Financial Risk Modelling in Vehicle Credit Portfolio

U Bhuvaneswari *
bhuvaneswari.u2013@vit.ac.in

P. James Daniel Paul *
jamesdanielpaul.p@vit.ac.in

Siddhant Sahu *
siddhant.sahu2011@vit.ac.in

* VIT University, Chennai, Tamil Nadu, India

Abstract — **Luxury cars are a segment of vehicles which are usually bought by people with a higher purchasing power. Still, majority of people make this luxury investment through vehicle finance services. The people from this segment tend to have a good credit record and thus are granted credit by vehicle finance service providers. Despite the good credit record and high purchasing power, a certain amount of risk is associated with these credit portfolios. This study deals with the analysis of a data set comprising of opulent vehicle credit portfolios characterized by relevant variables. It aims at assessing the risk associated with these portfolios and finally presents a predictive model which highlights the important variables and depicts the combination of those variables that classify a client under defaulter or non-defaulter. The study starts with the use of conventional statistical techniques and subsequently presents machine learning approach using three different decision tree classifiers.**

Keywords- Credit Risk; Vehicle Finance; Decision Tree Classifiers; Machine Learning

I. INTRODUCTION

Despite being movable assets, capital expenditure on passenger cars have assumed higher order of importance next to housing. Luxury cars indicate the wealth of a household and are thus an important tangible asset. Banks and various credit assessment agencies evaluate the wealth of individuals based on the passenger cars owned by them. If the risk of default in the automotive finance is 10-15%, the risk involved in lending in case of luxury segment is around 5-7%. Interestingly, since the value of the luxury passenger cars is significant, despite lower volume of default, the incidence of default is higher. Hence this study focuses on identifying the determinants of default using multiple methods like ANOVA, regression and decision trees on a luxury passenger car consumer data set in order to understand the variables and the methodology used in earlier studies, the study commenced with the review of empirical evidence.

II. REVIEW OF LITERATURE

L.G. Kabari et al. (2013) [1] in their research considered 1000 observations with 600 defaulter and 400 non-defaulter observations. They used variables like accommodation, job experience, place of work, loan size, account type, debt balance ratio, income, residency, guarantor, collateral, social security, age and marital status. The system used an integration of decision trees and artificial neural networks with a hybrid of decision tree algorithm and multilayer feed-forward neural network with back propagation learning algorithm to build up the proposed model. The research work thus proposed decision tree-neuro based credit risk evaluation system with 88% accuracy for decision support.

H.C. Koh et al. (2006) [2] in their research considered 1,000 observations, 70% good and 30% bad risk, 7 numerical and 13 categorical. They used variables like gender, age, credit history, credit amount, account, savings account/bonds, duration, other debtors/guarantors, installment plan and purpose. They used logistic regression, neural network and decision trees. Their key finding was that credit scoring will continue to be a major tool in predicting credit risk in consumer.

Raquel Florez-Lopez (2012) [3] used a data set of 2470 Taiwanese high-technology companies that are traded on the Taiwan's security market. The variables used were return on total assets, interest expense to net sales ratio, debt–equity ratio, asset growth, liabilities to assets ratio, inventory turnover ratio, gross profit margin, sale profit margin, return on equity, operating profit margin, earnings per share, current ratio, debt ratio, cash flow adequacy ratio, fix assets turnover ratio, debt payment ability, proportion of shares held by board and directors, proportion of mortgage shares held by board member, proportion of board members who are executive, proportion of independent directors on the board, proportion of shares held by the block stockholders. The model used by them included decision trees namely C 4.5, random forest, CART and pure SVM model. The key finding was that KMV are an important ex-ante indicator of corporation's credit risk.

Cheng-Lung Huang et al. (2007) [4] used 307 instances of creditworthy applicants and 383 instances where credit is not creditworthy, 6 nominal, 8 numeric attributes, and 1 class attribute. The models used was support vector machine (SVM) + grid search; SVM + grid search + F-score; SVM + GA. According to their study, a hybrid SVM-GA system is a good alternative for optimizing the parameters and feature subset. With a small feature subset, a hybrid SVMGA system can obtain a good classification performance. However, when using SVM-GA strategy (as well as GP and BPN), one should avoid over-training.

