Predictive Lending AI

Kommaraju Rupa Kusuma[1],Dr. Manne Suneetha[2].

[1]M.Tech Student, Data Science, Dept. of Information Technology, V R Siddhartha EngineeringCollege,Kanuru,India(Email:rupakommaraju23@gmail.com)

[2]Head of the Dept, Professor, Dept. of Information Technology, V R Siddhartha EngineeringCollege,Kanuru,India(Email:hodit@vrsiddhartha.ac.in

ArticleInfo
Page Number: 335-344
PublicationIssue:
Vol.72 No. 1 (2023)

ArticleHistory
Article Received: 12 October 2022
Revised: 24 November 2022
Accepted: 18 December 2022

***Abstract*:** The product Predictive Lending AI is a AI based Automatic Lending Solutionfor Non-Banking Financial Company (NBFC). It helps a global lending firm to reduce theloan processing time by 40%, minimize business risk and increase revenues with effectiveup-sellingoflendingproducts.Therearethreeuse cases(Applicationscoring,Defaulter Prediction and Churn Prediction). The first use case is Application Scoring inwhich we calculate a score based on the input and approve a loan based on that score. Thesecond use case is about Loan Defaulter Prediction. The person who takes loan from anorganization and doesn't repay the loan amount is called Loan defaulter. The third usecase is about Churn Prediction. Churn is defined as the movement of customers from oneprovidertoanother.

Keywords:Prediction,Non-BankingFinancialCompany(NBFC),Defaulters,ApplicationScoring,Churn Prediction.

Introduction

As, the count of the people applyingfor theloans has been increasingforvariousreasons in the recent years. The bank employees are unable to analyze or predict whether thecustomer can payback the amount or not (good customer or bad customer) for the given interestrate. In order to succeed in the stream of banking, one has to have an idea about the behavioralpatterns of variouscustomersbasedon theirtransaction history.Thisiswhat,ourmodelLending AI is doing, by predicting the cases of customers who may finally end up as a loandefaultornot.To predict the credit default, several methods have been proposed. The use of methoddepends on the complexity of banks and financial institutions, size, and type of the loan. Thecommonly used method has been discrimination analysis. By using scoring models that are AI-basedandusedeeplearning,banksandfinancialinstitutionscanaccessmorerealisticpredictions on credit risk, using customers' credit history and the power of big data. This waycredit can be approved to the right people and better pricing options offered to people whodeserve it. The output of the model will generate a binary value that can be used as a classifierthat will help banks to identify whether the borrower will defaultor not default. As the finalstepinthedirection,linear regressionmethod isalsogoingtobeperformedonthedataset.

RelatedWorks

InthepreviousversionstheframeworkisdevelopedtoeffectivelyidentifytheProbability of Defaultof a Bank Loan applicant. The metrics derived from the predictionsreveal the high accuracy and precision of the built model. The model proposed in an effectivepredictionmodelforpredictingthecrediblecustomerswhohaveappliedforbankloan.DecisionTree isappliedtopredicttheattributesrelevantforcredibility.

This prototype model can be used to sanction the loan request of the customers or not.The model proposed in has been built using data from banking sector to predict the status ofloans. This model uses three classification algorithms namely j48, Bayes Net and naive Bayes.The model is implemented and verified using Weka. The best algorithm j48 was selected basedon accuracy. An improved Risk prediction clustering Algorithm that is multi-dimensional isimplementedintodetermine badloanapplicants.

In this work, the Primary and Secondary Levels of Risk assessments are used and toavoid redundancy,Association Rule is integrated. In a decision tree model was used as aclassifierandforfeatureselectiongeneticalgorithmisused.ThemodelwastestedusingWeka.

SupportVectorMachine,DecisionTree,LogisticRegression,NeuralNetwork,Perceptronmodel,allthesetechniquesarecombinedinthismodel.Theeffectivenessofapplying the above techniques on creditscoringis studied.The analysis results show theperformance is outstanding based on accuracy. The aim of the study in is to introduce a discretesurvival model to study the risk of default and to provide the experimental evidence using theItalianbankingsystem.

Data mining in banking Due to tremendous growth in data the banking industry dealswith,analysisandtransformationofthedataintousefulknowledgehasbecomeataskbeyondoutstanding based on accuracy. The aim of the study in is to introduce a discrete survival modelto study the risk of default and to provide the experimental evidence using the Italian bankingsystem.

