Connectingdot Consutancy – A brief introduction

Connecting Dot

About CDPL

CONNECTINGDOT CONSULTANCY (CDPL) , headquartered in Kolkata is a privately **owned, debt free, profitable and cash positive** company where we provide platform based financial risk consultancy and training services to FS clients and institutions. Praloy Majumder, who is the founder of CDPL is also the founder and director of Disseminare Consulting which has huge experience of training banking professionals in India and Bangladesh

OUR PLATFORM

LADA
LOAN & ADVANCES DATA ANALYSIS

-LADA stands for Loan and Advances Data Analytics.
- ML/AI based platform which helps in prediction of risk categories though risk scoring of applicants or transactions

OUR VISION

-Make Credit accessible to unserved and underserved but deserving borrowers
-Solving the issue of high NPAs, particularly in the fast-growing retail, agriculture and MSME segments

OUR Co-FOUNDERS

Praloy Majumder
Experience - 25 years
Banks - Syndicate Bank and ICICI Bank
MBA – IIM Calcutta
Visiting Faculty – RBI College of Supervisors, IIM Calcutta & many Tier 1 MBA colleges
Trained 15000+ professionals

Soumya Dasgupta
Experience - 16 years
Firms – Accenture, Cognizant, TCS, Iris
MBA – IIM Calcutta
Visiting Faculty – IIM Lucknow & many other Tier 1 MBA colleges
***Certificates**: FRM , TOGAF*

OUR JOURNEY

2020
- Internal development of first model for Salaried and SENP Segment

2021
- LADA Go-live for first Indian NBFC client in Sourcing
- Enhancement of modelling techniques and software

2022
- Deployment of EWS for a large bank in Bangladesh
- Advanced data security addition
- Explainibility technique addition

2023
- UAT completion for second Indian NBFC client in credit sourcing
- Project start for CGTMSE

Next Steps
- Create a separate section 8 entity in India
- Integrating AA framework for enrichment of customer data used for model creation
- Collaboration with iSpirt

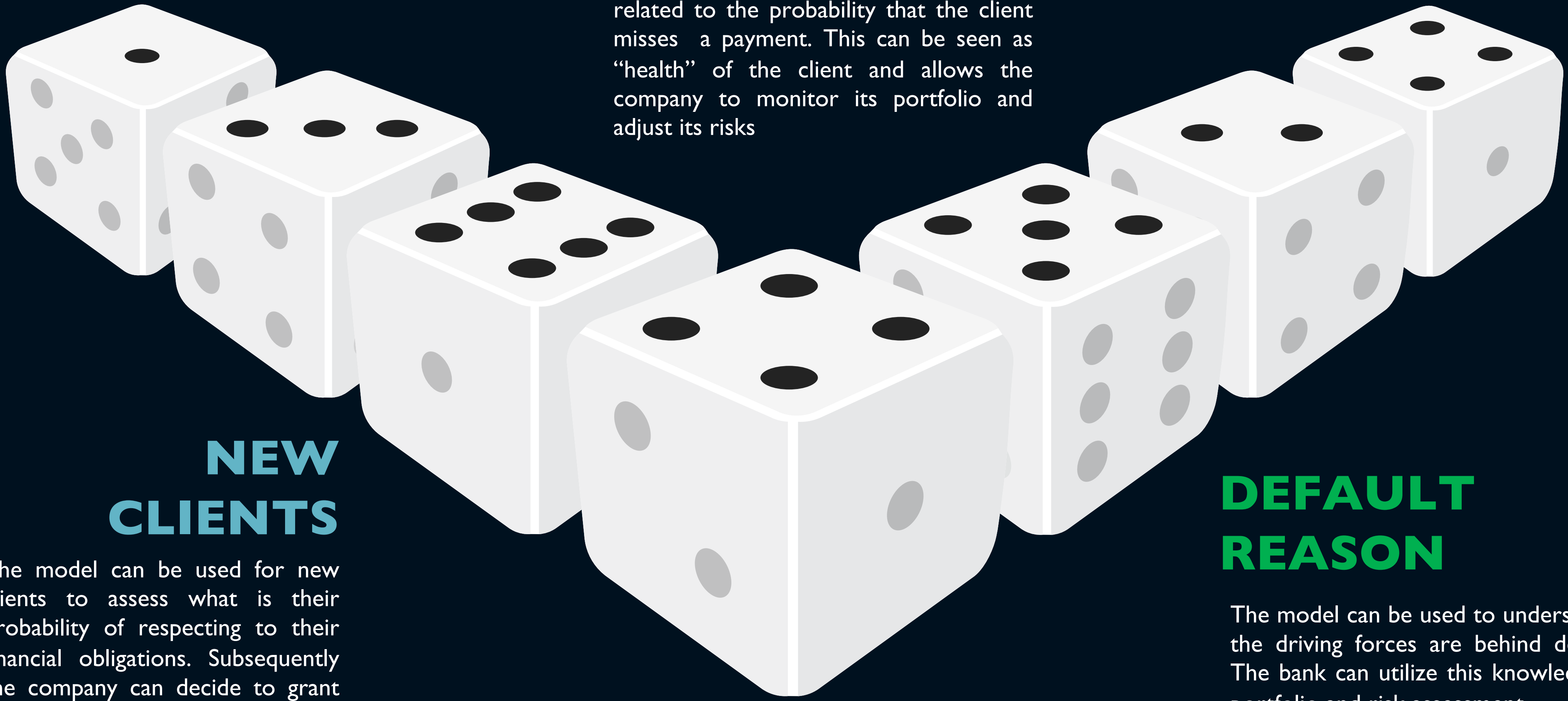
Risk Prediction Platform (LADA)