Junni L. Zhanga et al. (2010) [5] used variables like net income/total assets, net income/total sales, operating income/total assets, operating income/total sales, earnings before interest and tax/total assets, earnings before interest, tax, depreciation and amortization/total assets, earnings before interest and tax/total sales, own funds/total assets, own funds in-tangible assets/total assets intangible assets cash and cash equivalents, lands and buildings etc. The models were the SVM, gradient boosting and random forests.

978-1-4799-4674-7/14/$31.00 ©2014 IEEE

Stjepan Oreski et al. (2012) [6] in their research considered 1000 cases with a time horizon of 7 years. They collected 35 attributes divided into 5 groups. They used variables namely age, gender, postcode, telephone, time at present address, time with the bank, age of the clients current account, month of loan approval date, loan maturity in months (repayment period), purpose of loan, loan amount (in HRK), total payments to the account (monthly income), cash payments to the account/ total payments, regular payments (salaries) / total payments, contracted overdraft/ total payments, total withdrawals from the account (outcome), total payments/ total withdrawals, EFTPOS withdrawals/ total withdrawals, ATMs withdrawals/ total withdrawals, self-service withdrawals/ total withdrawals, contracted overdraft/ total withdrawals, installment to disposable income ratio, income on the loan approval date to the previous income year ratio, contracted overdraft, time deposits (balance on the loan approval date), balance of all accounts in the bank, balance/ contracted overdraft, number of times client exceeded contracted overdraft, interest on positive balance/ interest on overdraft, interest on overdraft history, number of loans with amount greater than current , number of loans with amount less than or equal to current, credit history and internal rating. They used genetic algorithm, forward selection, information gain, gain ratio, gini index and correlation neural network generic. They inferred that GANN model is significantly better in feature selection for classification.

Tian-Shyug Lee et al. (2006) [7] in their research work used a data set of credit card data comprising 8000 customers, used variables like good credit, bad credit, gender, age, marriage status, educational level, occupation, job position, annual income, residential status and credit limits on CART, MARS, discriminant analysis, logistics regression, neural networks model, support vector machine. They concluded that CART and MARS both have better average correct classification rate in comparison with discriminant analysis, logistic regression, neural networks, and support vector machine (SVM).

H.S. Kim et al. (2010) [8] in a research paper used data on SMEs supported on the basis of their technology scores evaluated during 1997–2002. They used variables like: stock market listing of company, age of company, venture company, external audit, foreign investment, expert manager, patents, joint company, financial ratios for fundamental valuation, management, technology, marketability, profitability. They applied their data on models like CART, MARS, discriminant analysis, logistics regression, neural networks model, support vector machine. They concluded that SVM outperformed the other methods. SVM can serve as the alternative method for the default prediction.

Ching-Chiang Yeh et al. (2012) [9] in their research work used 2470 Taiwanese high-technology companies that are traded on the Taiwan's security market. They incorporated some fundamental financial valuation ratios, proportion of shares held by board and directors, proportion of mortgage shares held by board, member, proportion of board members who are executive, proportion of independent directors on the board, proportion of shares held by the block stockholders. There predicted three ratings for each company from 2003-2008 using SVM, logistic regression, back propagation neural networks. They concluded that proposed model surpasses the listing methods in terms of both higher accuracy and fewer variables, and the output of proposed procedure is to generate a set of understandable rules that are easily interpretable model for a knowledge-based rating system.

Gang Wang et al. (2012) [10] in their research used the Australian and Germen credit datasets. Australian dataset includes 307 good customers and 383 bad customers. German dataset consists of 700 good customers and 300 bad customers. They use models like random forest, CART, SVM, rough set theory. They concluded saying that bagging decision tree is an efficient techniques for credit scoring.

Tsui-Chih Wu et al. (2012) [11] in their research used a dataset of 3069 public electronic companies from 2005-2009. They used variables fundamental financial valuation ratios and corporate governance assessment in an enhanced decision support model and SVM. The proposed model represents an advance over the relevance vector machine, whose practical application is impeded by its inability to provide comprehensive decision rules.

Christophe Mues, et al. (2004) [12] in their study used real-life credit-risk data enhanced decision support model, decision tree, support vector machine. They used datasets like German credit, Bene1 and Bene2. The Bene1 and Bene2 data sets were obtained from two major Benelux (Belgium, The Netherlands, and Luxembourg) financial institutions. They used models like neural network, multi-valued decision diagrams and found that MDD reduction mechanism was quite effective. In that, several isomorphic sub graphs were being shared which were otherwise replicated while using a decision tree representation.

