DA 204o: Data Science in Practice
Course Project Proposal

Credit Default Risk

Harshit Agarwal, harshita@iisc.ac.in
Sebin Kannampuzha, sebink@iisc.ac.in
Shreya Shrivastava, shreya3@iisc.ac.in
Tushar Srivastava, tushar1@iisc.ac.in

Image source: Internet

Problem Definition

Millions of people lack formal credit histories, making it hard for them to access fair loans and pushing them toward risky, unregulated lenders.

The challenge is to unlock safe credit access for the unbanked by accurately assessing their creditworthiness using innovative data sources.

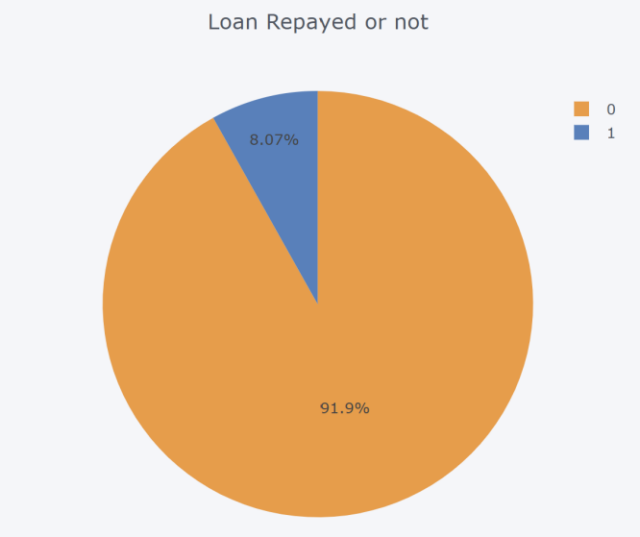
- Background of the problem
  - Traditional credit scores exclude millions without formal histories, especially in developing markets. This leads to loan denials or unsafe lending, fueling financial exclusion and exploitation. There is a pressing need for better ways to assess creditworthiness using alternative data sources.
- Why is it important?
  - Financial Inclusion: Fair loans empower the unbanked.
  - Reducing Exploitation: Safe lending protects borrowers.
  - Business Growth: New markets fuel lender growth.
  - Social Impact: Inclusive finance uplifts communities.
- Objectives of the project
  - Develop Predictive Models: Use alternative data to assess creditworthiness.
  - Improve Loan Approval Rates: Approve more deserving borrowers safely.
  - Promote Responsible Lending: Match loans to real repayment capacity.
- How can Data Science solve the problem?
  - Feature Engineering: Turn alternative data into actionable credit insights.
  - Machine Learning Models: Use AI to better predict and segment borrowers.
  - Risk Prediction: Accurately forecast defaults for safer lending.

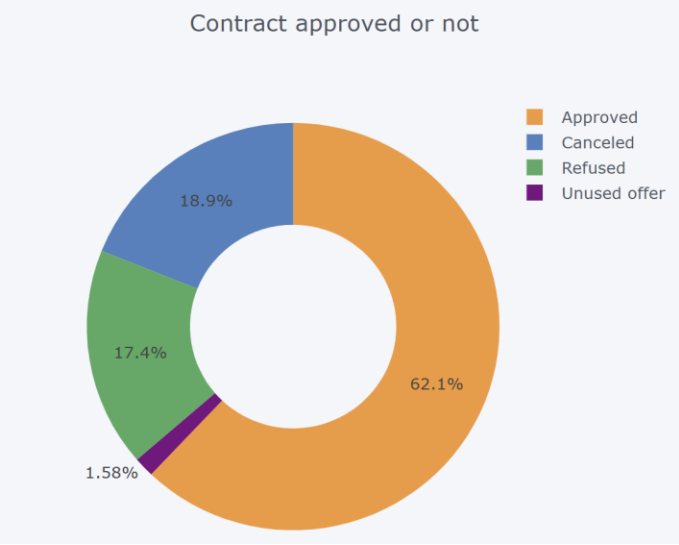
Data Collection and Preparation

Data source

Source : Home Credit Group – Kaggle Data

- The data was collected through Home Credit's operational activities. This includes:
  - **Application Data**: Information provided by clients when applying for a loan (e.g., income, age, family status).
  - **Internal Credit History**: Data on the applicant's past behavior with Home Credit, such as previous loans, credit card balances, and payment history.
  - **External Credit History**: Data from external Credit Bureaus regarding the applicant's credit history with other financial institutions.