Data mining in banking Due to tremendous growth in data the banking industry dealswith, analysis and transformation of the data into useful knowledge has become a task beyondhuman ability. Data mining techniques can be adopted in solving business problems by findingpatterns,associations and correlations which are hidden in the businessinformation stored inthe databases.

ProposedMethodology

In this paper different models like Random Forest, Logistic Regression and DecisionTree are built on the dataset. The input data is loaded from the database that is stored and themodel metrics should be identified such as whether the data is structured or unstructured andunderstanding of features. The data analysis includes what are the important features that arerequired for prediction and any features that are should be imputed or removed, etc. And thetarget variable should be identified. Next, the model is built by choosing the right Algorithm forprediction. The model should be able to predict for the incoming new data. After the model isbuilt the leads that are qualified. To increase the qualified leads, we have again feed the modelwithdataandcanincrease theperformance.

The Historical data from NBFC's is used to build a model in the RapidMiner tool. Aftercompleting EDA, a model is built using ML algorithms and predicts the results for unseen datausingthatmodel.

RapidMiner tool is used for providing a solution to this use case. RapidMiner is a datasciencesoftwareplatform developedby thecompany of thesamenamethatprovidesanintegrated environment for data preparation, machine learning, deep learning, text mining, andpredictive analytics. RapidMiner's data science platform delivers lightning-fast business impactforover 40,000+organizations ineveryindustrytodriverevenue, reducecosts,andavoidrisks.

Fig.1. ArchitectureDiagram

The product Lending AI is an AI based Automatic Lending Solution for Non-BankingFinancial Company (NBFC).Ithelps a global lendingfirm to reduce theloan processing timeby 40%, minimize business risk and increase revenues with effective up-selling of lendingproducts.

A. NBFC

A Non-banking financial company or non-banking financial institution is a financialinstitution thatdoes not have a full banking license or is notsupervised by a national orinternational bankingregulatory agency.Itisaninstitution,whichisacompany andhasprincipal business of receiving deposits under any scheme or arrangementin one lump sum orin instalments by way of contributions or in any other manner, is also a non-banking financialcompany(Residuarynon-bankingcompany).

B. EDA

EDA is a phenomenon under data analysis used for gaining a better understanding of dataaspects like, it identifies the main features of the data and also it identifies which variables areimportantforourproblem

C. DataPre-processing

Machine Learning algorithms don't work so well with processing raw data. Before wefeed such data to an ML algorithm, we must pre-process it. In other words, we must apply sometransformations on it. With data pre-processing, we converted raw data into a clean data set.Some ML models need information to be in a specified format. To pre-process data, we will usethe libraryscikit-learnorsklearn.

D. DataVisualization

Data visualization is the discipline of trying to understand data by placing it in a visualcontext so that patterns, trends and correlations that might not otherwise be detected can beexposed. Python offers multiple great graphing libraries that come packed with lots of differentfeatures are Histograms, Bar Charts, Scatter Plots, Using Seaborn, Horizontal Bar charts, Staticmaps,Networkdiagrams.

DataCleaningAndPreparation

A. DataDescription

The dataset from kaggle.com is used for statistical modelling. The dataset is divided intotraining dataset and testing dataset. The Training dataset contains of 4,96,307 examples with 1special attribute and 146 regular attributes. The Testing dataset contains of 55,143 exampleswith146regularattributes.

B. DataPreprocessing

In this stage, identified the attributes with more than 50% missing values and removedthose attributes. There are many observations they are there are 42 attributes with more than50% missing values, there are 6 attributes which have a greater number of values, the targetlabel loan status has a high-class imbalance problem, there are no duplicate data points in thegivendataset.

C. Data Cleaning

I) ImputingMissingvalues

In this process the missing values in Numerical attributes are replaced by mean and themissingvaluesinNominalattributes arereplacedbymode.

II) ConvertingNominaltoNumerical

In this process the Nominal attributes are converted to Numerical attributes by dummyencoding.

D. DataSampling

In case of splitting the data, Stratified sampling technique is used. Stratified sampling isa type of sampling method in which the total population is divided into smaller groups or stratatocomplete the sampling process. The strata is formed based on some common characteristicsin the population data. After dividing the population into strata, the researcher randomly selectsthe sampleproportionally.

E. ModelBuilding

After splitting the data in the 70:30 ratio, the training data is sent to the classificationmodel. The model which is build is applied on the testing data. Then check the actual resultswithpredicted values.Keep onchanging the attributestillbestperformanceisachieved.

