


LOAN DEFAULT PREDICTION

MACHINE LEARNING CLASSIFICATION

FINAL CAPSTONE PROJECT

SUHANI PATEL

4/14/2024

A solid blue horizontal bar spanning the width of the slide, located at the bottom.

EXECUTIVE SUMMARY

KEY TAKEAWAYS

- Manual loan approval process - \$ profit loss due to default loans
- Automated approval process feasible - Machine learning model
- Random Forest model is most suitable
- Key data predictors are flagged
- Dataset is inherently imbalanced (RF model addresses skewness)
- Enables real time detection of high risk loan applications
- Deploying new model will increase profits, increase customer and investor goodwill

KEY NEXT STEPS

- Model deployment and testing
- Modify or improve model if necessary
- Robust and mandatory data gathering for key features during loan application
- Performance monitoring after model deployment
- Stakeholder actionables
 - Approve additional engineering resources?
 - Decide on deployment strategy?
 - Agree on deployment timeline?

PROBLEM STATEMENT

The bank loses profit due to a manual loan approval process which is prone to bias and error in human judgment.

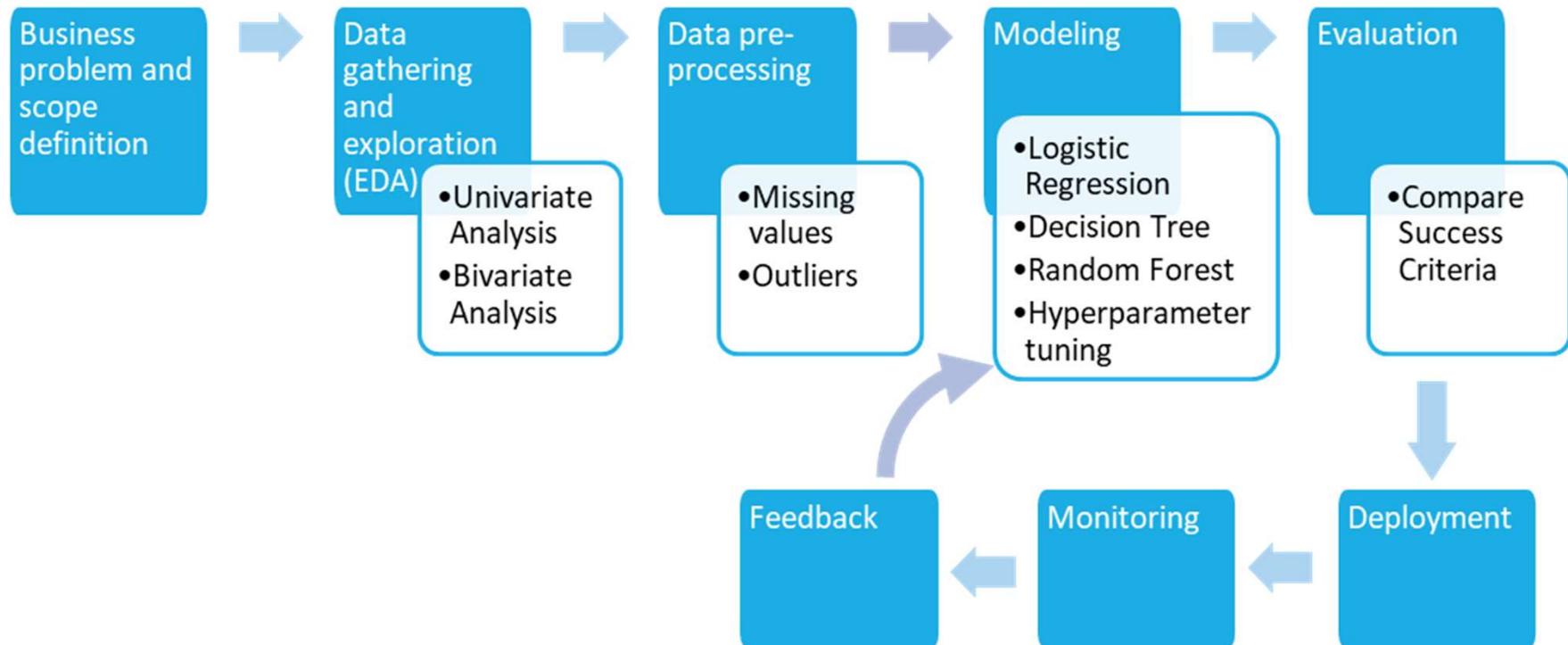
Current data shows that there are 12 out of 100 approved loans that default every year (~\$225,000 for every 100 approved loans). It is imperative for the business to reduce the default loans \$ loss by atleast 30% this year.

