北京工业大学

本科生课程考试答题纸

考试课程：统计类项目设计基础

学生学号：21068121

学生姓名：司徒婉儿

题号	分数	任课教师签名
一		
二		
三		
四		
五		
六		
七		
八		
九		
十		
总分		

考试时间：2024 年 12 月 31 日

基于借款人特征的贷款审批预测：Logistic 回归模型分析

摘要

本研究旨在利用贷款数据集，通过构建 Logistic 回归模型，预测贷款批准状态，并分析借款人特征与贷款批准及违约风险之间的关系。报告使用了一个包含丰富借款人信息和贷款特征的数据集，涵盖年龄、收入、房屋所有权等多个变量。通过对 58645 个样本的 13 个自变量进行描述性统计分析，给出了值型变量的五数统计，热力图和箱线图，分析变量间的相关性和分布形态。在模型构建过程中，移除了与预测目标关联性不显著的变量，并利用剩余变量进行 Logistic 回归分析。数据集被随机划分为 70%的训练集和 30%的测试集。模型的性能通过 ROC 曲线和 AUC 值进行评估，其中 AUC 值为 0.9002，表明模型具有很高的区分能力。

模型结果显示，贷款信用等级、贷款金额占收入的比例、租房情况、贷款用于家庭改善等因素是贷款获得批准的正向预测因素，而年收入、工作年限、贷款用于教育、医疗、个人消费、创业等因素则是负向预测因素。尽管模型表现出色，但存在信用等级降低，贷款批准概率却提高的反直觉现象，这可能归因于数据质量、分类变量编码、多重共线性或模型设定等问题。

该研究为贷款机构评估风险和做出贷款决策提供了有价值的参考，但需要进一步的数据审查和模型诊断，以确保模型的准确性和可靠性。

关键词：信用风险评估；Logistic 回归；贷款审批；ROC 曲线；AUC 值

目录

一、数据来源及基本背景描述

本报告所使用的数据集是一个专注于贷款批准预测的数据集。该数据集通过深度学习模型生成，旨在模拟真实的贷款审批流程，以帮助金融机构评估和预测借款人的信用风险。数据集中包含了丰富的借款人个人信息和贷款特征，如年龄、年收入、房屋所有权、工作年限、贷款目的、贷款信用等级、贷款金额、贷款利率、贷款金额与收入比例、违约记录、信用历史长度以及贷款批准状态等关键变量。这些信息为我们提供了一个全面的视角，用以分析借款人特征与贷款批准及违约风险之间的关系。变量说明表如下所示，包含信息有字段名、变量含义、单位、取值等。

表 1 变量说明表

字段名	含义	说明	取值
id	样本 ID	用于标识每个借款人的记录	数值型
person_age	借款人年龄	单位：岁	数值型
person_income	借款人年收入	单位：美元	数值型
person_home_ownership	借款人房屋拥有情况	RENT（租房）、OWN（拥有）、MORTGAGE（按揭）、OTHER（其他）	分类型
person_emp_length	借款人工作年限	单位：年	数值型
loan_intent	贷款意图	包括：EDUCATION（教育）、MEDICAL（医疗）、PERSONAL（个人消费）、VENTURE（创业）、DEBTCONSOLIDATION（债务整合）、HOMEIMPROVEMENT（家庭改善）	分类型
loan_grade	贷款信用等级	从 A 到 G 表示不同的信用等级，A 最好，依次递减	分类型
loan_amnt	贷款金额	单位：美元	数值型
loan_int_rate	贷款利率	百分比表示	数值型
loan_percent_income	贷款金额占收入的比例	无单位	数值型
cb_person_default_on_file	是否有违约记录	Y：有，N：无	分类型
cb_person_cred_hist_length	信用历史长度	单位：年	数值型

loan_status	贷款批准状态	0：未获得批准，1：获得批准	数值型

二、描述统计

整个数据集一共有 58645 个样本和 13 个自变量，不存在缺失值。对其中的数值型变量进行五数统计，结果如下表二。

表 2 变量五数表

变量	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
person_age	20	23	26	27.55	30	123
person_income	4200	42000	58000	64046	75600	1900000
person_emp_length	0	2	4	4.701	7	123
loan_amnt	500	5000	8000	9218	12000	35000
loan_int_rate	5.42	7.88	10.75	10.68	12.99	23.22
loan_percent_income	0	0.09	0.14	0.1592	0.21	0.83
cb_person_cred_hist_length	2	3	4	5.814	8	30

