24.86	26.542	23.5562
Payment Status of Previous Credit	2	2	2
Purpose	0	0	3
Credit Amount	3938.1267	4576.1603	3443.5562
Value Savings/Stocks	1	1	1
Length of current employment	3	5	3
Instalment per cent	3.0967	3.1832	3.0296
Sex & Marital Status	3	3	2
Guarantors	1	1	1
Duration in Current address	4	4	2
Most valuable available asset	3	4	3
Age (years)	33.96	40.1756	29.142
Concurrent Credits	3	3	3
Type of apartment	2	2	2
No of Credits at this Bank	1.3667	1.4427	1.3077
Occupation	3	3	3
No of dependents	1.1533	1.2901	1.0473
Telephone	1	1	1
Foreign Worker	1	1	1

APRIORI

- ▶ To determine the association between attributes
- ▶ Association Algorithm applicable only to nominal, binary and unary data
- ▶ Numerical attributes in the data set were converted to Nominal using Discretize and NumericToNominal features in Weka
- ▶ Key Parameters
 - ▶ Support
 - ▶ Measures the popularity of an attribute in the dataset
 - ▶ Confidence
 - ▶ Measures the probability of attribute X being chosen, given attribute Y
- ▶ Association rules were established by performing tests on customers belonging to the good credit class

ASSOCIATION RULES USING APRIORI

Interpretation based on best rules found:

- ▶ Debtors seek smaller credit amount
- ▶ Good debtors have lower savings
- ▶ Good debtors paid back their old loans
- ▶ Loan seeking population mostly constituted of foreign works
- ▶ Most of the loan applications which were approved did not require a guarantor
- ▶ Debtor who were foreign workers required no guarantors in 611 cases
- ▶ In most cases, debtor has only 1 dependent
- ▶ Debtor owns his/her apartment in 528 instances
- ▶ Debtors consisted of skilled employees

Best rules found:

```
1. Foreign Worker=1 667 ==> Creditability=1 667    conf:(1)
2. Guarantors=1 635 ==> Creditability=1 635    conf:(1)
3. Guarantors=1 Foreign Worker=1 611 ==> Creditability=1 611    conf:(1)
4. No of dependents=1 591 ==> Creditability=1 591    conf:(1)
5. Concurrent Credits=3 590 ==> Creditability=1 590    conf:(1)
6. No of dependents=1 Foreign Worker=1 569 ==> Creditability=1 569    conf:(1)
7. Concurrent Credits=3 Foreign Worker=1 561 ==> Creditability=1 561    conf:(1)
8. Guarantors=1 No of dependents=1 539 ==> Creditability=1 539    conf:(1)
9. Guarantors=1 Concurrent Credits=3 538 ==> Creditability=1 538    conf:(1)
10. Type of apartment=2 528 ==> Creditability=1 528    conf:(1)
11. Guarantors=1 No of dependents=1 Foreign Worker=1 524 ==> Creditability=1 524    conf:(1)
12. Guarantors=1 Concurrent Credits=3 Foreign Worker=1 517 ==> Creditability=1 517    conf:(1)
13. Type of apartment=2 Foreign Worker=1 504 ==> Creditability=1 504    conf:(1)
14. Concurrent Credits=3 No of dependents=1 503 ==> Creditability=1 503    conf:(1)
15. Concurrent Credits=3 No of dependents=1 Foreign Worker=1 485 ==> Creditability=1 485    conf:(1)
16. Guarantors=1 Type of apartment=2 476 ==> Creditability=1 476    conf:(1)
17. Guarantors=1 Concurrent Credits=3 No of dependents=1 461 ==> Creditability=1 461    conf:(1)
18. Guarantors=1 Type of apartment=2 Foreign Worker=1 458 ==> Creditability=1 458    conf:(1)
19. Guarantors=1 Concurrent Credits=3 No of dependents=1 Foreign Worker=1 449 ==> Creditability=1 449    conf:(1)
20. Type of apartment=2 No of dependents=1 446 ==> Creditability=1 446    conf:(1)
21. Occupation=3 444 ==> Creditability=1 444    conf:(1)
22. Concurrent Credits=3 Type of apartment=2 444 ==> Creditability=1 444    conf:(1)
23. No of Credits at this Bank=1 433 ==> Creditability=1 433    conf:(1)
24. Type of apartment=2 No of dependents=1 Foreign Worker=1 429 ==> Creditability=1 429    conf:(1)
25. Occupation=3 Foreign Worker=1 428 ==> Creditability=1 428    conf:(1)
```

Best rules found:

```
1. Credit Amount='(-inf-3279]' 483 ==> Creditability=1 483    conf:(1)
2. Value Savings/Stocks=1 386 ==> Creditability=1 386    conf:(1)
3. Payment Status of Previous Credit=2 361 ==> Creditability=1 361    conf:(1)
4. Account Balance=4 348 ==> Creditability=1 348    conf:(1)
5. Duration of Credit (month)='(-inf-15.333333]' 342 ==> Creditability=1 342    conf:(1)
6. Duration of Credit (month)='(-inf-15.333333]' Credit Amount='(-inf-3279]' 303 ==> Creditability=1 303
```


CONCLUSION BASED ON APRIORI

- ▶ Based on the association rules derived using Apriori Algorithm, we recommend to the bank that good debtors are people:
 - ▶ Who do not require guarantors
 - ▶ Have no additional loans
 - ▶ Have less dependents
 - ▶ Own their household
 - ▶ Have low savings and are interested to borrow
 - ▶ Pay off loans in a timely manner
- ▶ Bank managers should not approve loans of people whose checking account balance is < 0 DM (Deutsche Mark) and average balance in savings and stocks is < 100 DM. These customers are risky and may not be capable of fulfilling their credit obligations.

RECOMMENDATIONS

- ▶ Attributes such as age, most valuable available assets, purpose, credit history, and employment skill levels are important factors to determine the credibility of a potential client
- ▶ Creditable borrowers don't need guarantors, have no other loans/credits with no existing instalment plans, have fewer dependents, own their own apartments, have low savings, and pay off loans on time. Low risk loans also consist of those that are lower in amount and duration. Based on these recommendations, prospective borrowers not profiled as creditable should have markers that deviate from the profile above.
- ▶ Some of the “top 6 features” like value savings/stocks yielded common results between good credit and bad credit individuals. As a result, additional features beyond ‘the top 6’ are highly recommended to be used by the bank in order to make the right decisions.