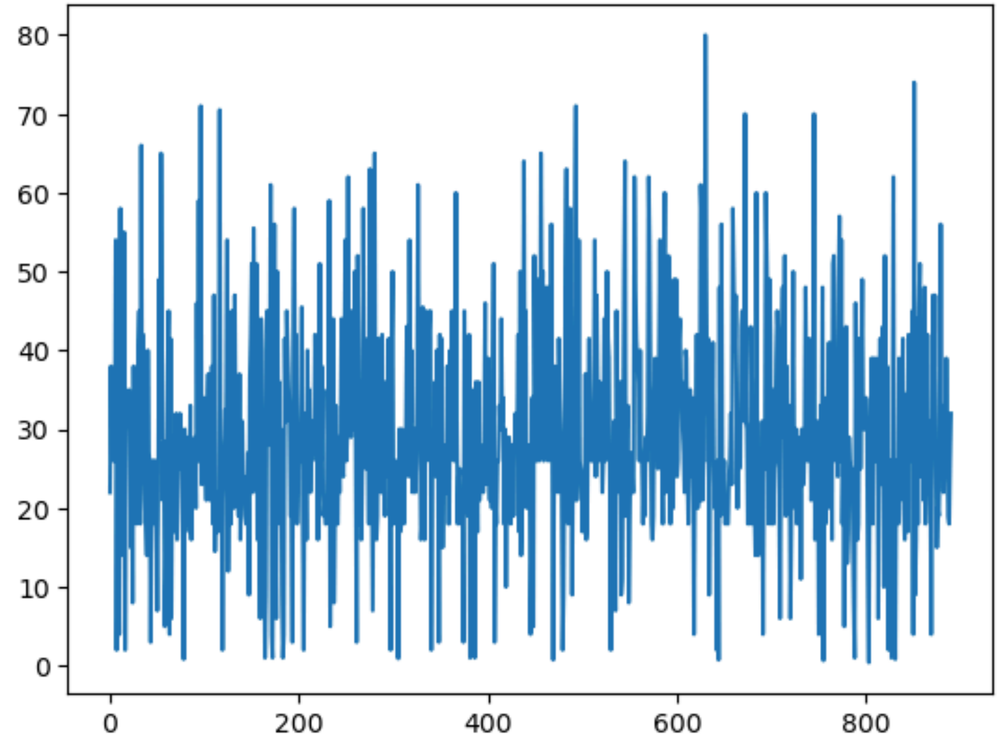
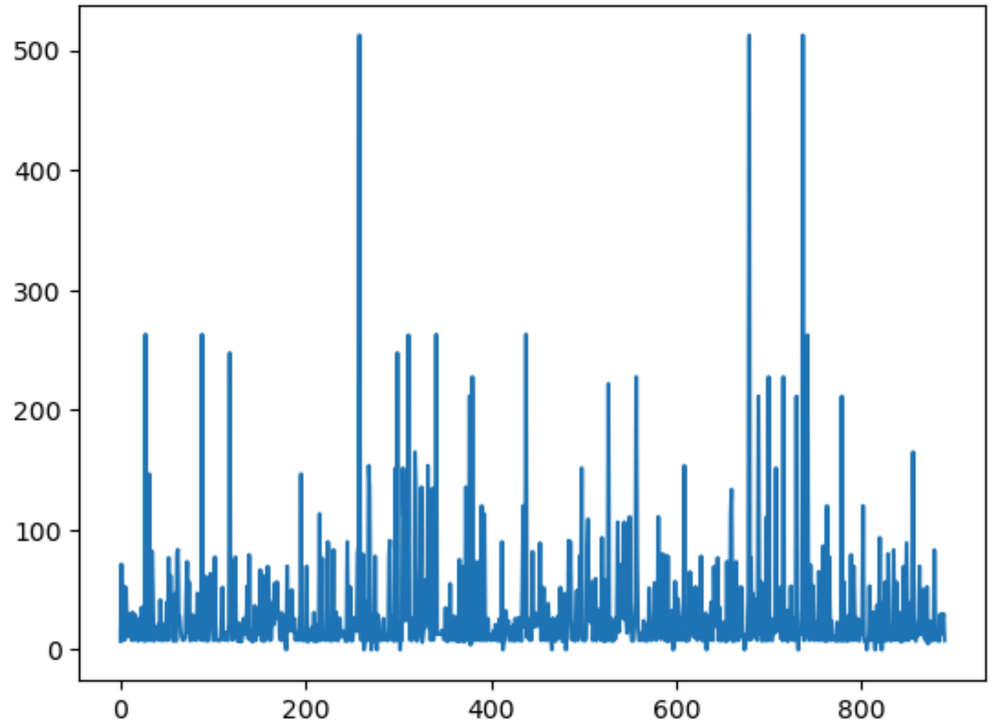
Titanic Dataset Classification related to passenger survival