对于除了 id 以外的数值型数据作热力图，分析其相关性。越接近红色代表相关性越大，越接近浅蓝代表相关性越差，变量间存在正相关和负相关，负相关的相关性很低。可以看出年龄与信用历史时长、贷款金额与贷款利率的相关性较高，贷款批准状态与贷款利率、贷款金额占收入的比例相关性较高。对于与预测变量相关性不高而存在变量间高相关的变量，可以考虑全部或部分删除。

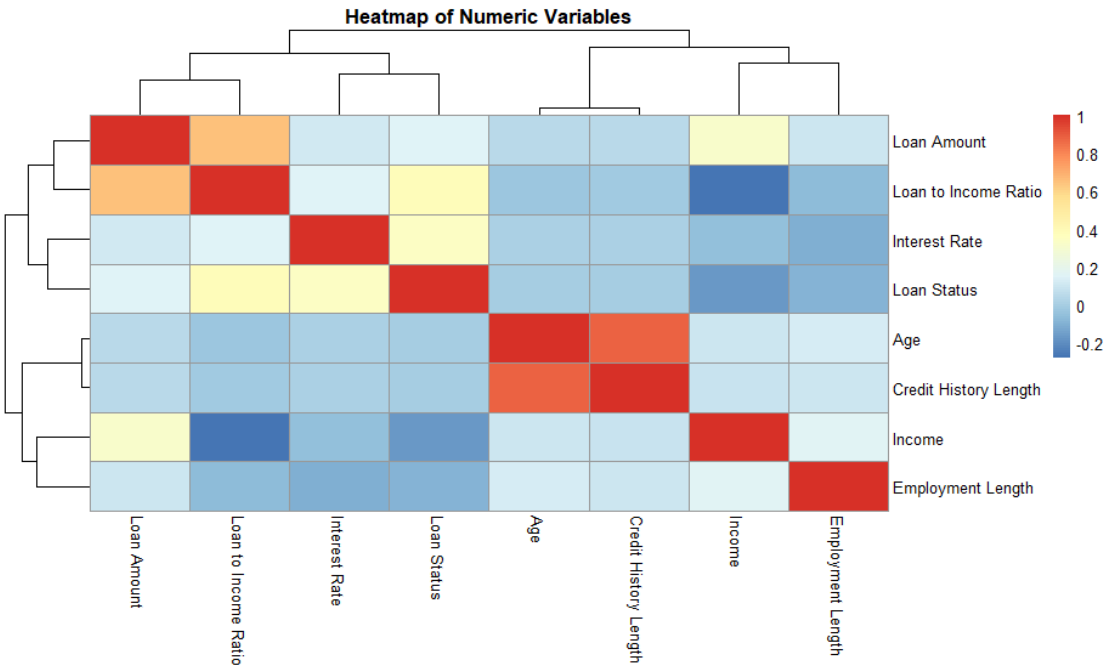

图 1 数值型变量相关关系热力图

贷款金额排布非常均匀且存在较高数额，部分变量的取值具有跳跃性，也存在一些异常值，为右偏分布。

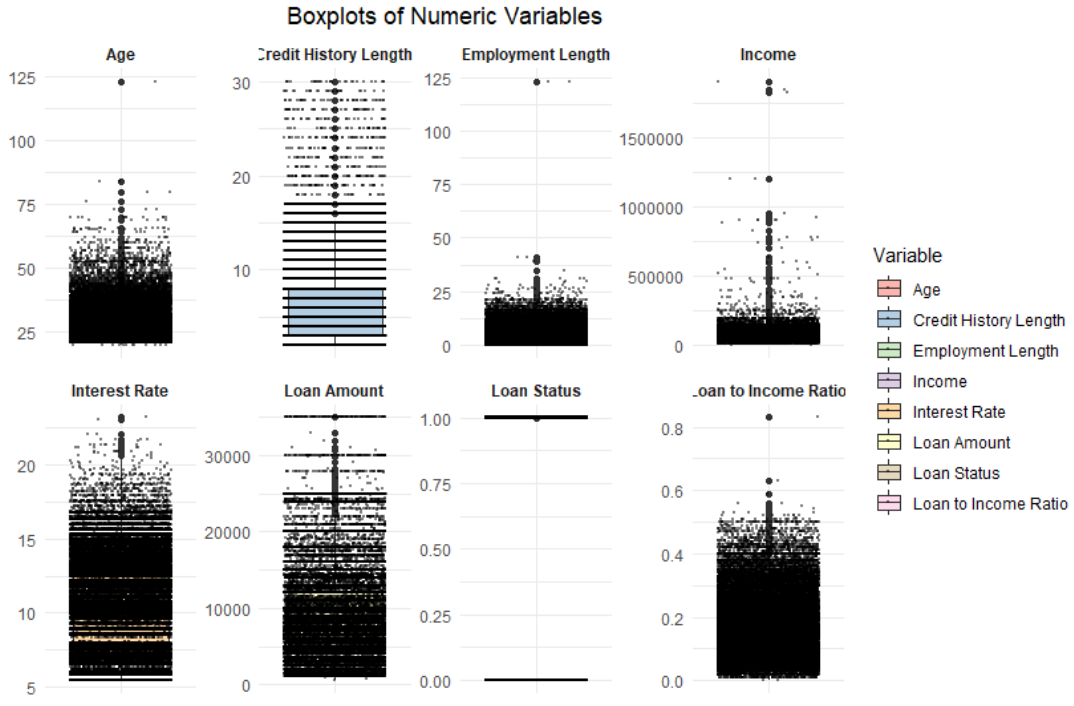

图 2 数值型变量分布形态箱线图

图 3 分类变量占比分布图

对于四个分类变量，贷款者房屋类型中其它房屋占比最小，租赁房屋类型最多。贷款意向六个类型占比较为均匀，贷款信用等级中，信用类型最高为 A，占比较多的是 A、D，C 次之，最小占比是信用等级超级低的 F 和 G。无违约记录的贷款者占比最高，大约占总体的 5/6。

三、模型建立

