Applied Data Science Program

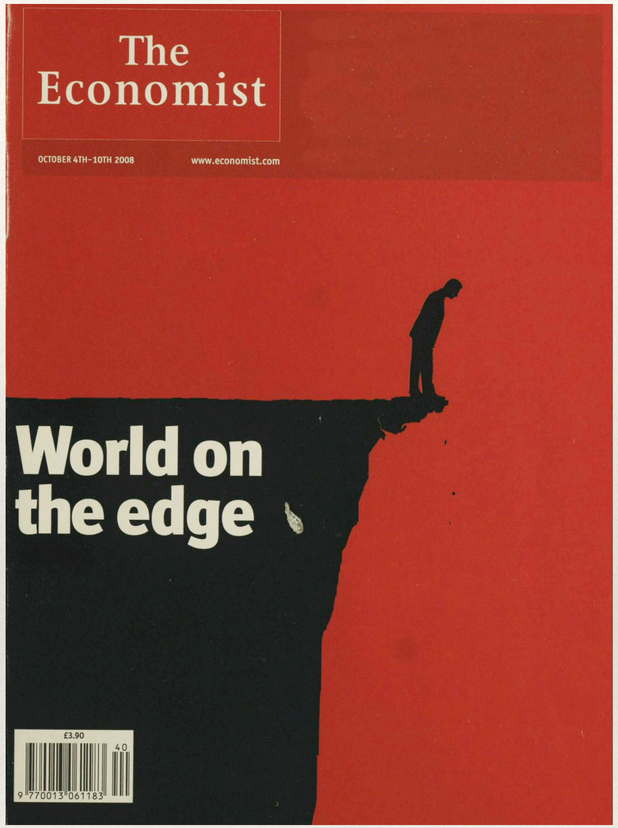
Capstone Project : Loan Default Prediction

Vivek S Magar
December, 21 2024

Understanding Loan Defaults

Why it Matters

- Home loan defaults are a **critical threat** to the financial stability of institutions and economies.
- The **2008 financial crisis** highlighted the devastating impact of such defaults on a global scale.



Challenges in Current Systems

- Errors in judgment and **human biases** hinder accurate evaluation.
- **Predatory lending** practices exacerbate risks.
- Reliance on external agencies for data without robust validation tools.
- Lagging responses to macroeconomic shifts.
- Erosion of **financial performance** and **reputation**.
- Exposure to **lawsuits, regulatory scrutiny**, and heavy fines.

The Promise of Machine Learning: A Smarter, Fairer Future

- Risk Mitigation: Improved **default predictions** and **revenue forecasting**.
- Enhanced Fairness: Reduced biases and greater equity in lending.
- Transparency: For management, shareholders, and regulators.

Key considerations

- Data Readiness: Do we have the **right datasets** and variables to assess creditworthiness?
- Predictive Power: How accurately can the system forecast defaults and financial outcomes?
- Fairness and Compliance: Can we ensure the system is **unbiased** and withstands **audit scrutiny**?

Call to Action

- Adopt a machine learning system with proper governance and controls.
- Drive a lending ecosystem that's efficient, equitable, and future-ready.

Understanding the data

Why it Matters

- Understanding data properties, structure, and **anomalies** is key to building effective predictive models with balanced performance