Nan-Chen Hsieh (2005) [13] in their research used a German credit data set of 1000 loan applicants with 700 samples of creditworthy applicants and 300 samples where credit should not be extended. Australian credit data set was a similar data set with 690 samples, in which 468 samples were accepted and maintain good credit and 222 samples were accepted, but became delinquent. They used variables like credit history, account balances, loan purpose, loan amount, employment status, and personal information on a hybrid (clustering + neural network) technique. They found that hybrid approach is simple but efficient in the application of credit market.

W. Chen et al. (2009) [14] in their research work used a 2000 Chinese credit card data set. They used variables like personal condition such as credit history, account balances, loan purpose, loan amount, employment status, personal information, age, housing and job. They used CART, MARS and SVM. They concluded that the hybrid SVM technique not only has the best classification rate, but also has the lowest type II error in comparison with CART, MARS and SVM and justified the presumptions that SVM having better capability of capturing nonlinear relationship among variables.

Shu-Ting Luo et al. (2009) [15] in their research work used a German credit data set consisting of a set of 1000 loan applicants including 700 creditworthy applicants and 300 non-creditworthy applicants. The Australian credit data set

consisted of 690 loan applicants out of which, 307 instances were creditworthy and rest 383 instances were non-creditworthy. They used 6 nominal, 8 numeric attributes. They used SVM and clustering launched classification in their article. They found that in most of the cases, the accuracy of CLC outperforms than the others.

Shu Ling Lin (2009) [16] in his research work used a data set from 37 listed banks in Taiwan. They used variables like risk-based capital adequate ratio, tier-I capital ratio, debt to equity ratio, equity ratio, non-performing loan, observable loans to total loans, allowance for doubtful accounts recovery rate, return on equity, return on assets, net interest income ratio, profit margins, per employee profit margins, current allowance ratio, loans to deposits ratio, demand deposits ratio, deposits growth ratio, loans growth ratio, investment growth ratio, current ratio, earnings per share, interest-sensitive assets to liabilities ratio, interest-sensitivity gap to equity ratio, cash flows ratio, director pledge ratio, ownership by directors and supervisors, ownership by block-shareholders in a logistic regression model, logarithm logistic regression, ANN and two-stage models. They find that the prediction of financially sound banks, two-stage hybrid model demonstrates a comparatively stronger prediction power than the conventional ones in the prediction of financially distressed banks, with two-stage model giving the best performance of 80%.

Ligang Zhou et al. (2010) [17] in their used German credit data set consists of 700 creditworthy applicants and 300 non-creditworthy instances and a financial services company in England with 1225 applicants, of which 323 are bad and others are good. They use variables like year of birth, number of children, number of other dependents, 'Is there a home phone?', spouse's income, applicant's employment status, applicant's income, residential status, value of home, mortgage balance outstanding, outgoings on mortgage or rent, outgoings on loans, outgoings on hire purchase, outgoings on credit cards in linear discriminant analysis, diagonal linear discriminant analysis, quadratic discriminant analysis, diagonal quadratic discriminant analysis, logistic regression, probabilistic regression, decision trees like C4.5, ID3 and CART, back propagation neural network with tangent sigmoid transfer function in hidden layer, probabilistic neural networks, Bayesian classifier and k-nearest neighbor. They found that several SVM ensemble models are proposed for credit risk assessment. Even though reliability-based ensemble models cannot achieve the best prediction performance for all the datasets, they can achieve performance as good as the best, with 95% confidence level.

Sungbin Cho et al. (2010) [18] in their research used financial data consisting of 1000 Korean manufacturing firms with an asset size of US$1 million to US$7 million in the fiscal year 2000–2002. They used fundamental financial valuation variables to assess the long term performance of the company and finally used models namely decision tree, case based reasoning model and chi square in their research. They concluded that proposed CBR model outperforms the currently used approaches.

Defu Zhang et al. (2010) [19] in their research used a German credit database same as used in [15]. They used variables like credit history, account balances, loan purpose, loan amount, employment status, personal information, age, housing, and job. They used vertical bagging decision trees. They found that vertical bagging decision tree model is more accurate and more robust than other types of model.