Once when the best performance is achieved, then we can predict the results on theunseendata.

F. ExperimentalSetup

TheAttributeselectionisdone,andtheAttributeselectionreducesmemoryrequirements andincreases the accuracy of the model. The random Forest, Decision Tree,Logistic regression, Support Vector Machine algorithms are applied on the same dataset forbuilding the model, and then the comparison is done between the four algorithms for identifyingthe bestapproach.

I) RandomForest

The most important aspect of the random forest algorithm is the variable importanceranking. It creates recursive partitioning trees using a majority vote. A number ‘m’ is specifiedwhich is much smaller than the total number of attributes. At each node, ‘m’ variables areselectedatrandomoutofthetotalnumberofattributes,andthensplitisperformed.

II) DecisionTree

The Recursive binary splitting technique can be used to perform split at a node. In thismethod, all the attributes are taken into consideration and various split points are tried andtested.Theyaretestedusing a costfunctionand the splitwiththebestcost isselected.

III) LogisticRegression

A logistic regression is run using each variable against the binary target variable for theresult. ROC curve for each variable is plotted. The variable containing the largest area under thecurve has the largestrelevancy and contributes the mostfor the result. The feature containingthe largest Information gain ratio has the lowest importance. The subset of optimal features isarrangedindescendingordertoobtainthehighestrelevancyfeaturesofthe dataset

IV) SupportVectorMachine

SVM is supervised machine learning model with learning algorithms which examine thedata and uses that data for regression and classification. This model uses a technique namely akernel trick to transform the data and based on these transforms of data, itfinds the bestoptimum results. Itis notconsidered as better as than the othermachine learning modelsbecauseitworksonless data set.

ResultandAnalysis

In this we calculate the results of predictions on all the four algorithms that I have used on thedataset in model building. The dataset is split into the training dataset and testing dataset in the70:30 ratio. The training dataset consists of the 70 % of the data from the dataset and the testingdataset consist of the 30% of the data from the dataset. The performance is calculated based onthe F1- Score, Gini Index, Accuracy, ROC. The f1- Score of the Logistic Regression, randomforest, DecisionTree,SupportVector Machineare96.1%,92.5%,86.5%,89.1%respectively.

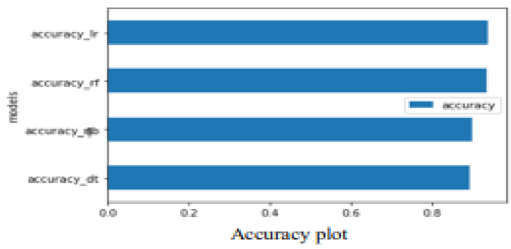

Fig.2.AccuracyPlot

Conclusion

The Logistic Regression is the best approach with 96% F1- Score and great accuracy. Thedefaulter prediction is used finally on applicants' data and the loan defaulters are identified andthe loan amount is safeguarded. The defaulter prediction use-case prevents the banking sectorfrom hugelossanddownfallby beingalertbeforesanction theloan amounttotheloandefaulters.This use-caseisveryuseful.

accuracy: 99.00%

	true Approved	true Rejected	class precision
pred. Approved	30491	541	98.26%
pred. Rejected	25	25559	99.90%
class recall	99.92%	97.93%	