Noah Onyebuchi

Handing and Removing Outliers



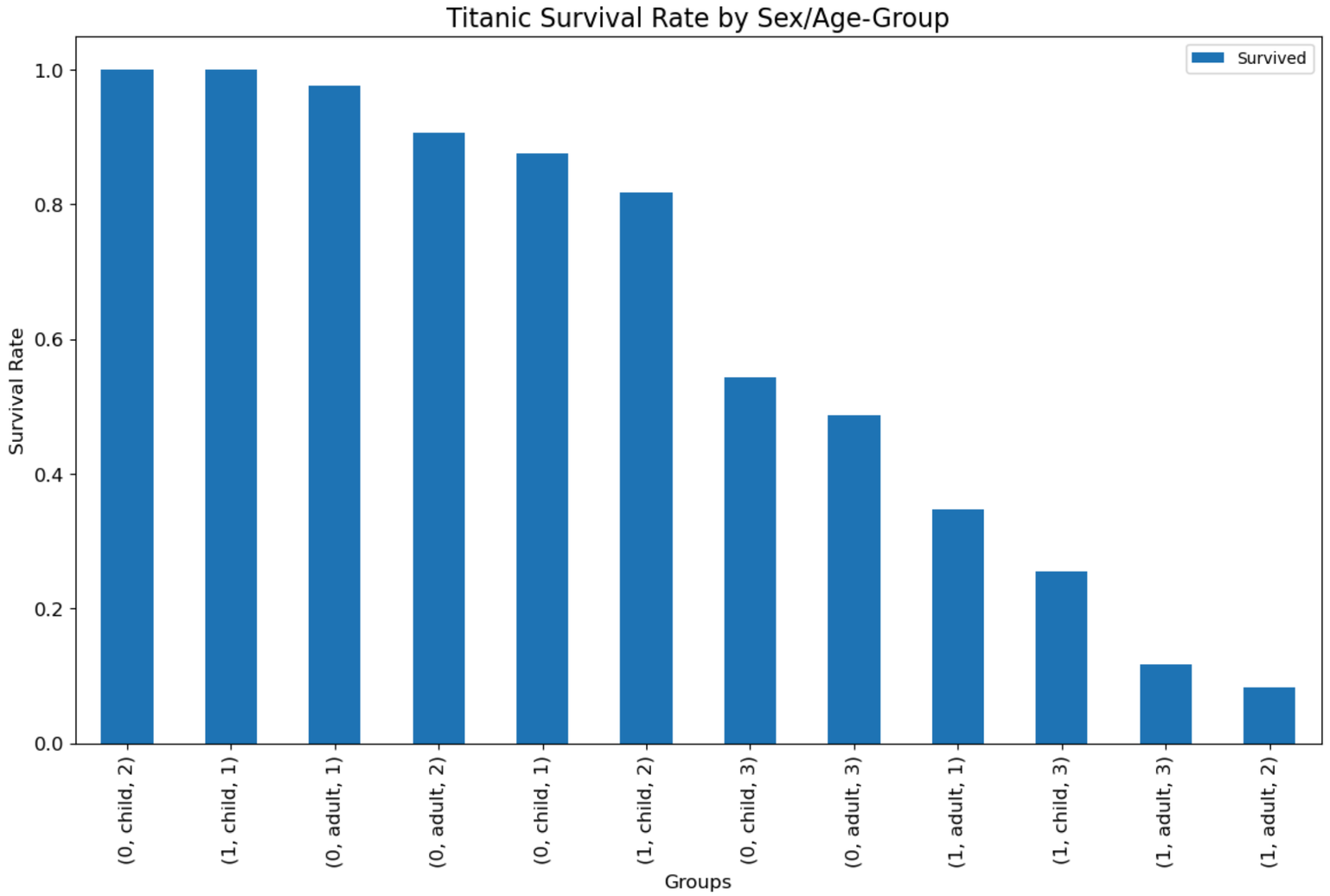
Age



Fare

data.loc[data.Fare > 250,"Fare"] = 250

Identifying relevant factors
that significantly Survival



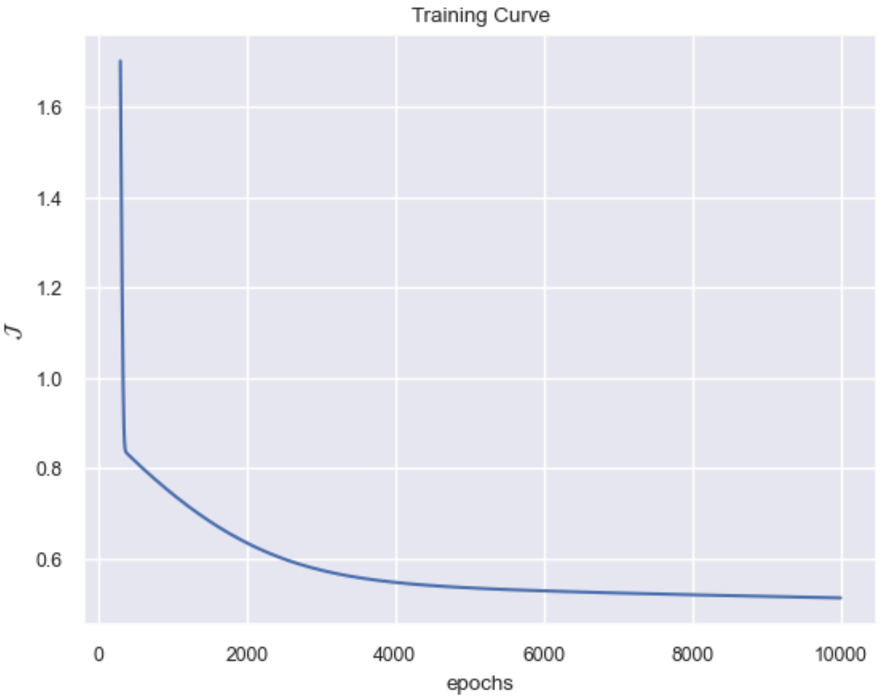
Survival rate : 0.38

Female:0.74 Male:0.17

Pclass1: 0.62 Pclass2: 0.47 Pclass3 0.24

Adult 0.36 Child 0.54

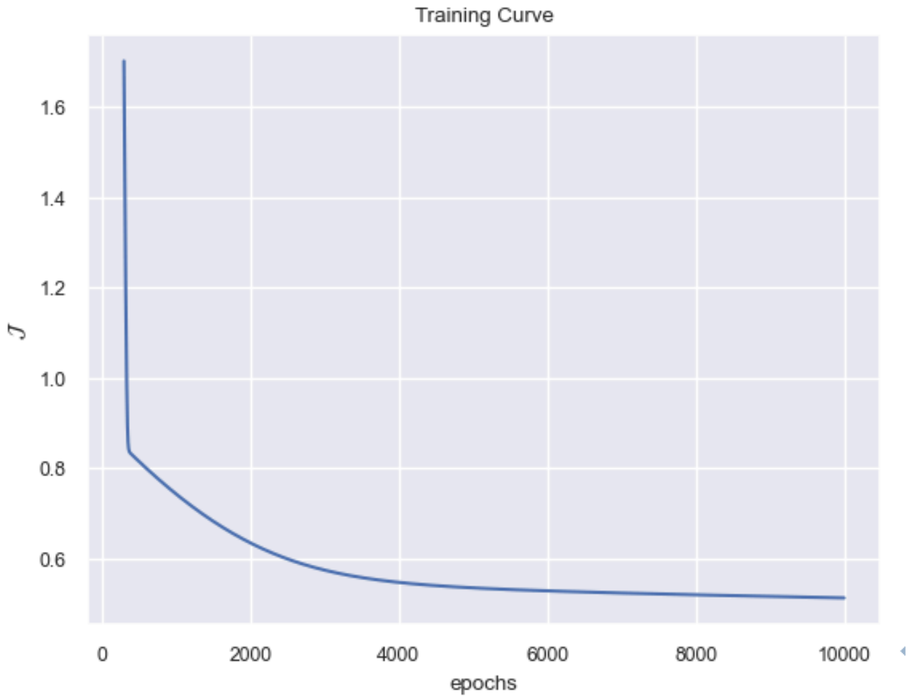
Binary Logistic Regression Class



eta=1e-3, epochs=1e4

Training Accuracy: 0.7778

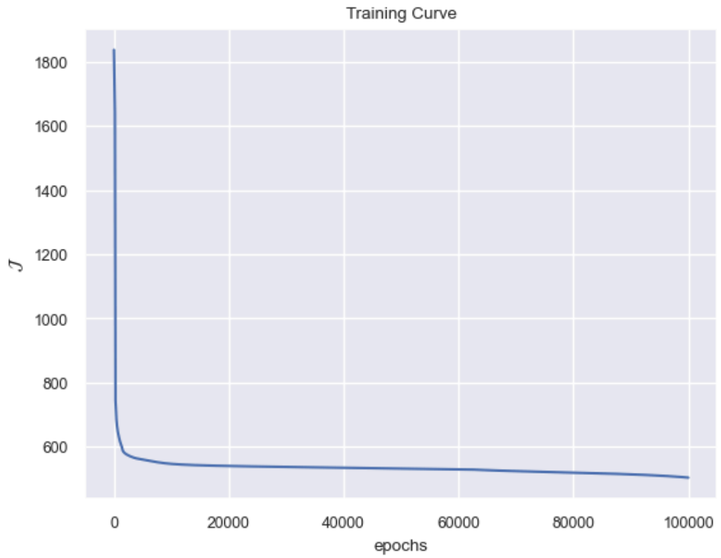
Multi Class Logistic Regression



eta=1e-4,epochs=1e4

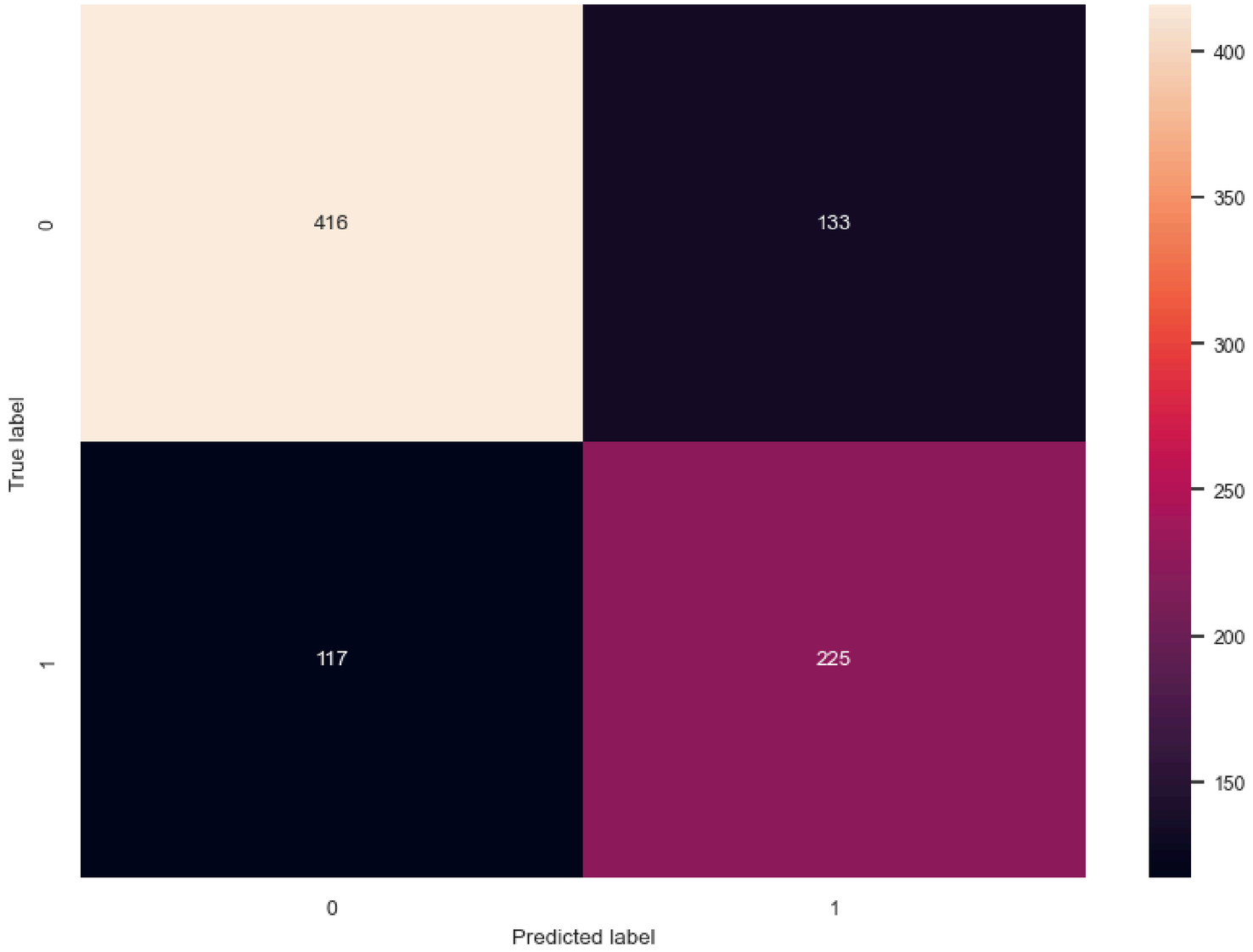
Training Accuracy: 0.6756