Bee Wah Yap et al. (2011) [20] in their article used 2765 data of credit 35% was defaulters while 65% was non-defaulters. They use variables like gender, age, district, occupation, race, marital status, number of dependents, number of cars, defaulters/non-defaulters, status and work sector. They use data mining, logistic regression, decision tree. They conclude that Logistic regression model has the highest sensitivity and the lowest Type II error (a defaulter misclassified as non-defaulter). The decision tree is the worst model as it has the highest Type II error and the lowest sensitivity.

This literature review gives us a detailed idea about the profound computational risk assessment models which have been used till now. This leads us to our study on a very specific sub-section of credit which is luxury vehicle finance. It can be seen that most of the credit data sets have a large proportion of non-creditworthy applicants. However, the nature of the data set used in this study is different as compared to the ones mentioned in the above literature. This is explained in the further sections

III. DATA AND ANALYSIS METHODOLOGY

This article attempts to analyse a longitudinal survey dataset consisting of 50 luxury passenger car customers from different dealerships across India. The sample is composed of 12 ladies and 38 men. It was distributed over 22 cities/states. The data was cloned for 100 customers in order to obtain a better fit of decision trees. The credit portfolios are characterized by certain relevant variables namely 'Gender', 'Age', 'City/State', 'Job Type', 'Asset Price', 'Down Payment', 'Loan Amount', 'EMI Tenure', 'EMI', 'Interest Rate' and 'Approximate Total Income'. Certain variables have been discretized to standardize them with the rest of the numeric variables. The variable 'Gender' has been coded as 0 and 1, where 0 stands for female and 1 stands for male. The variable 'Job Type' has been coded as 1 and 2, where 1 stand for salaried job and 2 stands for business. The variable 'City/State' has been discretized in accordance to the key presented in Table I.

TABLE I. CODING FOR VARIABLE CITY/STATE

Value	City/State	Value	City/State
1	Bangalore	**12**	Kolkata
2	Chandigarh	**13**	Lucknow
3	Chennai	**14**	Mumbai
4	Chattisgarh	**15**	Nasik
5	Cochin	**16**	New Delhi
6	Coimbatore	**17**	Noida
7	Goa	**18**	Orissa
8	Gujarat	**19**	Patiala
9	Hyderabad	**20**	Pune
10	Jalgaon	**21**	Punjab

The customers are classified into 2 classes namely defaulters and non-defaulters. The data comprises of 10 % defaulters and 90 % non-defaulters. Thus, it is majorly dominated by creditworthy applicants. Thus the final data obtained is a multi-variate and multi-class data with the class variable being the loan payer characteristic.

The sample characteristics are shown in Table II.

TABLE II: DESCRIPTIVE STATISTICS

Variable	Minimum	Maximum	Mean	Standard Devaition
Age	29	66	45.97	9.050
Asset Price	2.36E6	7.66E6	3.7750E6	1.16942E6
Down Payment	0%	81.4%	25.930%	17.9071%
Loan Amount	5.00E5	7.00E6	2.8927E6	1.24446E6
Loan Tenure	36	84	55.68	15.853
EMI Amount	16560.00	1.71E5	6.2071E4	28161.55750
Interest Rate	11.50%	15.00%	12.4898%	.93273%
Approximate Total Income	1.95E5	2.86E7	2.9755E6	5.35424E6

The age group of the luxury car customers in the sample is between 29 and 66. The interest rate varies between 11.5-15%. Loan tenure varies between 36 to 84 months. The study starts with the use of certain conventional statistical techniques for analysis. The first technique is the regression analysis with the variables as mentioned above. The second technique used is Carl Pearson's partial correlation which provides us with the inter-correlation between the variables. The third one is the analysis of variance (ANOVA) which is a conventional statistical technique but lays the foundation for this study.

The fourth technique used for the analysis is machine learning. This study uses three decision tree classifiers namely 'J48 Decision Tree', 'Random Tree' and 'REP Tree', the theory and details of which have been explained in detail in section IV. The classification was done using a ten-fold cross validation in order to check the efficiency of the classifier and the respective decision trees, confusion matrices and classification accuracies of all the classifiers were obtained. The results and discussion have been discussed in section V.