Data overview

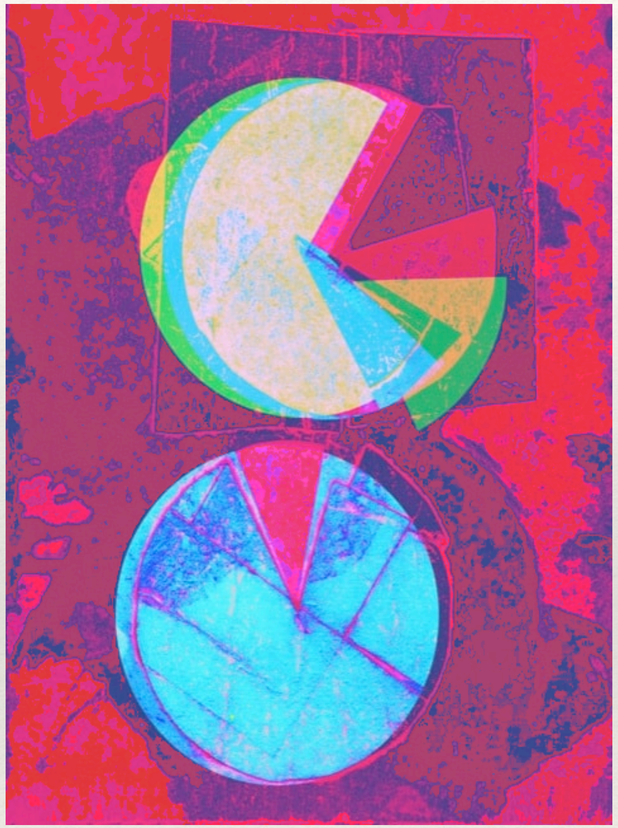
- 5,960 records, 12 features, 1 target variable. **~20% default rate**, suggesting bias or synthetic data.
- Missing values (e.g., 21.3% debt-to-income ratio), significant outliers and no duplicates.

Key Observations

- Unusual **high loan-to-value** ratios in several records.
- Job type and tenure may risk discriminatory litigation if not handled carefully.
- Default trends: High among clients with minimal credit or 30–41% debt-to-income ratio.
- Latent features such as "age of mortgage" may add predictive power

Insights for Prediction

- Key Features: Debt-to-income ratio, derogatory reports, recent inquiries, credit lines.
- Treat **missing values** statistically; cap and normalize skewed features.
- Combine **highly correlated variables** (e.g., mortgage due & property value).
- Outlier Impact: Debt-to-income ratio and mortgage amounts. Apply capping and transformation.
- Address bias: 70% loans for debt consolidation & legally sensitive features (e.g., job type).



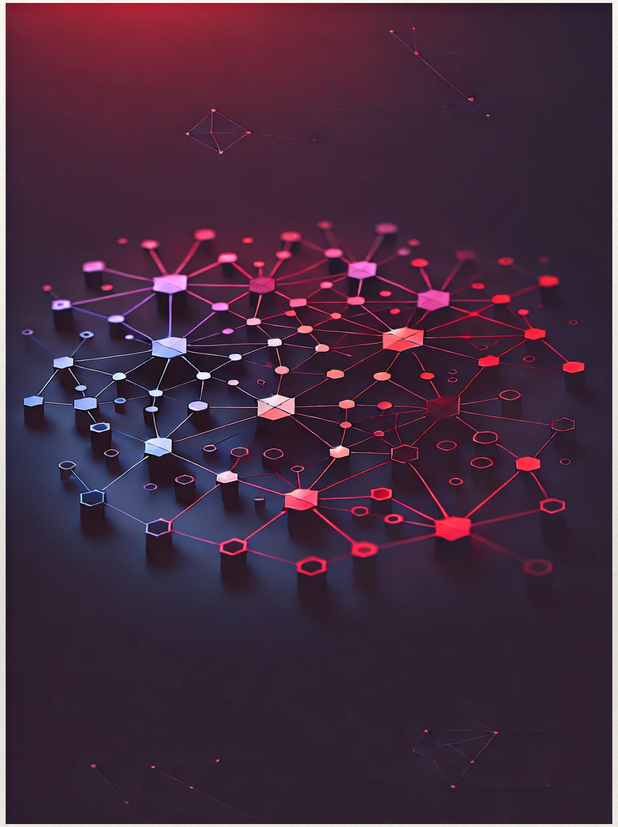
Predictive Modeling Approach

Why it Matters

- Selecting the right model ensures alignment with expected outcomes, and business KPIs.
- Evaluating multiple models is key to finding one that balances performance, business goals, constraints, and regulatory needs

Key considerations for modeling decisions

- Predict potential customer loan defaults is a binary classification problem
- Dataset contains customer profiles with known outcomes on defaults allowing for training of supervised learning models
- The model must deliver high accuracy in identifying potential defaulters, especially for high-value loans
- It should minimize the risk of misclassifying actual defaulters as non-defaulters to avoid significant financial losses
- Model predictions should be explainable for legal requirements and also to improve the model



