大数据引论中期汇报

银行信用卡用户流失情况数据分析

数据集来源：https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m

刘琮璟　　徐朱玮

目录 CATALOGUE

01

背景介绍

INTRODUCTION OF THE BACKGROUND

01 背景介绍——数据集简介

INTRODUCTION OF THE BACKGROUND

数据集来源于Kaggle，包含了10127个银行信用卡用户的23个特征，其中包括1627个已经流失的用户和8500个未流失的用户。

链接：https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m

这一数据集包含某银行从消费信用卡组合中收集的丰富客户信息。这23个特征包括全面的人口统计信息，如年龄、性别、附属卡数量（可近似家属人数）、受教育程度、婚姻状况和收入类别；以及每位客户与信用卡提供商关系的信息，如卡片类型、与银行交互的频率、持有银行产品的总数、过去一年中的不活跃月份数等。此外，它还包含了关于客户流失前消费行为的关键数据，如信用额度、总循环余额、过去12个月开放购买的信用额度；以及其它一些可分析指标如第4到第1季度的总变化金额、过去12个月的总交易额、过去12个月的总交易数、平均利用率和朴素贝叶斯分类器的流失标志(信用卡类别与12个月期间的联系人数量、依赖数量、教育水平和不活跃月份相结合)。

01 背景介绍——选题意义

INTRODUCTION OF THE BACKGROUND

- 帮助银行阻止客户流失

如果有一个良好的分类模型能够通过对信用卡用户各项特征的分析，提前预知该客户是否在近期可能流失，那么银行就能够在客户流失前采取行动、主动联系客户并为他们提供更好的服务，以期客户能够留下。

我们希望能够做出这样一个较好的模型，随着输入数据集的改变能应用于包括银行在内的需要维持客户的业务，如美容店、保险业等。

- 利于银行投资组合的稳定

一般银行的资金会用于进行再投资，以此获得收益。倘若某信用卡用户在前6个月每月向银行贷款20万，但是在第7个月只贷款了15万，第8个月贷款了10万，但是第9个月的时候银行仍然判断该客户会持续贷款20万，而实际第9个月的时候该客户流失了，那么就会影响银行投资组合的决定。

因此我们希望帮助银行提前预判以做出更为明智和稳健的决策。

预期目标

首先，我们希望能够找出不同特征对于客户是否会流失的重要程度排序，并且获得其中相关性较高的特征，进行初步分析。

然后，我们希望能够采用不同算法对该数据集做分类预测。目前我们计划采用的模型有：逻辑回归、KNN、决策树、随机森林、SVM和神经网络。

最后，我们将对不同模型应用于该数据集的预测性能做比较，并挑选找到预测准确率最高的模型。

挖掘每个特征的重要程度
并找到相关性最高的特征

采用不同算法做分类预测

对不同模型的性能做比较
并挑选找到最好的模型

02

数据预处理

DATA PREPROCESSING

02

数据预处理——缺失值处理和去除无效特征

DATA PREPROCESSING

缺失值处理

首先调用python中的info()函数对导入的数据集data做缺失值查询，结果如右图所示。

可以看到，除了用户流失与否、性别、受教育程度、婚姻状况、收入情况、信用卡类型的数据为object以外，其它数据都为整数或浮点数。

且该数据集没有任何缺失值。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

去除无效特征

经分析，在所有的特征中：“CLIENTNUM”（客户编号）、以及最后两列朴素贝叶斯（即通过朴素贝叶斯分析客户是否会按照某些特定特征流失）是无效的特征，因此可以去掉。

