Excellent — here is a **clear, professional, submission-ready explanation report** you can directly use in your project file, assignment, or internship submission.

You can paste this into a document as **“Credit Risk Modeling Report”**.

📘 Credit Risk Modeling and Expected Loss Estimation

Probability of Default (PD) Model Report

---

1. Introduction

The objective of this project is to build a machine learning–based credit risk system that estimates the **Probability of Default (PD)** for loan borrowers using historical financial data. Using the predicted PD and assuming a fixed recovery rate of 10%, the system also computes the **Expected Loss (EL)** for any given loan.

Expected loss is a key financial risk metric used in banking and lending institutions and is defined as:

[
\text{Expected Loss (EL)} = \text{PD} \times \text{LGD} \times \text{EAD}
]

Where:

- PD = Probability of Default
- LGD = Loss Given Default = 1 − Recovery Rate
- EAD = Exposure at Default (loan amount outstanding)

---

2. Dataset Description

The dataset contains borrower-level financial and credit attributes:

Feature	Description
customer_id	Unique customer identifier
credit_lines_outstanding	Number of active credit lines
loan_amt_outstanding	Outstanding loan amount
total_debt_outstanding	Total current debt
income	Annual income
years_employed	Employment history (years)
fico_score	Credit score
default	Target variable (1 = default, 0 = non-default)

The dataset was verified to contain **no missing values** and is fully numerical, making it suitable for statistical modeling.

---

3. Feature Engineering (Financial Risk Ratios)

To enhance predictive power and reflect real-world banking practices, additional **financial ratios** were derived:

- **Debt-to-Income Ratio** = total debt / income
- **Loan-to-Income Ratio** = loan amount / income
- **Utilization Ratio** = loan amount / total debt
- **Credit Lines per Year** = credit lines / years employed

These ratios capture borrower leverage, repayment capacity, and financial stress, which are fundamental indicators of default risk.

---

4. Exploratory Data Analysis

Several plots were generated to understand the data and risk drivers:

a) Correlation Heatmap

Used to identify relationships between variables and default. Strong correlations were observed between default and financial strength indicators such as FICO score and debt ratios.

b) Distribution Plots

- FICO score vs default
- Debt-to-income ratio vs default

These plots showed strong separation between defaulters and non-defaulters, validating the relevance of the chosen features.

c) Boxplots

Boxplots highlighted the median differences and spread of high-risk variables, confirming their discriminatory power.

---

5. Model Development

Two classification models were trained:

🔹 Logistic Regression (Baseline PD Model)

- Industry-standard method for PD modeling
- High interpretability
- Suitable for regulatory and financial risk applications

🔹 Random Forest (Challenger Model)

- Non-linear ensemble model
- Captures complex interactions
- Used to benchmark performance improvements

The dataset was split into **80% training** and **20% testing**, preserving default ratios. Standardization was applied for logistic regression.

---

6. Model Evaluation

Model performance was evaluated using **ROC-AUC**, which measures the ability to distinguish between defaulters and non-defaulters.

Both models achieved very high ROC-AUC values, indicating excellent discriminatory power. ROC curves were plotted to visually compare performance.

A **PD distribution plot** further demonstrated that defaulters were assigned significantly higher default probabilities.

---

7. Feature Importance

Random Forest feature importance analysis was conducted to identify key default drivers. The most influential features included:

- FICO score
- Debt-to-income ratio
- Loan-to-income ratio
- Total outstanding debt

This aligns with financial risk theory and validates the model's learning behavior.

---

8. Expected Loss Engine

An expected loss engine was built using the trained models.

Assumptions:

- Recovery Rate = 10%
- LGD = 90%
- EAD = Loan amount outstanding

The system provides real-time outputs:

- **Predicted Probability of Default**
- **Expected Loss value**

This enables financial institutions to quantify portfolio risk and make informed lending decisions.

---

9. Business Relevance

This system can be applied in:

- Loan approval decision support
- Credit risk monitoring
- Risk-based pricing
- Capital adequacy assessment
- Portfolio stress testing

The project follows industry-aligned risk modeling practices and demonstrates how machine learning can support financial risk management.

---

10. Conclusion

The developed credit risk system successfully integrates:

- Financial feature engineering
- Statistical and machine learning models
- Visualization-driven risk analysis
- An expected loss estimation engine

The results show that borrower defaults are strongly driven by credit quality and debt burden metrics. Logistic regression was selected as the primary PD model for its stability and interpretability, with Random Forest serving as a strong challenger model.

This project demonstrates a complete, production-style pipeline for default prediction and financial loss estimation.

---

11. Key Formula Used

[
\text{Expected Loss} = \text{PD} \times (1 - \text{Recovery Rate}) \times \text{Loan Amount}
]

If you want, I can also prepare:

✔ A **short one-page abstract**
✔ A **PowerPoint-style summary**
✔ A **viva/interview explanation sheet**
✔ Or a **PDF formatted report layout**

Just tell me your submission format.