Model Development and evaluation

- Prepare the data using the observation from data analysis (Refer to slide no 3)
- Define a custom metric to evaluate the expected revenue based on model predictions
- Train, test and compare the performance of the following models
  - Logistic regression, Decision Tree and Random Forest
- Use GridSearch to optimize the parameters for each of the above models
- Compare the performance of all models and fine-tune the best one to improve performance
- Review the relative importance of features used in prediction
- Analyze the misclassification of actual defaulters as non-defaulters

Strategic Roadmap for Loan Default Prediction

Model evaluation summary (Refer to appendix for details)

- Tree-based models outperform logistic regression, indicating non-linearity and outliers in the dataset.
- DecisionTree model, optimized for recall, excels in revenue performance while RandomForest models show superior metric results.
- Fine-tuning prediction thresholds, to reduce False Negatives substantially improved performance
- Debt to Income ratio, delinquent credit lines, age of the oldest credit line, approved loan amount and mortgage due to property value are the top 5 aspects driving the default prediction
- Customers wrongly classified as non-defaulters had the value of the above features close to average of all customers in the data set.

Proposal for final solution

- Random Forest model optimized for recall through GridSearch and adjusted threshold, is the best choice for future default predictions. It offers a strong balance of recall and expected revenue with high confidence.
- Use SHapley Additive Explanations (SHAP) technique to provide explanations for predictions.

Recommendations for Effective Model Implementation:

- Capture missing details (e.g., job type, loan reason) using drop-downs to reduce errors.
- Include financial data like income, expenses, savings, and credit scores for better assessments.
- Ensure fair representation of demographics and diversify loan types in the dataset.
- Regularly review bias and fairness to stay compliant.

These actions will boost model accuracy, fairness, and favorable business outcomes!

***Definitions**: **recall** - Number of correct default predictions. **Threshold** - Probability value, above which the model will classify the record as default. **False Negatives**: Wrongly classified as non-defaults*

Appendices

Model Comparison

Model Name	Model Type	Grid Search (Yes/No)	Recall	Precision	f1	Accuracy	FPs	FNs	Revenue Prediction	Confidence
lg	Logistic Regression	No	0.274510	0.748092	0.401639	0.836047	33	259	2.698671E+06	0.837825
log_reg_accuracy	Logistic Regression	Yes	0.274510	0.759690	0.403292	0.837170	31	259	2.668632E+06	0.837108
log_reg_f1	Logistic Regression	Yes	0.274510	0.748092	0.401639	0.836047	33	259	2.694613E+06	0.837809
log_reg_recall	Logistic Regression	Yes	0.274510	0.748092	0.401639	0.836047	33	259	2.694613E+06	0.837809
log_reg_precision	Logistic Regression	Yes	0.182073	0.783133	0.295455	0.825940	18	292	1.760434E+06	0.819371
dt	Decision Tree	No	0.565826	0.675585	0.615854	0.858506	97	155	7.501718E+06	1.000000
dtree_estimator_grid_accuracy	Decision Tree	Yes	0.591036	0.674121	0.629851	0.860752	102	146	7.346832E+06	0.995723
dtree_estimator_grid_f1	Decision Tree	Yes	0.591036	0.674121	0.629851	0.860752	102	146	7.346832E+06	0.995723
dtree_estimator_grid_recall	Decision Tree	Yes	0.565826	0.629283	0.595870	0.846154	119	155	7.458708E+06	1.000000
dtree_estimator_grid_precision	Decision Tree	Yes	0.459384	0.803922	0.584670	0.869175	40	193	4.874592E+06	0.911351
rf	Random Forest	No	0.630252	0.922131	0.748752	0.915216	19	132	5.562403E+06	0.922049
rf_estimator_grid_accuracy	Random Forest	Yes	0.680672	0.900000	0.775120	0.920831	27	114	6.276275E+06	0.944284
rf_estimator_grid_f1	Random Forest	Yes	0.680672	0.900000	0.775120	0.920831	27	114	6.276275E+06	0.944284
rf_estimator_grid_recall	Random Forest	Yes	0.680672	0.900000	0.775120	0.920831	27	114	6.276275E+06	0.944284
rf_estimator_grid_precision	Random Forest	Yes	0.521008	0.953846	0.673913	0.898933	9	171	3.285727E+06	0.862235