去除后，这一数据集变为10127行、20列，也即包含了10127个银行信用卡用户的信息，每个用户有20个相关特征，且无缺失值。

02 数据预处理——对数据集的简单考察

DATA PREPROCESSING

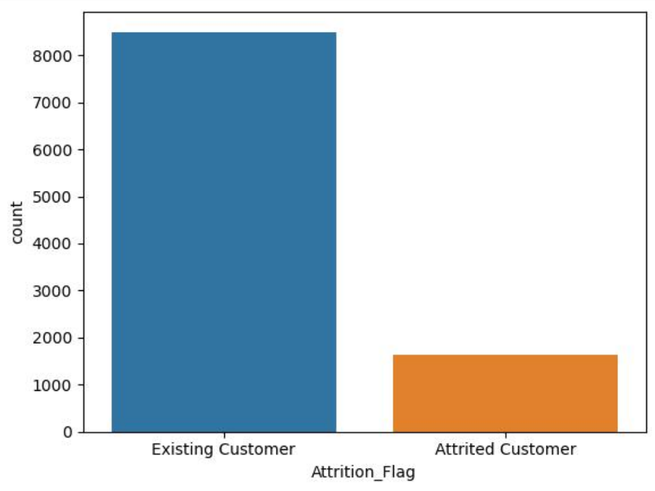


图1：流失客户和未流失客户数量柱状图

由上图可以看到，在所有数据中，未流失的客户要远多于流失的客户，两者差异很大。因此在模型训练的时候，可能会由于流失客户的样本过小，而导致最后的训练结果准确率不高。

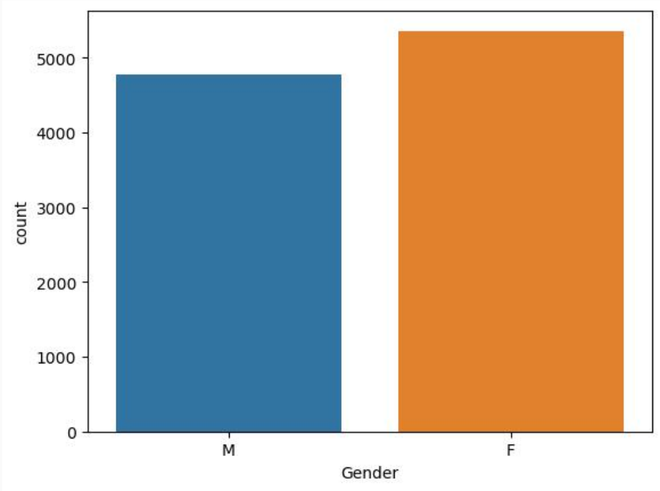


图2：男性客户和女性客户数量柱状图

从上图可知，虽然总体而言女性客户较多，但是男女客户的性别比例较为均衡。因此性别因素不会对最后的训练产生影响。也就是说，不论是男性客户还是女性客户，模型都将较好地对其做预测。

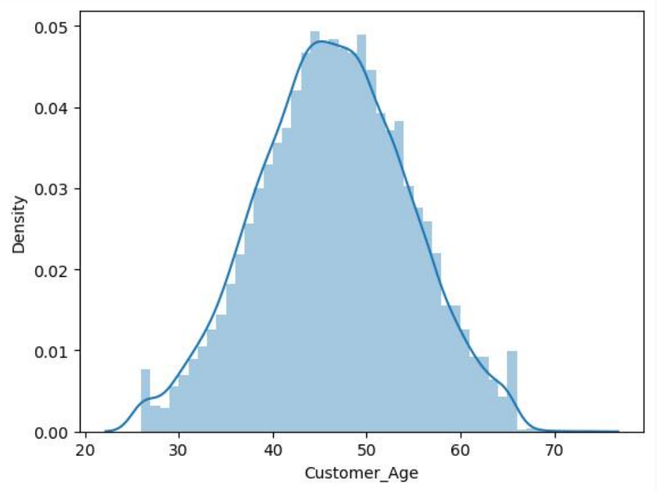


图3：用户年龄分布图

由图可知，该银行信用卡用户的年龄大致成正态分布。因此，可以在正态分布假设下进一步使用年龄特征。

02

数据预处理——数据集格式转换

DATA PREPROCESSING

Attrition_Flag	Attrited Customer（流失客户）：0	Existing Customer（未流失客户）：1
Gender	F（女性）：0	M（男性）：1
Education_Level	Unknown（未知）：1	Uneducated（未受教育）：2
	High School（高中）：3	College（本科在读）：4
	Graduate（本科毕业）：5	Post-Graduate（硕士）：6
	Doctorate（博士）：7	
Marital_Status	Unknown（未知）：0	Single（单身）：1
	Divorced（离异）：2	Married（已婚）：3
Income_Category	Unknown（未知）：1	Less than $40K（少于4万美金）：2
	$40K - $60K（四万-六万美金）：3	$60K - $80K（六万-八万美金）：4
	$80K - $120K（8万-12万美金）：5	$120K +（大于12万美金）：6
Card_Category	Blue（蓝卡）：1	Silver（银卡）：2
	Gold（金卡）：3	Platinum（铂金卡）：4

由于在训练模型时需要输入数字，因此我们将数据为“object”的列全部转化为数字，数字和含义的对应如上图所示。进行格式转化的列分别有“客户流失与否”、“性别”、“受教育程度”、“婚姻状况”、“收入状况”、“信用卡类别”。

02 数据预处理——特征归一化

DATA PREPROCESSING

	0	1	2	3
Attrition_Flag	1.000	1.000	1.000	1.000
Customer_Age	45.000	49.000	51.000	40.000
Gender	1.000	0.000	1.000	0.000
Dependent_count	3.000	5.000	3.000	4.000
Education_Level	3.000	5.000	5.000	3.000
Marital_Status	3.000	1.000	3.000	0.000
Income_Category	4.000	2.000	5.000	2.000
Card_Category	1.000	1.000	1.000	1.000
Months_on_book	39.000	44.000	36.000	34.000
Total_Relationship_Count	5.000	6.000	4.000	3.000
Months_Inactive_12_mon	1.000	1.000	1.000	4.000
Contacts_Count_12_mon	3.000	2.000	0.000	1.000
Credit_Limit	12691.000	8256.000	3418.000	3313.000
Total_Revolving_Bal	777.000	864.000	0.000	2517.000
Avg_Open_To_Buy	11914.000	7392.000	3418.000	796.000
Total_Amt_Chng_Q4_Q1	1.335	1.541	2.594	1.405
Total_Trans_Amt	1144.000	1291.000	1887.000	1171.000
Total_Trans_Ct	42.000	33.000	20.000	20.000
Total_Ct_Chng_Q4_Q1	1.625	3.714	2.333	2.333
Avg_Utilization_Ratio	0.061	0.105	0.000	0.760