Connecting Dot

LADA Use Case: Credit Scoring Using Data Analytics

HEALTH SCORE

The model provides a score that is related to the probability that the client misses a payment. This can be seen as "health" of the client and allows the company to monitor its portfolio and adjust its risks



NEW CLIENTS

The model can be used for new clients to assess what is their probability of respecting to their financial obligations. Subsequently the company can decide to grant or not the requested loan

DEFAULT REASON

The model can be used to understand what the driving forces are behind default are. The bank can utilize this knowledge for its portfolio and risk assessment

Model USP

Innovative fundamental concept involving Cost of default and Benefit of default of borrowers

Domain driven variable selection , fine tuned by data science

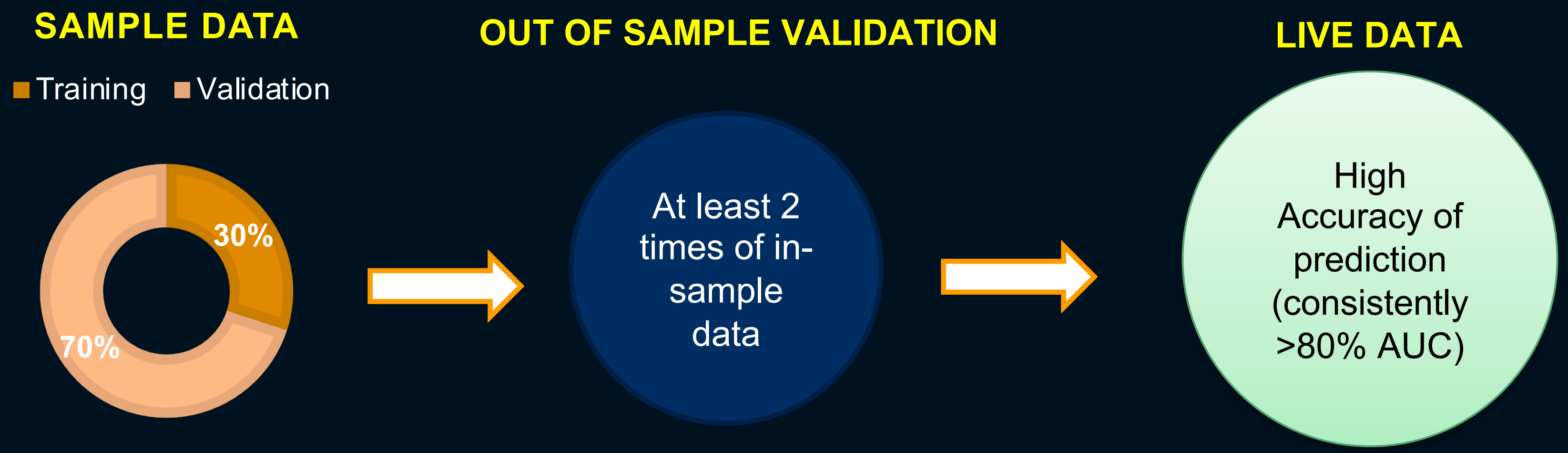
Segment specific unique model customized to quintessential banking needs