IV. DECISION TREE CLASSIFIERS

A decision tree works on the basis of certain classification rules which distinguish data into various classes. The branches are the classification paths which pass through a specific node which is representative of the classification attribute. This type of decision tree learning algorithm is known as a C4.5 algorithm. (Olson, 1974) [21]

A. J48 Decision Tree

The basis for a J48 decision tree algorithm is the C4.5 model. The algorithm has two phases namely 'Building Phase' and 'Pruning Phase'.

The building phase leads to the building of the decision tree by training a sample set with attributes [21]. The C4.5 classifier uses information gain and entropy reduction as the selection criteria for classification. Entropy is defined as the measure of homogeneity in the set of examples. Information gain is defined as the reduction in entropy in a set of data caused by the partitioning of the example according to the given feature. Information *Gain(S,A)* of a feature *A* relative to a collection of examples *S*, is given by the following expression:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

Where *Values(A)* is the set of all possible values for attribute A, S_v is the subset of *S* for which feature *A* has the value *v*

$$(i.e\ S_v = \{s \in S \mid A(s) = v\})$$

Entropy is given by:

$$Entropy\ (S) = \sum_{i=1}^{c} -P_i log_2 P_i \quad (2)$$

Where P_i is the proportion of *S* belonging to class '*i*' and c is the number of classes. The second term in the equation is the expected value of entropy after *S* is partitioned, using feature *A*. The total entropy can be evaluated by adding up entropies of each subset S_v, weighted by the fraction of examples $|S_v|/|S|$ that belongs to S_v . Hence, *Gain(S,A)* is the expected reduction in entropy caused by knowing the value of feature *A*.

The 'Pruning Phase' is the phase where less reliable branches are removed in order to gain larger classification accuracy

B. Random Tree

In the random tree algorithm multiple trees are constructed simultaneously and the algorithm picks one of the remaining features randomly at each node expansion, without any purity function check (such as information gain etc.). but it chooses a feature, which is not chosen previously in a particular decision path starting from the root of tree to the current node. This is because it is has little or no effect to pick an already used feature again, on the same path. A tree stops growing any deeper if a node becomes empty or there are no more examples to split in the current node or the depth of tree exceeds some limits.

C. REP Tree

This algorithm builds a decision/regression tree using information gain/variance and prunes it using reduced-error pruning (REP) (with back fitting). Reduced error pruning is one of the simplest forms of pruning. Starting at the leaves, each node is replaced with its most popular class. If the classification accuracy is not affected then the change is kept, else a different class is chosen. Reduced error pruning has the advantage of simplicity and speed.

V. RESULTS AND DISCUSSION

The data was subjected to various methods of analysis which are presented in the following sections:

A. Regression Analysis

A non-linear regression model was built with the variables as mentioned in section IV. In the model, the log of interest rate was included as there was correlation between the interest rates and other explanatory variables.

TABLE III: FIT OF REGRESSION MODELS

Model	R	R^2	Adjusted R^2	Std. Error of Estimate
1	.452	.204	-.054	.311

The identified regression model did not have a predictive fit as seen from table III. However, it was useful in establishing the causalities between the default and the available variables. The regression coefficients are presented in table IV.

TABLE IV: REGRESSION COEFFICIENTS

Coefficients	Unstandardized Coefficients		Standard Coefficient	T	Sig.
	B	*Std. Error*	*Beta*		
(Constant)	-3.793	2.368		-1.602	.118
Gender	-.029	.118	-.041	-.246	.807
Age	.000	.006	-.024	-.131	.896
City/State	-.006	.009	-.128	-.680	.501
Job Type	**-.214**	**.127**	**-.285**	**-1.684**	**.101**
Asset Price	8.354E-9	.000	.032	.045	.964
Downpayment	-.007	.017	-.406	-.393	.696
Loan amount	-2.010E-8	.000	-.083	-.092	.927
EMI Tenure	-.002	.004	-.120	-.550	.586
EMI Amount	8.964E-7	.000	.083	.244	.808
Approximate Total Income	-7.469E-9	.000	-.132	-.755	.455
Downpayment	.000	.000	.496	.721	.476
Interest Rate	**1.770**	**.867**	**.415**	**2.043**	**.048**

B. Carl Pearson's Rank Correlation

Correlation coefficients were estimated using Carl Pearson's partial correlation. The proximity is too large to be included in this paper. However, the significant inter-correlations are presented in Table V.