Shallow ANN Class



neurons= 3, eta=1e-3, epochs = 1e5,

Training Accuracy: 0.7194



True Negatives 416
Correctly predicted ‘Not Survived’.
• True Positives 225
Correctly predicted ‘Survived’.
• False Positives 133
Incorrectly predicted ‘Not Survived’.
• False Negatives 117
Incorrectly predicted ‘Survived’

Loan Pricing Model for a Fintech Startup

Understanding the Task

While the Titanic dataset is primarily used for classification tasks related to passenger survival, we can adapt it to model loan default by drawing parallels between passenger survival and loan repayment.

Key Parallels:

- Survival vs. Repayment: Both involve predicting a binary outcome (survived/deceased or repaid/defaulted).

Use Case Statement

A small fintech startup seeks to optimize loan pricing for small businesses by leveraging historical data and machine learning techniques. The goal is to develop a predictive model that accurately assesses loan risk and recommends appropriate interest rates.

Key Objectives

Minimize default risk:

Accurately predict the likelihood of loan default and account for any losses associated with it. The model should consider time to event and use survival analysis techniques, including the Cox proportional hazards model and Kaplan-Meier estimator

Minimize prepayment risk:

Reduce the likelihood of early repayment, which can impact profitability.

Profit maximization:

Attract and retain customers by offering competitive interest rate and fair loan terms that maximizes profits while minimizing risk.

key Considerations

Historical Data: The availability and quality of historical data on customer loans, repayments, defaults, and prepayments are crucial for model development.

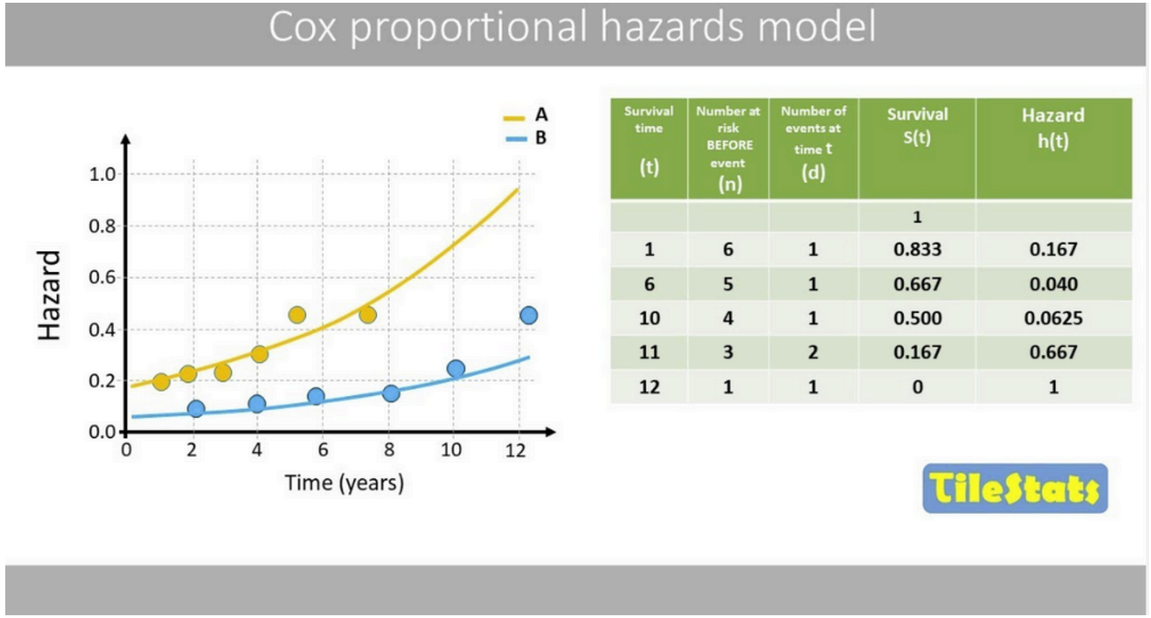
Factor Selection: Identifying relevant factors (e.g., credit history, loan amount, loan term, industry type) that significantly influence default and early repayment is essential.