在模型构建中，主要目标是预测贷款批准状态（loan_status），这是一个典型的二元分类问题，其中 1 代表贷款获得批准，而 0 则代表未获得批准。为了构建一个有效的预测模型，首先对数据集进行了筛选，移除了 person_age（借款人年龄）和 cb_person_cred_hist_length（信用历史时长）这两个变量。这一决策基于它们与预测目标之间的关联性似乎并不显著。在筛选后，保留了包括借款人年收入、房屋拥有情况、工作年限、贷款意图、贷款信用等级、贷款金额、贷款利率以及贷款金额占收入比例等关键变量，这些变量将被纳入 Logistic 回归模型中。

考虑到自变量中既包含数值型数据又包含分类数据，而预测变量是一个 0/1 编码的二元变量，结合描述性统计分析得出的结论，为提高预测精度，删除年龄和信用历史时长信息，对剩余变量进行准确的 Logistic 回归分析。

为了估计模型，首先将数据集按照 7:3 的比例随机分为训练集和测试集，即有 70%的数据用于训练模型，30%的数据用于测试模型的性能。

通过在测试集上进行预测并评估其性能。通过 ROC 曲线和 AUC 值来评估模型的预测能力。ROC 曲线是一个图形工具，用于展示在不同阈值下模型的真正例率（TPR）和假正例率（FPR）。AUC 值是 ROC 曲线下的面积，它提供了一个单一的数字来衡量模型的整体性能。AUC 值越接近 1，表示模型的区分能力越强。