TABLE V: INTER-CORRELATION COEFFICIENTS

Variable 1	Variable 2	Correlation Coefficient
Asset Price	Loan Amount	0.85
Asset Price	EMI Amount	0.70
Loan Amount	EMI Amount	0.82

The proximity matrix was estimated for similarities between variables.

C. Analysis of Variance (ANOVA)

In order to estimate the impact of the demographic factors on the default, a dichotomous variable called default was created with 0 representing non-default and 1 representing default. Analysis of variance was carried out to estimate if there is a significant influence of the available variables on default. The *F* value of corresponding variables is shown in table VI.

TABLE VI: F-VALUES IN ANOVA

Variable	F Value	Variable	F Value
Gender	.047	Down Payment	.539
Date of Birth	.458	Loan Amount	1.076
Age	.458	EMI Tenure	.005
City/State	.056	EMI Amount	.616
Job Type	1.371	Interest Rate	4.490
Asset Price	.282	Approximate Total Income	.290

D. J48 Decision Tree Algorithm

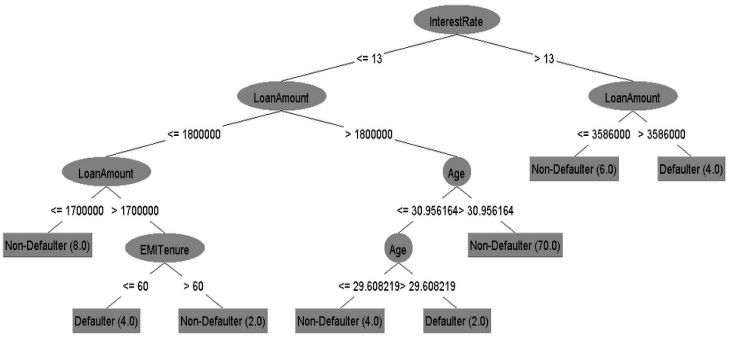

Fig. 1: J48 Decision Tree

Figure I represent the decision tree risk model using the J48 decision tree classifier. The confusion matrix for this algorithm is shown in Table VII.

TABLE VII: CONFUSION MATRIX FOR J48 CLASSIFICATION

Classified as ->	Defaulter	Non-Defaulter
Defaulter	4	6
Non-Defaulter	10	80

The confusion matrix shows that 4 defaulters are correctly classified and 6 are misclassified. Other classification results are presented in Table VIII.

TABLE VIII: CLASSIFICATION DETAILS FOR J48 ALGORITHM

Parameter	Value
Kappa Statistic	0.2453
Mean Absolute Error	0.167
Root Relative Squared Error	0.3622

E. Random Tree Algorithm

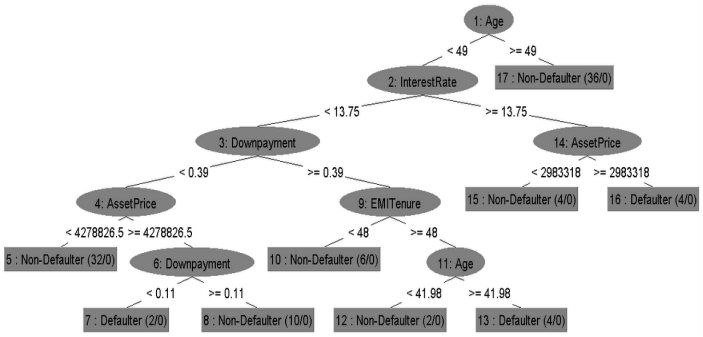

Fig. 2: Random Tree

Fig. 2 represents the decision tree risk model using a random tree classifier. The confusion matrix for this algorithm is shown in Table IX.

TABLE IX: CONFUSION MATRIX FOR RANDOM TREE

Classified as ->	Defaulter	Non-Defaulter
Defaulter	10	0
Non-Defaulter	2	88

The confusion matrix shows that all the defaulters are correctly classified. However, 2 non-defaulters are misclassified. The classification details for this analysis are shown in table X.