- Industry wise /Region wise/ Promoter wise
- Perception of cost of default and benefit of default is different

Ability to create models with scarce data.Transactional data used to derive behavioral data patterns and perform missing data treatment

Novel approach of creating model on 30 percent of sample data and do testing on 70% of sample data.

Ability to create models with very few financial variables and still proving high model accuracy



SAMPLE DATA

OUT OF SAMPLE VALIDATION

LIVE DATA

At least 2 times of in-sample data

High Accuracy of prediction (consistently >80% AUC)

Use of advanced statistical models (parametric and non-parametric) : SVM, KNN, LR

LADA Utility

Connecting Dot

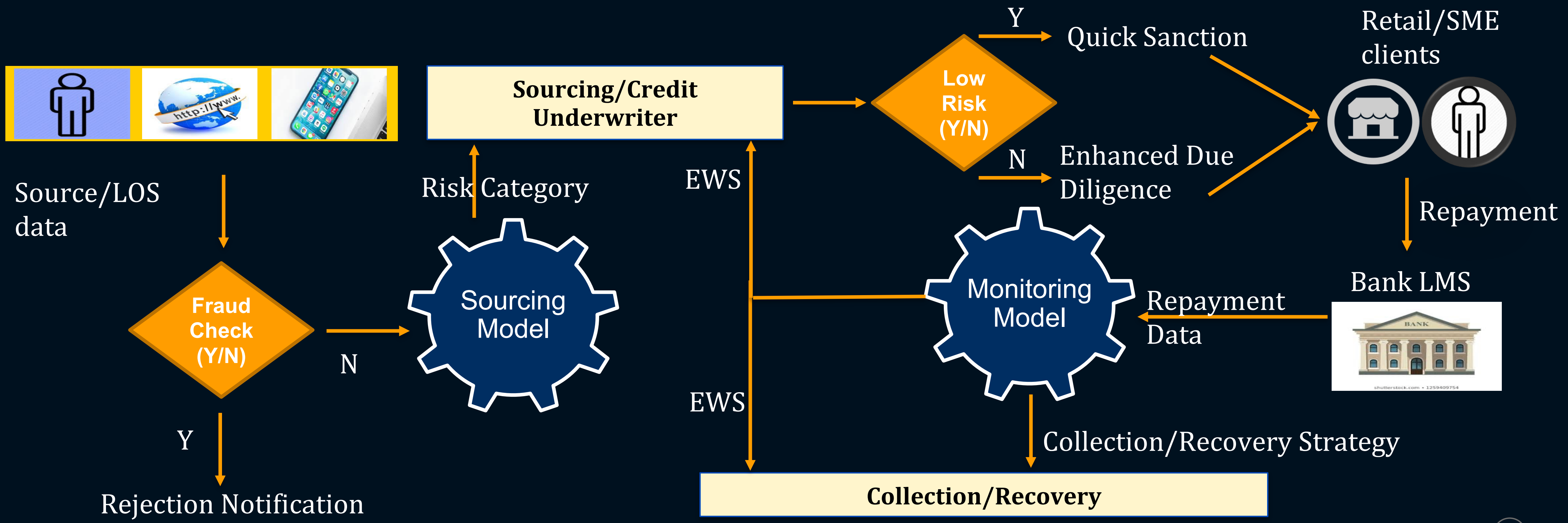
Classify the customer loan portfolio into high , medium and row risk category with high degree of accuracy

Can be readily used for quick loan sanction as low-risk bad rate (defined as 30 DPD+once in loan lifecycle for sourcing model) has been at least 4 times less than high risk bad rate

Input fraud data handling through domain driven intelligent solution

Facility/Borrower monitoring for non repayment (30 DPD+/60 DPD +) and generating Early Warning signals

