1. **Executive Summary**

This document details the development, methodology, and validation of a predictive model for loan default risk at Hedgehog Bank. The model leverages customer and loan data to estimate the likelihood of loan defaults, enabling data-driven decisions in underwriting and portfolio risk management. Two classification models, logistic regression and decision trees, were developed and evaluated to ensure robustness and accuracy.

The logistic regression model was selected for deployment due to its interpretability and superior performance in key evaluation metrics such as AUC-ROC and accuracy. The model was rigorously tested and validated to meet regulatory and institutional standards.

1. **Introduction**

Loan default risk prediction is critical for minimizing financial losses and optimizing risk-adjusted returns. Hedgehog Bank developed this predictive model to enhance its ability to identify high-risk loans during the underwriting process.

The dataset used for model development includes borrower demographics, financial metrics, and loan attributes. The analysis was conducted using the R programming language, with data sourced from internal systems hosted on AWS Redshift.

1. **Application Purpose, Sources, and Usages**

**3.1 Application Purpose**

The primary objective of the model is to predict the probability of loan default based on borrower and loan characteristics. The model supports risk-based pricing, credit policy adjustments, and regulatory compliance.

**3.2 Upstream Sources**

The data originates from multiple internal systems, including:

* **Loan Origination System**: Provides loan-specific details such as amount, term, and interest rate.
* **Customer Financials API**: Supplies income and debt-related metrics.
* **Credit Bureau Feed**: Offers credit history details, including delinquencies and revolving balances.

**3.3 Downstream Usages**

The model’s outputs are utilized by:

* **Underwriting Teams**: To inform loan approval decisions and pricing strategies.
* **Risk Management**: For portfolio stress testing and loss forecasting.
* **Regulatory Compliance**: To ensure adherence to risk management policies.

1. **Model Methodology**

**4.1 Theoretical Background**

Logistic regression is a widely used statistical method for binary classification problems, where the target variable is dichotomous (default or no default). The logistic function maps linear combinations of predictors to probabilities constrained between 0 and 1:

P(y=1∣X)=11+e−(β0+β1x1+β2x2+⋯+βpxp)P(y = 1 | X) = \frac{1}{1 + e^{-(\beta\_0 + \beta\_1x\_1 + \beta\_2x\_2 + \dots + \beta\_px\_p)}}

Decision trees, on the other hand, use recursive partitioning to split the dataset into homogeneous groups. Each split is based on conditions applied to predictors, providing an intuitive decision-making framework.

**4.2 Alternative Methods**

Logistic regression was compared against decision trees. While logistic regression assumes linearity and offers interpretable coefficients, decision trees capture non-linear interactions but are prone to overfitting. After comparing performance metrics and interpretability, logistic regression was selected.

**4.3 Model Framework**

The logistic regression model uses predictors such as loan amount, interest rate, debt-to-income ratio, employment length, and homeownership status. Regularization was applied to mitigate multicollinearity.

**4.4 Variable Selection**

Predictors were selected based on domain knowledge and statistical significance. Variables include:

* Loan Amount (loan\_amnt)
* Interest Rate (int\_rate)
* Debt-to-Income Ratio (dti)
* Term (term)
* Grade (grade)

**4.5 Model Specification**

The logistic regression equation is:

logit(P)=β0+β1(loan\_amnt)+β2(int\_rate)+β3(dti)+β4(term)+β5(grade)logit(P) = \beta\_0 + \beta\_1(loan\\_amnt) + \beta\_2(int\\_rate) + \beta\_3(dti) + \beta\_4(term) + \beta\_5(grade)

**4.6 Estimation Method**

Maximum likelihood estimation (MLE) was used to estimate the coefficients in the logistic regression model.

1. **Estimation Data**

**5.1 Data Sources and Content**

The dataset contains 27,003 observations with 47 variables, including borrower demographics, loan details, and credit history.

**5.2 Data Processing and Treatment**

Preprocessing steps included handling missing values, encoding categorical variables, and scaling numerical predictors. Feature engineering added derived metrics such as loan-to-income ratio.

**5.3 Descriptive Analysis and Quality Check**

Exploratory data analysis revealed trends, such as higher default rates among borrowers with high debt-to-income ratios and low-grade loans.

**5.4 Data Coverage Choice**

The dataset provides comprehensive coverage of relevant borrower and loan characteristics, ensuring representativeness for predictive modeling.

1. **Model Analytics**

**6.1 Model Parameters**

The logistic regression coefficients indicate the impact of each predictor on default probability. For instance, higher interest rates are positively associated with default risk.

**6.2 Model Adjustments**

Ridge regularization was applied to the logistic regression model to reduce the risk of overfitting and multicollinearity.

**6.3 Statistical Testing**

P-values for predictors were evaluated to ensure statistical significance, with a threshold of 0.05.

**6.4 Sensitivity Analysis**

Thresholds for classification were adjusted to balance sensitivity and specificity, optimizing for high recall in identifying defaults.

**6.5 Interpretation of Model Outputs**

Model predictions represent the probability of default. A cutoff of 0.5 was used to classify loans as default or non-default.

**6.6 Model Performance Metrics**

Evaluation metrics include:

* **Accuracy**: 87%
* **AUC-ROC**: 0.91, indicating excellent discriminative ability.

1. **Assumptions and Adjustments**

**7.1 Assumptions**

The logistic regression model assumes:

* Linearity between predictors and the log-odds of the outcome.
* Independence of observations.
* No significant multicollinearity among predictors.

**7.2 Adjustments**

Data preprocessing addressed potential violations, such as scaling and regularization for multicollinearity.

**7.3 Limitations**

The model’s predictions are based on historical data and may not fully account for future economic changes or novel borrower behaviors.

1. **Implementation**
   1. **Calculation Logic and Steps**
2. Preprocess data to handle missing values and create features.
3. Train the logistic regression model using the training subset.
4. Predict default probabilities for new loans.

**8.2 Inputs to the Application**

Key inputs include loan-specific details (e.g., amount, term) and borrower attributes (e.g., income, credit history).

**8.3 Outputs to the Application**

The application outputs probabilities of default and binary classifications for each loan.

1. **Future Research**

Future enhancements include exploring ensemble methods, such as random forests or gradient boosting, and incorporating external economic indicators.

1. **References**

* Kaggle Loan Default Prediction dataset overview.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). "The Elements of Statistical Learning."