Data Preprocessing and Feature Engineering

1. Data Integration & Preprocessing

- Dataset is a collection of 7 CSV
- The main files are application_train.csv and application_test.csv. All other files provide supplementary, historical information that can be linked back to the main files using identifier keys.

2. **Feature Engineering Strategies**

- Polynomial Features: Generated interactions of 3 external credit scores (EXT_SOURCE_1, 2, 3) with DAYS_BIRTH to capture non-linear relationships
- Domain insights: Age (DAYS_BIRTH) shows strongest positive correlation with repayment; external scores negatively correlate with default risk
- Multi-level aggregations: Computed count, mean, max, min, sum statistics for numeric columns; aggregated one-hot encoded categoricals

3. **Feature Engineering Strategies**

- Polynomial Features: Generated interactions of 3 external credit scores(EXT_SOURCE_1, 2, 3) with DAYS_BIRTH to capture non-linear relationships
- Domain insights: Age (DAYS_BIRTH) shows strongest positive correlation withrepayment; external scores negatively correlate with default risk
- Multi-level aggregations: Computed count, mean, max, min, sum statistics fornumeric columns; aggregated one-hot encoded categoricals

Proposed Methodology

Algorithms Used

- **Random Forest**: Selected for its robustness against overfitting, ability to handle non-linear relationships, and capacity to work with a large number of features. It provides good predictive performance and can also be used for feature importance ranking.
- **XGBOOST:** Used as a high-performance gradient boosting algorithm that handles non-linear relationships, missing values, and many sparse features efficiently. It is well-suited for tabular credit-risk data, offers built-in regularization to reduce overfitting, and provides interpretable feature importance scores.
- **Light Gradient Boosting Machine (LightGBM)**: Chosen due to its reputation for speed, efficiency, and scalability when handling large datasets. It is particularly effective for classification problems like credit risk assessment, often demonstrating high accuracy and outperforming other models.
- **Blended (50-50 XGB + LGBM):** A simple ensemble that averages the predicted probabilities from XGBoost and LightGBM. This 50–50 blend achieved the best validation AUC among all tested models.

Tools/Technologies (e.g., Python, libraries)

- Pandas, numpy, scikit-learn (for Random Forest, data preprocessing, model selection, evaluation metrics), lightgbm, matplotlib, seaborn, plotly (for interactive visualizations), xgboost, lightgbm.

Implementation Plan

Resources Required

- **Tools:** Python, Jupyter, pandas, scikit-learn, matplotlib
- **Platforms:** Kaggle (data), GitHub (version control), Google Colab / cloud compute
- **Team Roles:** Data acquisition & cleaning lead, modeling lead,visualization/documentation lead
- **Hardware/Compute:** Moderate compute (GPU optional), 8–16GB RAM recommended

- **Project Phases**
  - **Setup & Data Acquisition:** Define problem, objectives, roles and download datasets, understand schema and relationships
  - **Data Cleaning & Preprocessing:** Handle missing values, encode variables, remove outliers, engineer basic features
  - **Exploratory Data Analysis (EDA):** Visualize data, analyze distributions and correlations, identify trends and drift
  - **Feature Engineering:** Create and select meaningful features for modeling
  - **Baseline Modeling:** Build simple models (logistic regression, random forest) and establish benchmark
  - **Advanced Modeling & Tuning:** Implement boosted models, tune hyperparameters, evaluate stability
  - **Evaluation & Insights:** Assess performance, analyze feature importance, derive insights
  - **Final Report & Delivery:** Prepare report, slides, and code submission
- **Milestones**
  - Data acquisition and schema understanding completed
  - Preprocessed dataset ready
  - EDA findings and insights documented
  - Final feature set selected
  - Baseline models evaluated
  - Advanced models tuned and validated
  - Stability and interpretability analysis completed
  - Final report and presentation delivered

Challenges and Risks

- Potential risks or challenges

  - **Data Quality & Completeness**: Alternative data may be incomplete or inaccurate, affecting prediction accuracy.
    - Around 30 columns with Null values around 50% or more
  - **Bias & Fairness**: Models may unintentionally favor or disadvantage certain groups, resulting in unfair lending decisions.
    - Use the best F1 Score across the training dataset since the dataset is imbalanced for non-defaulters (approx. 80%)
  - **Feature Selection & Engineering**: Choosing irrelevant or too many features can increase complexity and reduce model performance.
    - Applied feature importance for all algorithms to maximise the AUC score
    - Used the best features between LGBM and XGBoost model for training and prediction

Expected Outcome

	model	AUC	num_features
0	Blend 50/50 (intersection)	0.787173	569
1	Blend 50/50 (union)	0.787099	629
2	XGB (union)	0.785690	629
3	LGBM (intersection)	0.784666	569
4	XGB (all)	0.784619	750
5	XGB (intersection)	0.784082	569
6	Blend 50/50 (all)	0.783932	750
7	LGBM (all)	0.783245	750
8	LGBM (union)	0.782987	629
9	RF (union)	0.754971	629
10	RF (all)	0.753699	750

What do you expect to achieve?