Define collection and recovery strategy for stressed assets



Case Study 1 : Early Warning Signals - Credit Card Transactions

Background

- A top bank in Bangladesh had more than 1.2 lakh credit card customer accounts. The bank was targeting to increase the credit card penetration in Bangladesh but were having the following pain points:
- Absence of any framework to categorize credit card customers into risk categories
- Inability to create strategy for collections team by assigning priority of collections on the basis the customers' transactional patterns
- Lack of score-based decision making for Credit Limit Increase/Decrease for customers

Solution Provided

- - Developed a comprehensive framework for categorization of customers into three EWS categories namely "High Risk ", "Medium Risk "and "Low Risk "
- - Used non-parametric method of supervised learning using Machine Learning (ML) methodology to segregate customers
- - Model created over small data set and tested over 15 times larger data set in UAT

  Model creation set ~5000 customer accounts
  Model validation set ~15000 customer accounts
  UAT phase I data set ~54000 customer accounts
  UAT phase II data set ~74000 customer accounts

- High AUC during Model creation and validation was achieved as ~97% and ~84% respectively
- VAPT (security) standards of client maintained in solution
- Developed a robust and automated software solution which run remotely at the end of every month
- Developed reporting dashboard with risk categorization of customer accounts and also publishing collection strategy for collections team
- Solution running in live for last 16 months from Jan 2022 without bug

Output

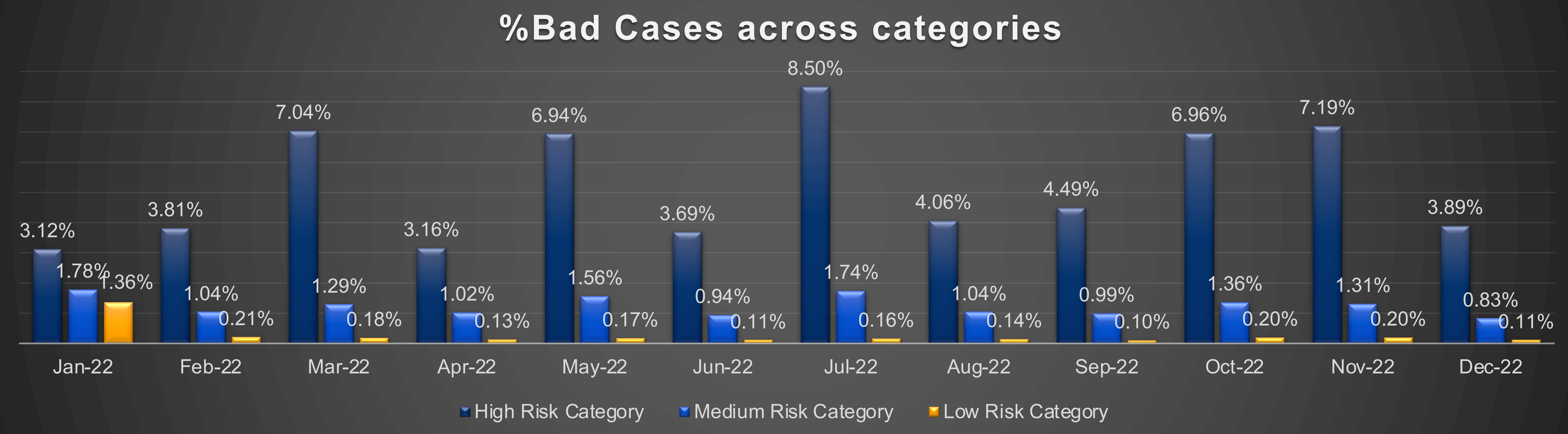
- - High level of accuracy in predicting risky accounts over ~74000 UAT data. The category defined through model as
- High Risk segment (11011 accounts) had 85.53% actual bad cases ,
  Medium Risk segment(31238 accounts) had 43.14% bad cases
  Low Risk segment (31682 accounts) had just 9.5% bad cases (The overall population had bad case around 35%)
- - Model has hold consistent over large out of time data of ~15.5 lacs for a period of 12 months in live
- - Considerable reduction of bad rate in low-risk category - Bad rate in low risk category is just around 0.1%-0.2% in live
- - Streamlined collection in bank through automated assignment of personnel
- - Bank is able to leverage the automated collection prioritization generated from the software