TABLE X: CLASSIFIATION DETAILS

Parameter	Value
Kappa Statistic	0.898
Mean Absolute Error	0.02
Root Relative Squared Error	0.1414

F. REP Tree Algorithm

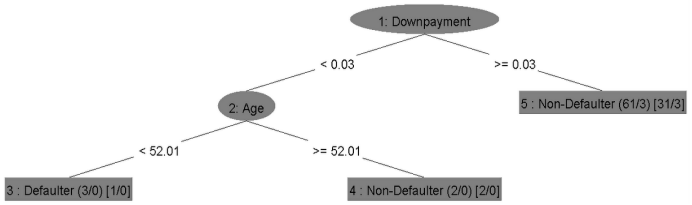

Fig. 3: REP Tree

Fig. 3 represents the REP tree risk decision models using the REP tree algorithm. The confusion matrix for this algorithm is given in table XI.

TABLE XI: CONFUSION MATRIX FOR REP TREE

Classified as ->	Defaulter	Non-Defaulter
Defaulter	2	8
Non-Defaulter	0	90

The confusion matrix in table XI shows that only 2 defaulters are correctly classified and rest 8 are misclassified. However, all the non-defaulters are correctly classified. Further classification details are presented in Table XII.

TABLE XII: CLASSIFIFCATION DETAILS FOR REP ALGORITHM

Parameter	Value
Kappa Statistic	0.3103
Mean Absolute Error	0.1491
Root Relative Squared Error	0.2762

The classification accuracies for the 3 classifiers are presented in Table XIII.

TABLE XIII: CLASSIFICATION ACCURACIES

Algorithm	Accuracy
J48	84 %
Random Tree	98 %
REP Tree	92 %

After analyzing the confusion matrices of all the three trees and considering Table XIII, it can be inferred that random tree proves to be the best tree for classification since it classifies all the defaulters correctly and has the highest classification accuracy. The top variables that classify the data in all the three are represented in table XIV.

TABLE XIV: MOST SIGNIFICANT VARIABLES

Algorithm	Best Features
J48 Tree	Interest Rate Loan Amount Age
Random Tree	Age Interest Rate Down Payment Asset Price
REP Tree	Down Payment Age

A bank or vehicle finance service company would be interested in finding out the features of loan payers that classify them under a defaulter. Thus, the combination of variables which classifies an instance under defaulters for all the three decision trees is shown in table XV.

TABLE XV: COMBINATION OF VARIABLES FOR DEFAULTERS

Algorithm	Combination
J48 Tree	Interest Rate > 13, Loan Amount > 35,86,000
	Interest Rate <= 13, Loan Amount > 18,00,000, 29.6 < Age < 30.95
	Interest Rate <= 13, 17,00,000 < Loan Amount < 18,00,000, EMI Tenure < 60
Random Tree	Age < 49, Interest Rate >= 13.75, Asset Price >= 29,83,318
	41.98 <= Age < 49, Interest Rate < 13.75, Down Payment >= 0.39, EMI Tenure >= 48
	41.98 <= Age < 49, Interest Rate < 13.75, Down Payment < 0.39, Asset Price >= 42,78,826.5
REP Tree	Down Payment < 0.03, Age < 52.21

Thus after all the types of analysis done on the dataset, we can infer the following things:

a. ANOVA results indicate that the loan amount, job type and interest rates have a significant impact on the default decisions of the luxury car customers.

b. The regression model indicates that the job type interest rates are the two most important factors that determine the default of the luxury passenger car customers.

c. The J48 decision tree signifies that interest rate, loan amount and age are the important factors that influence the default.

d. Random tree signifies that age, interest rate, asset price and down payment are the key factors for classification.

e. REP tree signifies that down payment and age are the key factors for classification.

VI. CONCLUSION

From the study conducted above, it can be inferred that gender, city, interest rates, job type, loan amount, age, asset price, down payment are the key factors that influence the risks in the luxury passenger car segment in India. The analysis that mapped most of these variables is random tree. It is evident that the customers who are younger, and who pay higher interest rate along with a high asset price with lower down payment tend to be defaulters. The credit risk analysis of the luxury passenger car customers in India requires validation of the models using various new techniques to map the customer risk profile. Thus, this study was successful in finding the significant variables and their combinations which classify them as defaulters or non-defaulters.