* Expected loss = Loss given Default (LGD) = Probability of default (PD) × Exposure at default (EAD).

Reference: <https://www.investopedia.com/terms/l/lossgivendefault>.

* December 31 2023 - 11.7 million residential mortgage loans, \$ 2.9 trillion in unpaid principal balances. 8320 new foreclosures in Q4 2023. Reference: <https://www.occ.gov/pub-mortgage-metrics-q4-2023.pdf>

SOLUTION SUMMARY



EDA and Data Pre-processing

Unique values in BAD are :

0	80.05
1	19.94

Name: BAD, dtype: float64

Unique values in REASON are :

DebtCon 65.90

HomeImp 29.86

Name: REASON, dtype: float64

Unique values in JOB are :

Other 40.06

ProfExe 21.40

Office 15.90

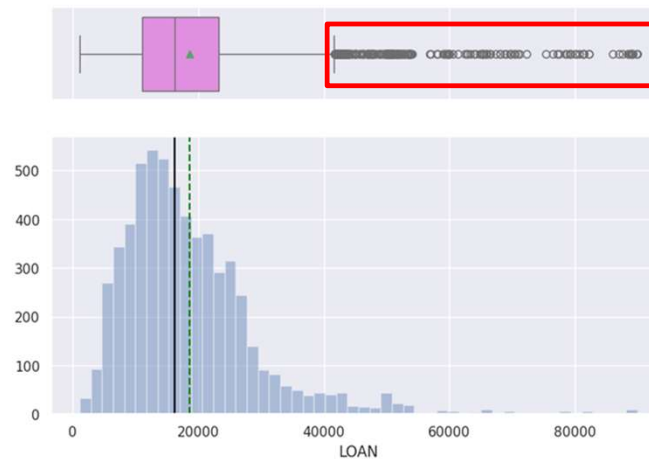
Mgr 12.86

Self 3.23

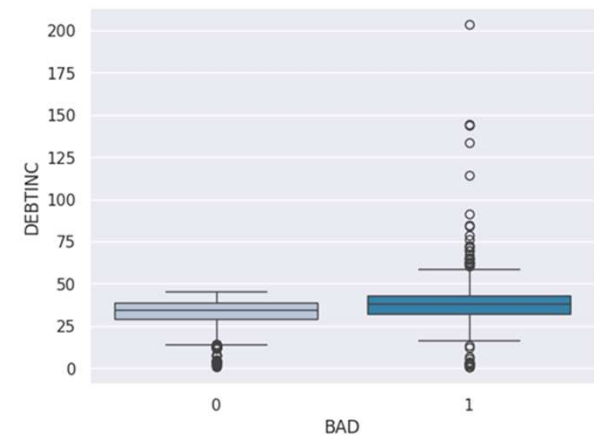
Sales 1.82

Name: JOB, dtype: float64

Outliers



Relationship between DEBTINC and BAD (default loan)



MISSING VALUE TREATMENT – Impute based on data type (Categorical = Mode, Numerical = Median). Create new binary variables for each column to flag missing values

OUTLIER TREATMENT – Capping to upper and lower whisker (based on 1.5 IQR) for logistic regression

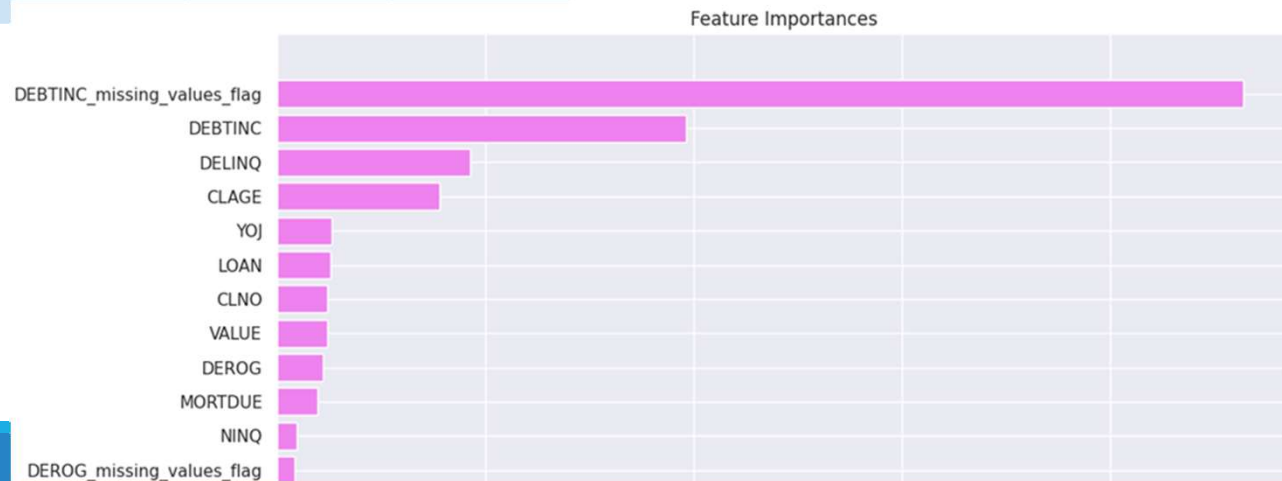
DATA SCALING – Data was not scaled for this project. Models used are not impacted by non-scaled data