- Client has initiated LADA project for other Lines of Business such as Personal Loan and SME Loan

Please refer to next slide for definition of bad case and looking at the model results in live post Jan 2022

Case Study 1 : EWS Performance post Go-Live

- Bad Case definition = 30 DPD + on current outstanding
- Model ran with end-of-month application data for January to December 2022



%Bad Cases across categories

Jan 2022	Feb 2022	Mar 2022	Apr 2022	May 2022	June 2022
Total =121077	Total =119564	Total =120909	Total =122294	Total =122997	Total =125922
High Risk=2915	High Risk=2388	High Risk=2572	High Risk=2724	High Risk=2754	High Risk=2545
Med Risk=73510	Med Risk=69720	Med Risk=71040	Med Risk=72309	Med Risk=71657	Med Risk=70835
Low Risk=47834	Low Risk=49469	Low Risk=49961	Low Risk=49145	Low Risk=51046	Low Risk=52542

July 2022	Aug 2022	Sep 2022	Oct 2022	Nov 2022	Dec 2022
Total =131955	Total =132915	Total =135962	Total =139573	Total =141030	Total =142502
High Risk=2859	High Risk=2562	High Risk=2896	High Risk=3061	High Risk=3074	High Risk=3165
Med Risk=74728	Med Risk=74842	Med Risk=76713	Med Risk=78883	Med Risk=79621	Med Risk=80473
Low Risk=54368	Low Risk=55511	Low Risk=56353	Low Risk=57629	Low Risk=58335	Low Risk=58864

Case Study 2 : Housing Loan Model for Sourcing

Connecting Dot

Background

An Indian NBFC wants to register healthy and sustainable business growth in mortgage lending. With the existing customer data, the NBFC want to create a framework which can enable effective screening of low and medium risk borrowers to whom loans could be sanctioned quickly

Solution Provided

- CDPL developed a comprehensive framework for categorization of customers into four categories namely "Very High Risk", "High Risk ", "Medium Risk "and "Low Risk "respectively. Risk score was generated at a customer application level on a runtime basis.
- Used non-parametric method of supervised learning using Machine Learning (ML) methodology to segregate customers
- 3 separate models were created for 3 segments based on data analysis and bank-management input.
- Model created over small data set and tested over >10 times larger data set in UAT
    - Model 1 for Salaried borrowers availing housing loan (Sal-HL)
    - Model 2 for Non-Salaried borrowers availing housing loan (Non-Sal-HL)
    - Model 3 for borrowers availing non-housing loan products (non-HL)
- High AUC during Model creation and validation
- VAPT (security) standards of client maintained in solution

	Salaried – HL	NonSalaried – HL	Non HL
Training Data	2649	5000	5000
Training AUC	99.7%	99.12%	99.12%
Validation Data	850	1000	1000
Validation AUC	87.1%	81.2%	81.2%
UAT Data Size	62940	39848	26002

- Developed an API based solution which can generate the risk scores and risk category of loan applicants on the click of a button in the bank's LOS portal.
- Developed reporting dashboard with risk categorization of customers
- Ready for going-live, as the LOS of the NBFC has changed, go-live is planned post new LOS installation at NBFC
- Only 3 financial variables (FOIR ,average bank balance, LTV) were available, for modeling, remaining 21 variables were non-financial attributes. With integration with AA in future , the model predictions is bound to improve even further.