REFERENCES

[1] L. G. Kabari and E. O. Nwachukwu, "Credit Risk Evaluating System Using Decision Tree – Neuro Based Model," International Journal of Engineering Research & Technology, vol. 2 issue 6, June 2013.

[2] Hian Chye Koh, Wei Chin Tan,and Chwee Peng Goh, "A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques", International Journal of Business and Information, vol. 1 issue 1, pp 96-118, 2006.

[3] Raquel Florez-Lopez, "Credit risk modelling facing complex datasets. A computational approach based on cooperative decision forests and bootstrapping strategies", Knowledge-Based Systems, vol33, pp. 166–172, 2012.

[4] Cheng-Lung Huang, Mu-Chen Chen, Chieh-Jen Wang, "Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications, vol. 33, pp 847-856, 2007.

[5] Junni L. Zhanga, Wolfgang K. Härdle, "The Bayesian Additive Classification Tree applied to credit risk modelling", Computational Statistics and Data Analysis, vol 54, pp 1197-1205, 2010.

[6] Stjepan Oreski, Dijana Oreski, Goran Oreski, "Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment", Expert Systems with Applications, vol. 39, pp 12605-12617, 2012.

[7] Tian-Shyug Lee, Chih-Chou Chiu,Yu-Chao Chou, Chi-Jie Lu, "Mining the customer credit using classification and regression tree and multivariate adaptive regression splines", Computational Statistics & Data Analysis, vol. 60, pp 1113-1160, 2006.

[8] Hong Sik Kim, So Young Sohn, "Support vector machines for default prediction of SMEs based on technology credit", European Journal of Operational Research, vol. 201, pp 838-846, 2010.

[9] Ching-Chiang Yeh, Fengyi Lin, Chih-Yu Hsu, "A hybrid KMV model, random forests and rough set theory approach for credit rating", Knowledge-Based Systems, vol. 33, pp 136-172, 2012.

[10] Gang Wang , Jian Ma, Lihua Huang , Kaiquan Xu, "Two credit scoring models based on dual strategy ensemble trees", Knowledge-Based Systems, vol. 26, pp 61-68, 2012.

[11] Tsui-Chih Wu, Ming-Fu Hsu, "Credit risk assessment and decision making by a fusion approach", Knowledge-Based Systems, vol. 35, pp 102-110, 2012.

[12] Christophe Mues, Bart Baesens, Craig M. Files, Jan Vanthienen, "Decision diagrams in machine learning: an empirical study on real-life credit-risk data", Expert Systems with Applications, vol. 27, 257-264, 2004.

[13] Nan-Chen Hsieh, "Hybrid mining approach in the design of credit scoring models", Expert Systems with Applications, vo. 28, pp 665-665, 2005.

[14] Weimin Chen , Chaoqun Ma, Lin Ma, "Mining the customer credit using hybrid support vector machine technique", Expert Systems with Applications, vol. 36, pp 7611-7616, 2009.

[15] Shu-Ting Luo , Bor-Wen Cheng , Chun-Hung Hsieh, "Prediction model building with clustering-launched classification and support vector machines in credit scoring", Expert Systems with Applications, vol. 36, pp 7562-7566, 2009.

[16] Shu Ling Lin, "A new two-stage hybrid approach of credit risk in banking industry", Expert Systems with Applications, vol. 36, pp 8333-8341, 2009.

[17] Ligang Zhou , Kin Keung Lai , Lean Yu, "Least squares support vector machines ensemble models for credit scoring", Expert Systems with Applications, vol. 37, pp 127-133, 2010.

[18] Sungbin Cho, Hyojung Hong, Byoung-Chun Ha, "A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance: For bankruptcy prediction", Expert Systems with Applications, vol. 37, pp 3482-3488, 2010.

[19] Defu Zhang, Xiyue Zhou, Stephen C.H. Leung, Jiemin Zheng, "Vertical bagging decision trees model for credit scoring", Expert Systems with Applications, vol. 37, pp 7838-7843, 2010.

[20] Bee Wah Yap, Seng Huat Ong , Nor Huselina Mohamed Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models", Expert Systems with Applications, vol. 38, pp 13274-13283, 2011.

[21] Olson, C. L., "Comparative robustness of six tests in multivariate analysis of variance", Journal of the American Statistical Association, vol, 69, pp 19894-19908, 1974.