四、模型结果与结论

对于最终模型，分析如下。正系数表示随着变量值的增加，贷款获得批准的概率增加；负系数则表示相反。模型的截距项为-5.745，这表示在所有预测变量为 0 时，贷款获得批准的对数比值，id 的系数非常小，表明样本 ID 对贷款批准状态的影响微乎其微。年收入的系数也非常小，年收入每增加 1 美元，贷款获得批准的概率略有下降。

借款人的房屋拥有情况较为丰富，与其他房屋拥有情况相比，拥有其他类型房屋的人贷款获得批准的概率更高，系数为 0.8129，拥有自己房屋的人贷款获得批准的概率显著降低，系数为-3.012。租房的人贷款获得批准的概率显著提高，系数为 1.091。工作年限每增加 1 年，贷款获得批准的概率略有下降。同时，贷款用于教育的人贷款获得批准的概率显著降低，贷款用于家庭改善的人贷款获得批准的概率略有提高，贷款用于医疗的人贷款获得批准的概率显著降低，系数为-0.2621，用于个人消费的人贷款获得批准的概率显著降低，为-0.6311，用于创业的人贷款获得批准的概率最低，系数为-1.163。

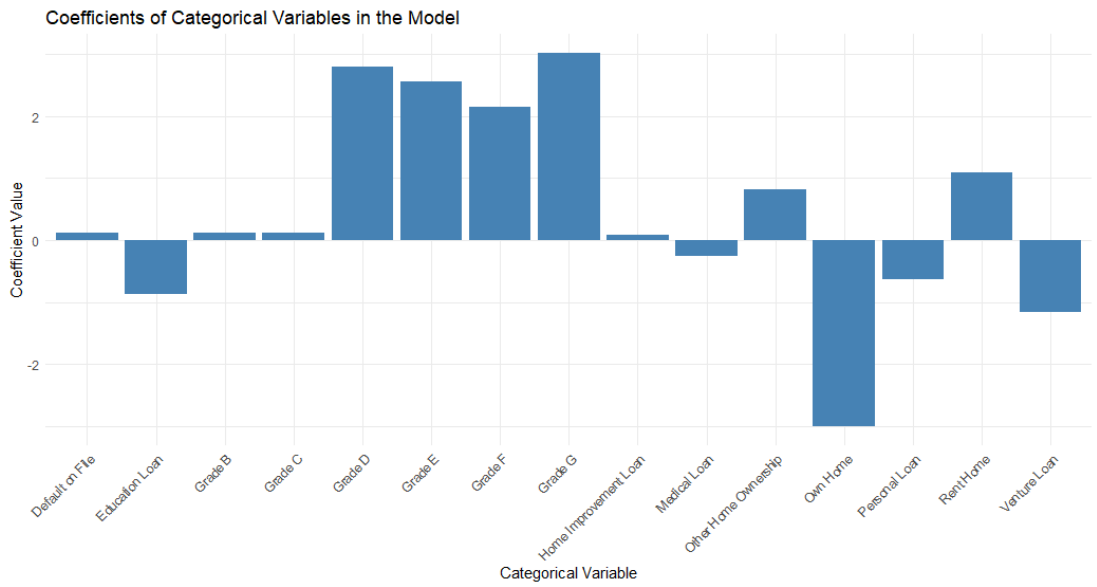

图 4 分类变量系数图

对于贷款信用等级，贷款获得批准的概率基本上随着信用等级的降低而提高。贷款金额每增加 1 美元，贷款获得批准的概率略有下降。贷款利率每增加 1%，贷款获得批准的概率略有提高。贷款金额占收入的比例每增加 1%，贷款获得批准的概率显著提高。

可以得出以下结论：贷款信用等级、贷款金额占收入的比例、租房情况、贷款用于家庭改善等是贷款获得批准的正向预测因素。年收入、工作年限、贷款用于教育、医疗、个人消费、创业等是贷款获得批准的负向预测因素。贷款金额、贷款利率等对贷款获得批准的影响较小。