Output

- High level of accuracy in predicting risky accounts on ~1.3 lac customer data in UAT.
    - Bad rate in low risk category and medium risk category is significantly less than the overall sample bad rate for each of the 3 three models in UAT.
    - Bad rate in high risk category and very high risk category is significantly more than the overall sample bad rate for each of the 3 three models in UAT.
- Model showed high level of differentiation of bad rates across higher and lower risk segments both for CIBIL > 730 and for CIBIL<= 730
- Models were validated successfully against the approval rate of manual process of past bank loan approval. It is seen that Very High-Risk category (as defined by model) has the least approval rate and the Low-Risk Category (as defined by model) has almost 4 times more approval rate than the Very High-Risk Category.
- Models were validated successfully against the rejected past bank loans. It is seen that for the Low-Risk category, the customers made very good repayment over 6 months period while the borrowers from Very High-Risk Category (as defined by model) performed almost 6 times worse over 6 months period than the Low-Risk Category.
- Bank will be able to leverage the automated credit-scoring api-based solution which can generate credit score and category of a new transaction on a near real time basis directly from the LOS system
- Projected High Operational efficiency and proposed cost savings of ~$0.8 million

Please refer to next slide for UAT Results

Case Study 2 : Test Results across CIBIL

- Bad Case definition = 30 DPD + over a period of latest 18 months

Model Testing Results across all CIBIL Scores

Total Size = 128790	Salaried – HL	NonSalaried – HL	Non HL
Testing Data Size	62940	39848	26002
Bad case Rate in total test data	5.59%	12.18%	11.01%
Bad case Rate in Very High risk category	11.75%	23.91%	24.46%
Bad case Rate in High risk category	4.01%	13.54%	12.34%
Bad case Rate in Med risk category	2.69%	7.59%	6.40%
Bad case Rate in Low Risk category	1.66%	3.25%	3.60%

Model Testing Results for CIBIL Scores >= 730

Total Size = 47860	Salaried – HL	NonSalaried – HL	Non HL
Testing Data Size	23250	12971	11639
Bad case Rate in total test data	3.38%	8.33%	7.51%
% cases in Very High risk category	~19% (4450 cases)	~11.6% (1506 cases)	~10.7% (1243 cases)
% cases in High risk category	~12.5% (2916 cases)	~20% (2594 cases)	~19.5% (2278 cases)
Bad case Rate in Very High risk category	~9.5%	~21%	~22.6%
Bad case Rate in High risk category	~4%	~11.1%	~10.5%

Model Testing Results for CIBIL Scores <730

Total Size = 60811	Salaried – HL	NonSalaried – HL	Non HL
Testing Data Size	32760	17811	10240
Bad case Rate in total test data	6.79%	14.25%	13.05%
% cases in Low risk category	~28% (9166 cases)	~8.5% (1516 cases)	~10% (1031 cases)
% cases in Medium risk category	~21% (7018 cases)	~35% (6173 cases)	~35% (3613 cases)
Bad case Rate in Low risk category	~2.1%	~4%	~4%
Bad case Rate in Medium Risk category	~3.2%	~7.5%	~6.5%

Model Testing Results for Rejected Cases

Category defined through model over past rejected data	Total Decisions (Live HL/LAP Offus)	Performance of Rejected cases (% of 30 DPD in first 18 months)	Performance of Rejected cases (% of 30 DPD in first 36 months)
Very High	6018	6%	2%
High	2670	4%	1.1%
Medium	1879	3%	1%
Low	575	2%	0.7%

Model Testing Results for Approved Cases

Category defined through model over past data	Total Decisions	Approval Rate through incumbent process done on same data
Very High	6071	24%
High	2371	58%
Medium	3260	78%
Low	3561	95%

Case Study 2 : High Operational efficiency and cost savings

Connecting Dot

- Faster loan disbursement
- Credit decisioning process is faster. Near real-time generation of credit score
- Enablement of quick strategy formation by credit and collections team
- Potential to slash salary overhead and operational cost by at least 50%

- Figures from a recent implementation :

- The reject cases determined through extensive due diligence by our banking client's credit appraisal team were almost similar to the prediction of bad cases through LADA (without any manual intervention)

Without LADA:	**With LADA:**
No of underwriters required to use conventional process of underwriting for retail : 150 Average Salary = INR 700000 per year Total Salary per year = INR 105,000,000 Additional operations cost per year(~30% of salary) = INR 31,500,000 Total appraisal expenses per year = INR 135500000 = INR 13.5 crore = $1.6 million	No of underwriters required will be maximum 50%, so savings per LOB per year for a midsize housing finance bank = $0.8 million

Case Study 3 : CGTMSE Risk Categorization (Pilot Ongoing)

Background

Government of India launched Credit Guarantee Scheme (CGS) so as to strengthen credit delivery system and facilitate flow of credit to the MSE sector. To operationalize the scheme, Government of India and SIDBI set up the Credit Guarantee Fund Trust for Micro and Small Enterprises (CGTMSE). CGTMSE extends them helping hand by providing guarantee to enable them access credit leading to setting up viable micro and small enterprises. Integration with our LADA platform will help CGTMSE track the performance of the loans availing CGS Guarantee and can take preventive measures.