Fig.3. ConfusionMatrix

References

1. Altman,E.I,P.Narayanan,AnInternationalSurveyofBusinessFailureClassificationModels,FinancialMarkets,Instruments andInstitutions,1997
2. Altman,E.I,SaundersA,Creditriskmeasurement:Developmentoverthelast20years,JournalofBankingandFinance,1998,20:1721~1742
3. ShiXiquan,ZouXinyue,TheApplicationofcanonicaldiscriminate analysisincreditriskevaluationofenterprise,The StudyofFinance andEconomics,2001,27（10）：53~57
4. ZhangAimin,ZhuChunshan,XuDanjian,Principlecomponentspredictionmodelandtheempiricalstudyoffinancialfailureofpublic company,JournalofFinance,2001,3：10~25
5. A.B.Hussain,andF.K.E.Shorouq,“Creditriskassessmentmodelfor Jordaniancommercial banks: Neuralscoring approach”, Review of Development Finance, Elsevier, vol.4,pp.20–28, 2014.
6. A. Blanco, R. Mejias, J. Lara, and S. Rayo, “Credit scoring models for the microfinanceindustry using neural networks: evidence from Peru”, Expert Systems with Applications, vol.40,pp.356–364, 2013.
7. T. Harris, “Quantitative creditrisk assessment using support vector machines: BroadversusNarrow default definitions”, Expert Systems with Applications, vol. 40, pp. 4404–4413,2013.
8. A. Abhijit, and P.M. Chawan, “Study of Data Mining Techniques used for Financial DataAnalysis”, International Journal of Engineering Science and Innovative Technology, vol. 2(3),pp.503-509, 2013.
9. D. Adnan, and D. Dzenana, “Data Mining Techniques for Credit Risk Assessment Task”,inProceedings of the 4th International Conference on Applied Informatics and ComputingTheory(AICT'13),2013,p. 105-110.
10. G.Francesca,“ADiscrete-TimeHazardModelforLoans:SomeEvidencefromItalian
11. J H Cheon, D Kim, Y Kim and Y Song, "Ensemble Method for Privacy-Preserving LogisticRegressionBasedonHomomorphic Encryption",*IEEEAccess*,vol.6,pp.46938-46948,2018.
12. S Guo, H He and X Huang, "A Multi-Stage Self-Adaptive Classifier Ensemble Model WithApplication inCreditScoring",*IEEE Access*, vol.7,pp.78549-78559,2019.
13. Dayu Xu, Xuyao Zhang, Junguo Hu and Jiahao Chen, "A Novel Ensemble Credit ScoringModelBasedonExtremeLearningMachineandGeneralizedFuzzySoftSets",*Hindawi MathematicalProblemsinEngineering*,2020.

14. S.M.SandR.S.T,"LoanCredibilityPredictionSystemBasedonDecisionTreeAlgorithm",*InternationalJournalofEngineeringResearchand*,vol.V4,no.09,2018.
15. P. M. Addo, D. Guegan and B. Hassani, "Credit Risk Analysis Using Machine and DeepLearningModels",*SSRNElectronicJournal*,2018.
16. S Tabik, R.F. Alvear-Sandoval, M.M. Ruiz, J.L. Sancho-Gómez, A.R. Figueiras-Vidal and F.Herrera, "A Tutorial On Ensembles And Deep Learning Fusion With Mnist As GuidingThread:AComplexHeterogeneousFusionSchemeReaching10Digitserror",*arxiv.org*,January2020.
17. https://docs.rapidminer.com/latest/studio/getting-started/ui-overview.html
18. A.SaxenaandM.Prasad,AreviewofclusteringtechniquesanddevelopmentspublicationNeurocomputing, Elsevier, 2017
19. MAla'rajandM.FAbbod,"Classifiersconsensussystemapproachforcreditscoring",*Knowl.-Based Syst*,vol.104, pp. 89-105,Jul. 2018.
20. AdnanDželihodžićandDženanaĐonko,"ComparisonofEnsembleClassificationTechniques and Single Classifiers Performance for Customer Credit Assessment",*ModelingofArtificial Intelligence*,vol.11,no.3, pp.140-150,2018.
21. OkokpujieOkesola, O.Kennedy, Adewale and Samuel N. John, "An improved BankCreditScoringModelANaïveBayesianApproach",*InternationalConferenceonComputationalScienceandComputationalIntelligence*,pp.228-233,2017.
22. H He, W. Zhang and S. Zhang, "A novel ensemble method for credit scoring: Adaption ofdifferentimbalanceratios", *ExpertSyst.Appl*,vol.98,pp.105-117,May2018.
23. MI Wang, H Wang, J Wang, H Liu, R Lu, T Duan, et al., "A novel model for malariapredictionbasedonensemblealgorithms",*PLOSOne*,2019.
24. N. Mohammad and A. Onni, "Credit RiskGrading Model and Loan Performance OfCommercialBanksIn Bangladesh",*EuropianJournal of Business Management*,vol.7,no.13,Yuan Sun,WeifengJian,YufengFu, "Anewperspectiveofcreditscoringforsmall andmedium-sizedenterprisesbasedoninvoicedata",IEEEAcess,2021
25. Wenyu,Zhang,"ANovelMulti-StageEnsembleModelWithaHybridGeneticAlgorithmforCreditScoringonImbalancedData",IEEEAcess,2021
26. AdelHassan,RashidJayousi,"FinancialServicesCreditScoringSystemUsingDataMining",IEEEAcess,2020
27. ASafiyaParvin,BSaleena,"AnEnsembleClassifierModeltoPredictCreditScoring",
28. IEEEAcess,2020
29. DawnIrisCalibo,MelvinA.Ballera,"VariableSelectionforCreditRiskScoringonLoanPerformanceUsingRegressionAnalysis",IEEEAcess,2019
30. OlatunjiJ.okesola,KennadyO.Okokpujie,deyinkaA.Adewale,"AnImprovedBankCreditScoringModel",IEEEAccess,2018
31. WeisongChen,LiangShi,"CreditscoringwithF-scorebasedonsupportvectormachine",IEEEAccess,2018
32. Radha Vadala, Bandaru Rakesh kumar, "An application of Credit Scoring in E-