Survival Analysis: Using techniques like the Cox proportional hazards model and Kaplan-Meier estimator can help estimate the probability of an event (default or early repayment) occurring over time.

Loss Estimation: Quantifying potential losses due to default is necessary to incorporate into the loan pricing model.

Pricing Optimization: Set loan interest rates that reflect the estimated risk and maximize profitability while maintaining competitive pricing.

Cox proportional hazards model and Kaplan-Meier estimator



Survival time (t)	Number at risk BEFORE event (n)	Number of events at time t (d)	Survival S(t)	Hazard h(t)
			1	
1	6	1	0.833	0.167
6	5	1	0.667	0.040
10	4	1	0.500	0.0625
11	3	2	0.167	0.667
12	1	1	0	1

The Cox proportional hazards model is a statistical model used in survival analysis to analyze the time until a specific event (e.g., death, default, or early repayment) occurs. This model is a frequently used approach that allows the investigator to study relationships between the time to event outcome Y and a set of explanatory variables X1, X2, ..., Xp.

The Kaplan-Meier estimator is a non-parametric method used to estimate the survival function, which gives the probability of an individual surviving beyond a certain time point. It's particularly useful when you have censored data (e.g., when you know an individual has not experienced the event by a certain time but don't know their exact survival time).

Proposed Architecture

1. Data Preprocessing and Feature Engineering:

- Data cleaning: Handle missing values, outliers, and inconsistencies in the data.
- Feature engineering: Create new features or transform existing ones to improve model performance. For example, you could create a credit score based on various credit history factors.

2. Default Prediction:

- Classifier: Use a binary classifier to predict whether a loan will default.
- Algorithms: Consider Logistic Regression, KNN, or ANN as potential candidates.

3. Time-to-Event Prediction (if default = 0):

- Survival Analysis: Use a survival analysis model to predict the time until default.
- Algorithm: The Cox proportional hazards model is a common choice for survival analysis.
- Features: Input features would be similar to those used in the default prediction step

4. Early Repayment Prediction (if default = 1):

- Regression: Use a regression model to predict the time until early repayment.
- Algorithm: Linear Regression, Random Forest, or Gradient Boosting Machine could be considered.
- Features: Input features might include interest rates, loan terms, and economic indicators.

5. Loan Pricing:

- Risk Assessment: Combine the predicted probabilities of default and early repayment with loss estimates to assess the overall risk of the loan.
- Interest Rate Calculation: Set the interest rate based on the assessed risk and the desired profit margin.

Workflow

1. Data Preprocessing: Clean and prepare the data for modeling.
2. Default Prediction: Pass the data to a Logistic Regression classifier to predict whether the loan will default.
3. Time-to-Event Prediction (if default=0): Pass the data to a Cox proportional hazards model to estimate the time until default.
4. Early Repayment Prediction (if default=1): Pass the data to a Random Forest regressor to predict the time until early repayment.
5. Loan Pricing: Calculate the interest rate based on the predicted probabilities of default and early repayment, along with loss estimates.

Thank You