Solution Proposed

- Develop a comprehensive framework for categorization of customers into three categories namely "High Risk ", "Medium Risk "and "Low Risk "respectively. Risk score to be generated at a customer application level on a runtime or batch basis.
- Use non-parametric method of supervised learning using Machine Learning (ML) methodology to segregate customers
- Create 8 separate models for 8 segments based on data analysis and inputs from CGTMSE. Model to be created over small data set and tested over >10 times larger data set in UAT

Model 1	Model 2	Model 3	Model 4
• Non-Manufacturing sector for Working Capital Loans up to INR 50 lacs ticket size	• Non-Manufacturing sector for Term Loans up to INR 50 lacs as ticket size	• Non-Manufacturing sector for Working Capital Loans more than INR 50 lacs	• Non-Manufacturing sector for Term Loans more than INR 50 lacs as ticket size
Model 5	**Model 6**	**Model 7**	**Model 8**
• Manufacturing sector for Working Capital Loans up to INR 50 lacs ticket size	• Manufacturing sector for Term Loans up to INR 50 lacs as ticket size	• Manufacturing sector for Working Capital Loans more than INR 50 lacs	• Manufacturing sector for Term Loans more than INR 50 lacs as ticket size

Progress so far

- 8 model segments are identified

Model 1 created with 4050 data and tested over 10371 data points with satisfactory differentiation between various risk segments

- Bad rate (defined as NPA rate) 3.85 times in high risk category to that of low risk
- Following variables are available to be used in the model creation

□ **Borrower Behavioral data:** Family Constitution Type, Type of Activity of borrower, Industry Sector and Sub-sector, Greenfield/Existing business, Microenterprise or not

□ **Geographical Context**: State, District

□ **Borrower Financial Data**: None

□ **Facility information**: Bank name, Credit Guarantee amount availed, Loan amount, Loan Tenure, Collateral Amount, Promoter Contribution

□ **Sensitive variables** such as Caste ,Social Category are **not considered**

Please refer to next slide for Results

Case Study 3 : CGTMSE Model 1 Results (Pilot Ongoing)

- Bad Case definition = NPA cases

	Model 1
Training Data	4050
Training AUC	95.5%
Validation Data	450
Validation AUC	93.4%
UAT Data Size	10371
Bad case Rate in total test data	27.6% (2863 borrowers)
High risk category split	1265 (12.2%)
Bad case Rate in High risk category	46.72%
Medum Risk Category Split	6906 (66.6%)
Bad case Rate in Med risk category	29.03%
Low Risk Category Split	2200 (21.2%)
Bad case Rate in Low Risk category	12.13%
Bad Rate Ration in High Versus Low Risk Category	3.9

- **Looking ahead**
- - The model has given satisfactory results in ~2.5 times of testing data than that was used for model creation
- - The model shall be further tested in more data volume to check how it holds good in out-of-sample and out-of-time samples
- -Model created over NonCovid period has been tested over Covid+NonCovid period and is seen to be holding good
- - The model so far created has no financial variables of the customer, once AA data is integrated , the new financial variables is bound to give even superior performance to atleast 2 times

Connecting Dot

THANK YOU!

Address

4, Gouranga Mandir Lane,
Kolkata – 700086, West Bengal, India.