LendingPlatform”,IEEEAccess,2019

33. Syed Zamil Hasan Shoumo, Mir IshrakMaheerDhruba, Sazzad Hossain, “Application ofMachineLearninginCreditRisk Assessment”,IEEEAccess,2019
34. YashaswiRajSureshKumar,BrijeshKumarSingh“LoanRiskPredictionUsingTransactionInformation”JournalofAnalysis and Computation(JAC),2021
35. B Spoorthi, swethaS.Kumar , “Comparative Analysis of Bank Loan Defaulter PredictionUsingMachineLearning Techniques”,IEEE,2021
36. Jaya Sinha, Rani Asthya, “Machine Learning based Loan Allocation Prediction System forBankingSector”,IEEE, 2021
37. P.Maheswari,CH.V.Narayana,“PredictionsofLoanDefaulter-ADataSciencePerspective”,IEEE,2020
38. ApurvaDatkhile,KomalChandak, “StatisticalModellingonLoanDefaultPredictionUsingDifferentModels”,InternationalJournalofResearchinEngineering,ScienceandManagement,2020
39. AbhishekShivanna,DharmaPAgrawal,“PredictionofDefaultersusingMachineLearning onAzureML”,IEEE, 2020
40. Hafiz Ilyas Tariq, AsimSohail, “Loan Default Prediction Model Using Sample, Explore,Modify,Model,andAssess(SEMMA)”,JournalofComputationalandTheoreticalNanoscience,2019
41. T. Harris, “Quantitative credit risk assessment using support vector machines: Broadversus Narrow default definitions”, Expert Systems with Applications, vol. 40, pp. 4404–4413,2013.
42. A. Abhijit, and P.M. Chawan, “Study of Data Mining Techniques used for Financial DataAnalysis”, International Journal of Engineering Science and Innovative Technology, vol. 2(3),pp.503-509, 2013.
43. D. Adnan, and D. Dzenana, “Data Mining Techniques for Credit Risk Assessment Task”,in Proceedings of the 4th International Conference on Applied Informatics and ComputingTheory (AICT '13), 2013, p. 105-110. [1] D. J. Hand, G. Blunt, M. G. Kelly and N. M. Adams,"DataMiningForFunand Profit}",*StatisticalScience*,vol.15,pp.111-131,2020.
44. Dayu Xu,XuyaoZhang,JunguoHu and JiahaoChen,"A Novel Ensemble CreditScoringModelBasedonExtremeLearningMachineand Generalized Fuzzy SoftSets",*HindawiMathematicalProblems inEngineering*,2020.
45. S Tabik, R.F. Alvear-Sandoval, M.M. Ruiz, J.L. Sancho-Gómez, A.R. Figueiras-Vidaland F. Herrera, "A Tutorial On Ensembles And Deep Learning Fusion With Mnist As GuidingThread:AComplexHeterogeneousFusionSchemeReaching10Digitserror", *arxiv.org*,January2020.
46. Yuan Sun, Weifeng Jian, Yufeng Fu, “A new perspective of credit scoring for small andmedium-sizedenterprisesbasedoninvoice data”,IEEEAcess,2021
47. Wenyu,Zhang, “ANovelMulti-StageEnsembleModelWithaHybridGeneticAlgorithmforCreditScoringonImbalancedData”,IEEEAcess,2021
48. Adel Hassan, Rashid Jayousi, “Financial Services Credit Scoring System Using

DataMining",IEEEAcess,2020
49. A Safiya Parvin, B Saleena, "An Ensemble Classifier Model to Predict Credit Scoring",IEEEAcess,2020