这是一个违反直觉的现象：信用等级降低，贷款批准概率却提高。这可能是由于数据

质量问题，如异常值或选择偏差，导致模型学习到的关系不准确。另外，分类变量的编码方式可能不当，影响系数解释。多重共线性或模型设定不当也可能是原因，如缺少重要的预测变量或需要非线性关系。最后，业务逻辑的变化，如市场条件或政策调整，也可能影响这一关系。因此，需要进一步的数据审查、模型诊断和业务咨询来准确解释这一现象。根据该数据的分析，原因可能是类型占比不均，而相关类型样本中恰有其它强相关因素数值较大产生。

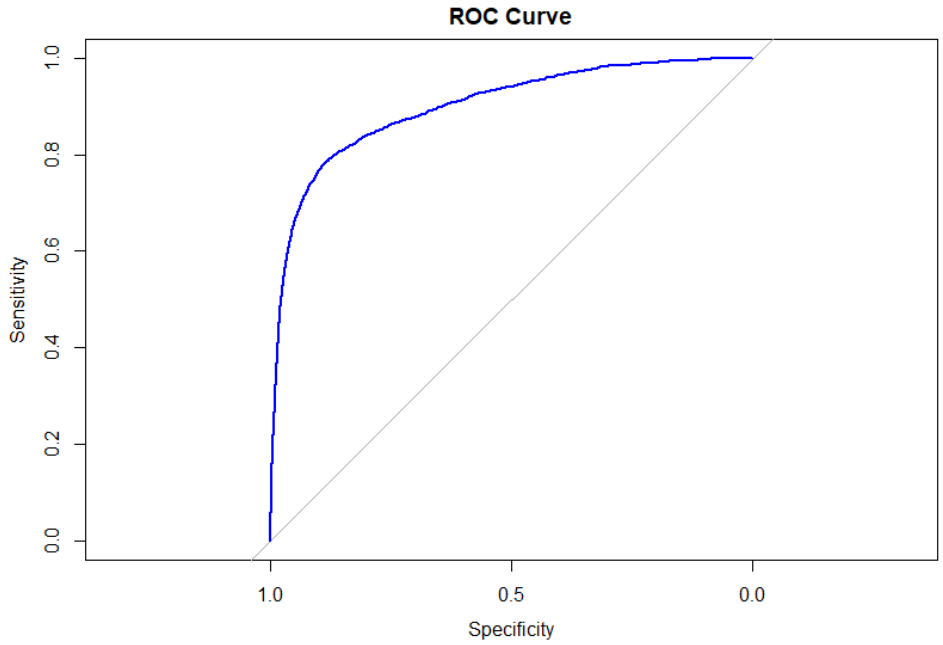

图 ROC 曲线

ROC 曲线如上图所示，AUC 值为 0.9002，意味着模型具有很高的区分能力，能够很好地区分出贷款是否会被批准。利用建立的模型，我们可以对新借款人的贷款批准状态进行预测。通过将新借款人的特征输入模型，我们可以得到其获得贷款批准的概率。这有助于贷款机构评估风险并做出贷款决策。