- **Comprehensive Data Understanding**: Gain deep insights into the credit dataset, including distributions, outliers, and relationships between features and loan default.
- **Identification of Key Risk Factors**: Pinpoint the most influential features contributing to credit risk, informing business strategies and credit policies.
- **Development of Robust Predictive Models**: Build and optimize accurate machine learning models (Logistic Regression, Random Forest, LightGBM) for creditworthiness assessment.
- **Improved Decision-Making**: Provide actionable insights and predictive capabilities to assist financial institutions in making informed loan application decisions.
- **Benchmarking Model Performance**: Compare the effectiveness of different models for the credit risk prediction task.

How will you measure success?

- **High Area Under the Receiver Operating Characteristic Curve (ROC AUC)**: Primary metric for discrimination between defaulting and non-defaulting customers.
- **F1-Score**: Balanced performance in minimizing false positives (financial losses) and false negatives (missed opportunities).
- **Clear Feature Importance Scores**: Ability to rank features by their influence on predictions.
- **Outperformance of Baseline Models**: Developed models should significantly exceed simple benchmarks.

Role and Responsibilities

- Harshit Agarwal:
  - Modelling and Deployment
- Sebin Kannampuzha:
  - Modelling and Feature Selection
- Shreya Shrivastava:
  - Exploratory Data Analysis
- Tushar Srivastava:
  - Preprocessing
  - Feature Engineering

Data Science Canvas

Project: Credit Default Risk

Team: Harshit, Sebin, Shreya, Tushar

Problem Statement

Business Case & Value Added

- Millions lack formal credit histories, leading to financial exclusion and unsafe lending.
- **Value:** Improved decision-making for financial institutions, identification of key risk factors.

Data Landscape

Required Data: Features like telco and transactional information.

Available Data: Application Data, Internal Credit History, External Credit History

Additional Data to be collected:

Model Selection

- **Logistic Regression** (baseline)
- **Random Forest** (robustness, non-linear)
- **LightGBM** (speed, efficiency, accuracy for large datasets)
- **XGBoost**
- **Ensemble Model**: LightGBM and XGBoost

Model Requirements

- High ROC AUC, Accuracy, Precision, Recall (F1-Score).
- Low Log Loss.
- Model Interpretability

Software & Libraries

- **Languages/Tools:** Python, Jupyter.
- **Libraries:** pandas,numpy, scikit-learn,lightgbm,matplotlib, seaborn, plotly,xgboo

Skills

- Data Acquisition & Cleaning
- Feature Engineering
- Machine Learning Modeling
- Model Tuning & Validation
- Data Visualization

Execution & Evaluation

Model Evaluation

- Measure success using ROC AUC, Accuracy, Precision, Recall, F1-Score, and Log Loss.
- Compare performance against baseline models.

Data Storytelling

The target group = **risk analysts + credit officers** who need:

- **Clear probability of default (PD)** for each applicant — not just a label.
- **Explainability** — why the model thinks someone may default (key features).
- **Reliability** — stable, validated predictions (AUC, calibration).
- **Actionability** — how the scores can support loan approval / rejection decisions.

Results can be communicated via **AUC metrics, key feature drivers , and simple visuals** that show how the model improves loan decisions.

Data Collection & Preparation

Data Selection & Cleansing

Relevant data: All applicant-level and history tables (application data, bureau data, previous loans, credit card balances, POS cash, installment payments).

Cleanup: the data must be cleaned for missing values, outliers, inconsistent categories, merging relational tables, and handling class imbalance before modeling.

Data Collection

Collect extra data through credit bureau checks, income verification, and transaction history, ensuring it is **accurate, up-to-date, consistent, and legally compliant** for reliable risk prediction.

Data Integration

Migrate all sources into a **centralized data warehouse or unified relational database (e.g., PostgreSQL/BigQuery)** so the data stays consistent, joinable, and ready for modeling.

Explorative Data Analysis

The EDA shows strong skewness and clear outliers in key financial fields (income, credit, annuity, days features), plus many missing-value structures that must be handled.

Descriptive statistics such as mean, median, percentiles, missing-value rates, and distribution plots reveal that defaulted clients typically have lower income, higher credit ratios, and more late payments.