Contact Info

Email: support@connectingdot.in

Telephone

Office Phone: +91 9831195208
Sales: +91 9686596800

Website

https://connectingdot.in/

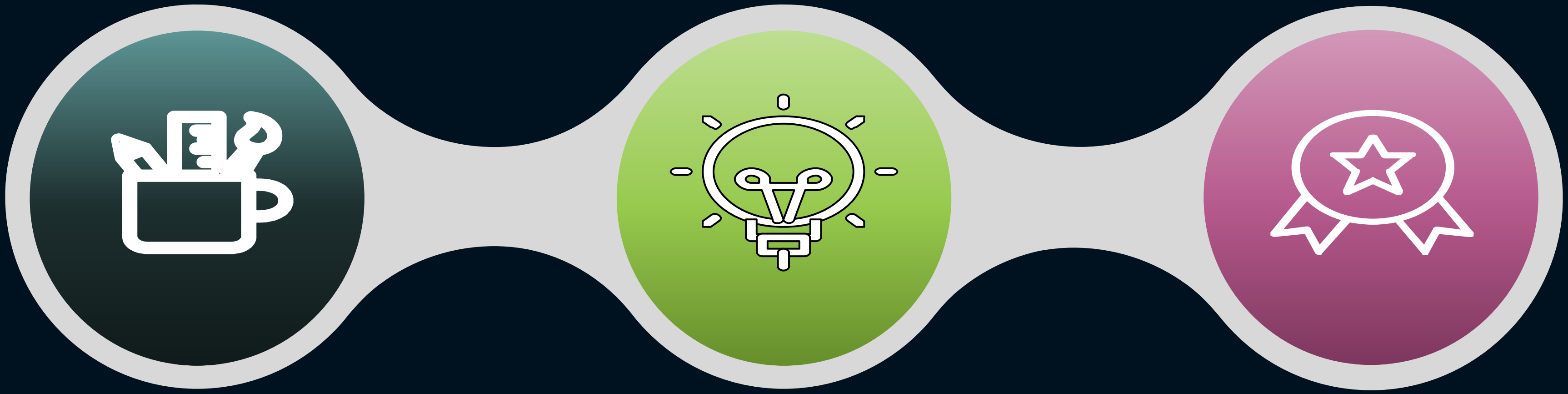
Appendix

LADA USP

CUSTOMIZED

ADVANCED MODEL

BANK READY



- **Customized Model :** "One model does not fit all" approach. Based on the internal bank data of bank only.
- **Customized Interface :** Flexible code design and enhanced parameterization leads to quick customization as per needs
- **Customized Roles :** Various roles can be configured with various levels of entitlement and authority

- **Going beyond the conventional techniques:** Uses advanced AI techniques such as SVM in addition to commonly used techniques such as logistics regression
- **Optimized Variable selection:** Based on deep domain experience supported by statistical techniques. Use of Game Theory to make variable selections
- **Proven Accuracy:** High accuracy rate of prediction already proven in multiple implementations. Accuracy rate remains valid even for huge volume of out-of-sample data

High level of security: VAPT cleared successfully for current client implementations. Code developed with security based design. Data security maintained through strict control over movement of sensitive data.

Multiversion support: Availability in web as well as API version

Efficient Architecture: Provides ability to process huge volume of data in less time and can be easily integration with bank internal systems (e.g. LOS)

LADA through various lens (1/3)

Explainabilty (qualitative assessment of model behavior)

- model-specific -- Feature importance
- model-agnostic – AUC, GINI
- Implementation of XAI techniques placed in different model lifecycle phases

Interpretibility (from a quantitative point of view)

- LADA can track the logic that governs the model's behaviour

Accuracy

- High Forecasting accuracy in in-sample as well as out-sample , out-of-time data
- Low Differentiation Bias over long period of time
- Proven success stories

LADA through various lens (2/3)

Performance

- Steady performance over time
- Efficiency In process management

Fairness

- Fairness through unawareness
- Statistical parity
- Supports customer right to enhanced information

Reliability

- Traceability of LADA’s functioning, enabled by the automatic recording of events in logs
- Use of techniques to detect anomaly /fraud in dataset
- Robustness and security features added
- Presence of a level of human control over LADA systems

LADA through various lens (3/3)

Bias handling

- Bias handling at at different phases of
  - *data collection ,*
  - *model specification and learning*
  - *output analysis.*

Governance and control

- User level control
- Reporting
- Staff training
- Documentation

Financial Inclusion

- Flexibility to leverage alternate data sources- enabling the assessment of the creditworthiness of entities otherwise excluded because of lack of standard financial data
- Uses non financial data as well as Financial data