RANDOM FOREST MODEL

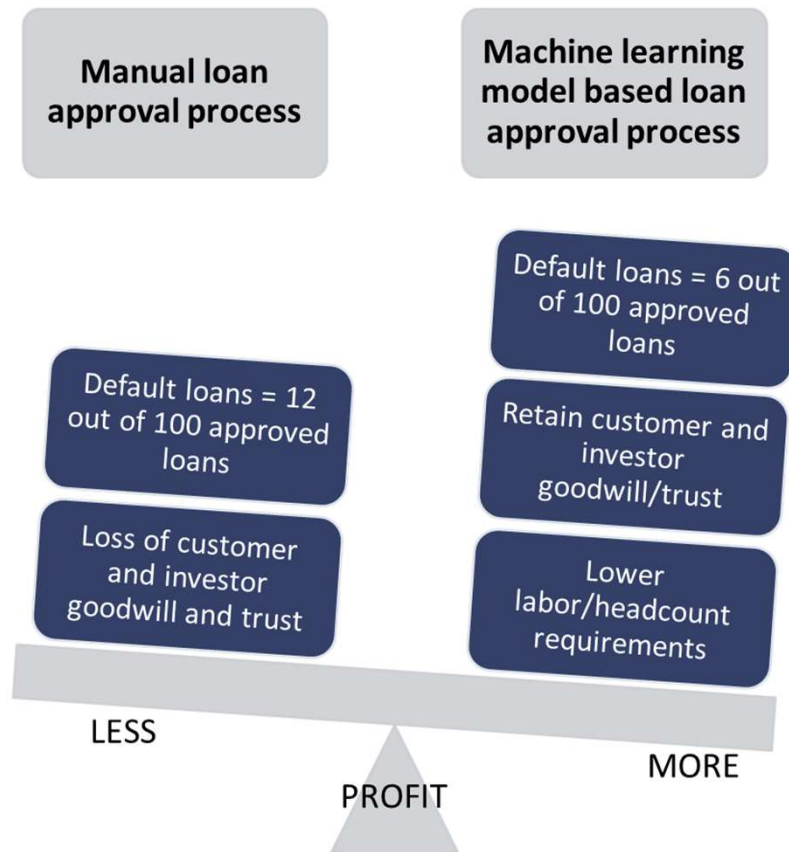
Model	Recall	Precision	F1 score
Logistic regression	0.50	0.65	0.45
Decision tree	0.77	0.81	0.79
Tuned decision tree	0.81	0.76	0.78
Random Forest	0.83	0.90	0.86
Tuned Random Forest	0.82	0.81	0.81

Maximize Recall =
Minimize false negatives

Actual	No Default	True negative	False positive
	Default	False negative	True positive
		No Default	Default
		Predicted	



BUSINESS VALUE – Current Vs Proposed



- Using this model reduces actual default loans to 6 out of 100 approved loans = \$118,635 for every 100 loans (gain in profit)
- This is 50% improvement compared to the manual loan approval process
- Reduced labor resources needed for manual approval process - additional savings ~\$100,000.

NEXT STEPS AND MODEL DEPLOYMENT

- Improve model performance
 - Collect more data
 - Use ensemble models such as Adaboost, Gradient boost, XGBoost
 - Collect data on geographic region to see impact on loan defaults
- Stakeholder actionables
 - Approve additional engineering resources – 2 Engineers
 - Decide on deployment strategy
 - Shadow mode
 - A/B testing
 - Canary deployment
 - Agree on deployment timeline – Total 8 months = 2 month (engineering) + 6 month (monitoring + feedback)
 - Create robust data collection for key features (dataset had upto 21% missing values)