Function to calculate the expected revenue

Expected Revenue=Revenue from True Negatives (TN)−Loss from False Positives (FP)−Loss from False Negatives (FN)

Revenue from True Negatives (TN): RevenueTN = PND·(1−Actual)·L·r

Where: • PND=1−PD: Probability of non-default. • Actual: Indicator of the actual outcome (0 = non-default, 1 = default). • L: Loan amount. • r: Interest rate.

Loss from False Positives (FP): LossFP = PD·(1−Actual)·L·r

Where: • PD: Probability of default. • (1−Actual): Ensures the outcome is non-default (Actual=0). • L·r : Opportunity cost (lost interest revenue).

Loss from False Negatives (FN): LossFN = PND·Actual·L·(1+r)

Where: • PND: Probability of non-default. • Actual: Ensures the outcome is default (Actual=1). • L·(1+r) : Total loss (principal + interest).

Final Formula:

Expected Revenue = (PND·(1−Actual)·L·r) - (PD·(1−Actual)·L·r) - (PND·Actual·L·(1+r))

Definitions: • PD: Probability of default. • PND=1−PD: Probability of non-default. • Actual: Binary variable (0=Non-default,1=Default) • L: Loan amount. • r: Interest rate.

Techniques used in data analysis and modeling

Functionality	Technique/Metric	Description	Python Library
Impute missing values	KNN Imputation	A method to impute missing values by finding the nearest neighbors and using their values.	sklearn.impute
Impute missing values	Iterative imputation	A multivariate approach to impute missing values by modeling each feature as a function of others iteratively.	sklearn.experimental
Feature correlations	Correlation matrix	A matrix showing pairwise correlation coefficients between features.	pandas, numpy
Variance Inflation due to correlations	Variance Inflation Factor	A measure to detect multicollinearity in regression models by quantifying how much the variance of a predictor is inflated due to linear dependence with other predictors	statsmodels
Feature significance	Chi squared test	A statistical test to determine the independence between categorical variables or the goodness of fit for observed data	scipy.stats
Feature variance	Principle Component Analysis	A dimensionality reduction technique that transforms data into a lower-dimensional space by identifying the principal components with the most variance	statsmodels
Prediction	Logistic Regression	A statistical method used for binary or multi-class classification by modeling the probability of a categorical dependent variable using a logistic function	sklearn.linear_models
Prediction	Decision Trees	A tree-based model used for classification or regression that splits data into subsets based on feature values to make predictions.	sklearn.tree
Prediction	Random Forest	An ensemble learning method that combines multiple decision trees to improve classification or regression accuracy through majority voting or averaging	sklearn.ensemble
Model evaluation	Classification report	A classification report is a summary of the precision, recall, F1-score, and support for each class in a classification model's predictions, providing insights into the model's performance for each target label.	sklearn.metrics
Model optimization	Grid search	A hyperparameter tuning technique that systematically tests combinations of parameter values to find the optimal configuration for a model	sklearn.model_selection
Data Normalization	Z Scaler	A normalization technique that standardizes features by removing the mean and scaling to unit variance, resulting in a zero mean and standard deviation of one	sklearn.preprocessing
Encoding of categorical variables	Get dummy columns	A technique to convert categorical variables into numerical format by creating binary (0/1) columns for each category.	pandas
Split the data in test and training set	Stratified sampling	Splitting data into training and testing subsets for modeling.	sklearn.model_selection
Model evaluation	Confusion matrix	A table used to evaluate the performance of a classification model by showing the counts of true positives, true negatives, false positives, and false negatives.	sklearn.metrics
Threshold tuning	Precision-recall curve	A graph that shows the tradeoff between precision and recall at different classification thresholds, commonly used to evaluate model performance on imbalanced datasets	sklearn.metrics
Data vizualization	Heat maps, histograms, bar plots, box plots	Visualization techniques used to explore and represent data distributions, relationships, and summaries effectively	matplotlib, seaborn

Q & A