五、附录：代码

```
loan<-read.csv("loanstats.csv")
str(loan)
summary(loan)
head(loan)
library(ggplot2)
library(reshape2)
library(pheatmap)
library(dplyr)
library(tidyr)
library(ggrepel)
library(caret)
readable_names<-readable_names <- c("Age", "Income", "Employment Length", "Loan Amount", "Interest Rate", "Loan to Income Ratio", "Credit History Length", "Loan Status")
numeric_data <- loan[, sapply(loan, is.numeric) & !(names(loan) %in% c("id"))]
```

```
names(numeric_data) <- readable_names
sum(is.na(loan))

cor_matrix <- cor(numeric_data)
pheatmap(cor_matrix,
             main = "Heatmap of Numeric Variables",
             scale = "none")
long_data <- pivot_longer(numeric_data, cols = everything(), names_to = "Variable",
values_to = "Value")
ggplot(long_data, aes(x = factor(1), y = Value)) +
  geom_boxplot() +
  facet_wrap(~ Variable, scales = "free", nrow = 2, ncol = 4) +
  labs(title = "Boxplots of Numeric Variables", x = "", y = "") +
  theme_minimal() +
  theme(axis.text.x = element_blank(), axis.ticks.x = element_blank(),
          strip.text.x = element_text(size = 9, face = "bold"),
          plot.title = element_text(hjust = 0.5))
process_outliers <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  x[x < lower_bound | x > upper_bound] <- median(x, na.rm = TRUE)
  return(x)}
loan <- loan[, !colnames(loan) %in% c("person_age", "cb_person_cred_hist_length")]
set.seed(123)
trainIndex <- sample(1:nrow(loan), 0.7 * nrow(loan))
trainSet <- loan[trainIndex, ]
testSet <- loan[-trainIndex, ]
model <- glm(loan_status ~ ., data = trainSet, family = "binomial")
predictions <- predict(model, newdata = testSet, type = "response")
library(pROC)
rocObj <- roc(response = testSet$loan_status, predictor = predictions)
auc(rocObj)
plot(rocObj, main = "ROC Curve", col = "blue")
summary(model)
coefficients <- coef(model)
categories <- c("person_home_ownershipOTHER", "person_home_ownershipOWN",
"person_home_ownershipRENT",
                "loan_intentEDUCATION", "loan_intentHOMEIMPROVEMENT",
"loan_intentMEDICAL",
                "loan_intentPERSONAL", "loan_intentVENTURE",
                "loan_gradeB", "loan_gradeC", "loan_gradeD", "loan_gradeE",
```

```
                    "loan_gradeF", "loan_gradeG",
                    "cb_person_default_on_fileY")
coef_values <- coefficients[categories]
category_data <- data.frame(
  Category = names(coef_values),
  Estimate = as.numeric(coef_values),
  ReadableCategory = c("Other Home Ownership", "Own Home", "Rent Home",
                       "Education Loan", "Home Improvement Loan", "Medical Loan",
                       "Personal Loan", "Venture Loan",
                       "Grade B", "Grade C", "Grade D", "Grade E",
                       "Grade F", "Grade G",
                       "Default on File"))
ggplot(category_data, aes(x = ReadableCategory, y = Estimate)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Coefficients of Categorical Variables in the Model",
       x = "Categorical Variable", y = "Coefficient Value") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) # 旋转 X 轴标签以便阅读
library(RColorBrewer)
categorical_vars <- loan[sapply(loan, is.character)]
super_pretty_colors <- c("#FF6384", "#36A2EB", "#FFCE56", "#4BC0C0", "#9966FF",
"#FF9F43", "#2ED573", "#E7E9BD")
# 定义变量的描述性名称
variable_titles <- c(
  "person_home_ownership" = "Home Ownership",
  "loan_intent" = "Loan Intent",
  "loan_grade" = "Loan Grade",
  "cb_person_default_on_file" = "Default on File")
plot_pie <- function(data, variable, title) {
  freq <- table(data[[variable]])
  freq_df <- as.data.frame(freq)
  names(freq_df) <- c("Category", "Freq")
  if (missing(title) || is.na(title)) {
    title <- paste("Pie Chart of", variable)
  }
  ggplot(freq_df, aes(x = "", y = Freq, fill = Category)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar(theta = "y") +
    scale_fill_manual(values = super_pretty_colors[1:min(length(freq_df$Category),
length(super_pretty_colors))]) + # 应用超级好看的颜色
    labs(title = title, x = NULL, y = NULL) +
    theme_void() +
    theme(legend.title = element_blank()) }
lapply(names(categorical_vars), function(var) {
```

```
    title <- variable_titles[var]
    plot_pie(loan, var, title)
})
```