左图展示了数据集中的部分数据。可以看到，不同特征的值跨越的范围很大，这会在模型训练中导致权重不一致的情况。

因此，需要对数据集做特征归一化。也就是将不同特征的取值范围统一到相同的尺度上，这里将特征的取值范围缩放到[0, 1]之间，以避免某些特征对模型的影响过大。

归一化之后的相应数据如右图所示。归一化后虽然特征的值处于[0, 1]之间，但是每一个特征的分布仍然是不变的。这不会影响特征本身，但更利于训练。

	0	1	2	3
Attrition_Flag	1.000000	1.000000	1.000000	1.000000
Customer_Age	0.404255	0.489362	0.531915	0.297872
Gender	1.000000	0.000000	1.000000	0.000000
Dependent_count	0.600000	1.000000	0.600000	0.800000
Education_Level	0.333333	0.666667	0.666667	0.333333
Marital_Status	1.000000	0.333333	1.000000	0.000000
Income_Category	0.600000	0.200000	0.800000	0.200000
Card_Category	0.000000	0.000000	0.000000	0.000000
Months_on_book	0.604651	0.720930	0.534884	0.488372
Total_Relationship_Count	0.800000	1.000000	0.600000	0.400000
Months_Inactive_12_mon	0.166667	0.166667	0.166667	0.666667
Contacts_Count_12_mon	0.500000	0.333333	0.000000	0.166667
Credit_Limit	0.340190	0.206112	0.059850	0.056676
Total_Revolving_Bal	0.308701	0.343266	0.000000	1.000000
Avg_Open_To_Buy	0.345116	0.214093	0.098948	0.022977
Total_Amt_Chng_Q4_Q1	0.392994	0.453636	0.763615	0.413600
Total_Trans_Amt	0.035273	0.043452	0.076611	0.036775
Total_Trans_Ct	0.248062	0.178295	0.077519	0.077519
Total_Ct_Chng_Q4_Q1	0.437534	1.000000	0.628164	0.628164
Avg_Utilization_Ratio	0.061061	0.105105	0.000000	0.760761

02 数据预处理——训练集和测试集的划分

DATA PREPROCESSING

```
from sklearn.model_selection import train_test_split
# 将特征列赋值给X，将目标变量（流失与否）赋值给y
X = data.drop('Attrition_Flag', axis=1)
y = data['Attrition_Flag']
# 使用分层抽样划分训练集和测试集，设置stratify参数为y，确保训练集和测试集中的流失和非流失客户的比例与整个数据集中的比例相似
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, stratify=y, random_state=42)
```

```
file_path = f"D:\YOLANDA\FuDan\课程\大四上\大数据引论\PJ_new\BankChurners.csv"
data = pd.read_csv(file_path)
Attrition_Flag = list(set(list(data['Attrition_Flag'])))
num_attrited = len(data[data['Attrition_Flag'] == 'Attrited Customer'])
num_existing = len(data[data['Attrition_Flag'] == 'Existing Customer'])
print('用户流失数：{}'.format(num_attrited))
print('用户未流失数：{}'.format(num_existing))
print('未流失用户占比：{}'.format(num_existing / (num_attrited + num_existing)))
```

```
用户流失数：1627
用户未流失数：8500
未流失用户占比：0.8393403772094401
```

```
num_y_existing = len(y_train[y_train==1.0])
num_y_all = len(y_train)
print('训练集中用户未流失数：{}'.format(num_y_existing))
print('训练集中未流失用户占比：{}'.format(num_y_existing / num_y_all))
```

```
训练集中用户未流失数：6799
训练集中未流失用户占比：0.8392791013455129
```

由于实验目标是完成客户“流失”还是“未流失”的二分类预测，但是在总体数据中，未流失的用户数量有8500，已经流失的用户数量只有1627，两者的比例严重不均衡。

为了增加模型的鲁棒性，在划分训练集和测试集时，我们采用分层抽样的方式，也即设置stratify参数为y。由此，在训练集和测试集中，两类的比例都和原数据集中两类的比例相同。我们希望可以通过这个方式减少两类数据巨大差值对模型的影响。

02 数据预处理——探索性分析

DATA PREPROCESSING

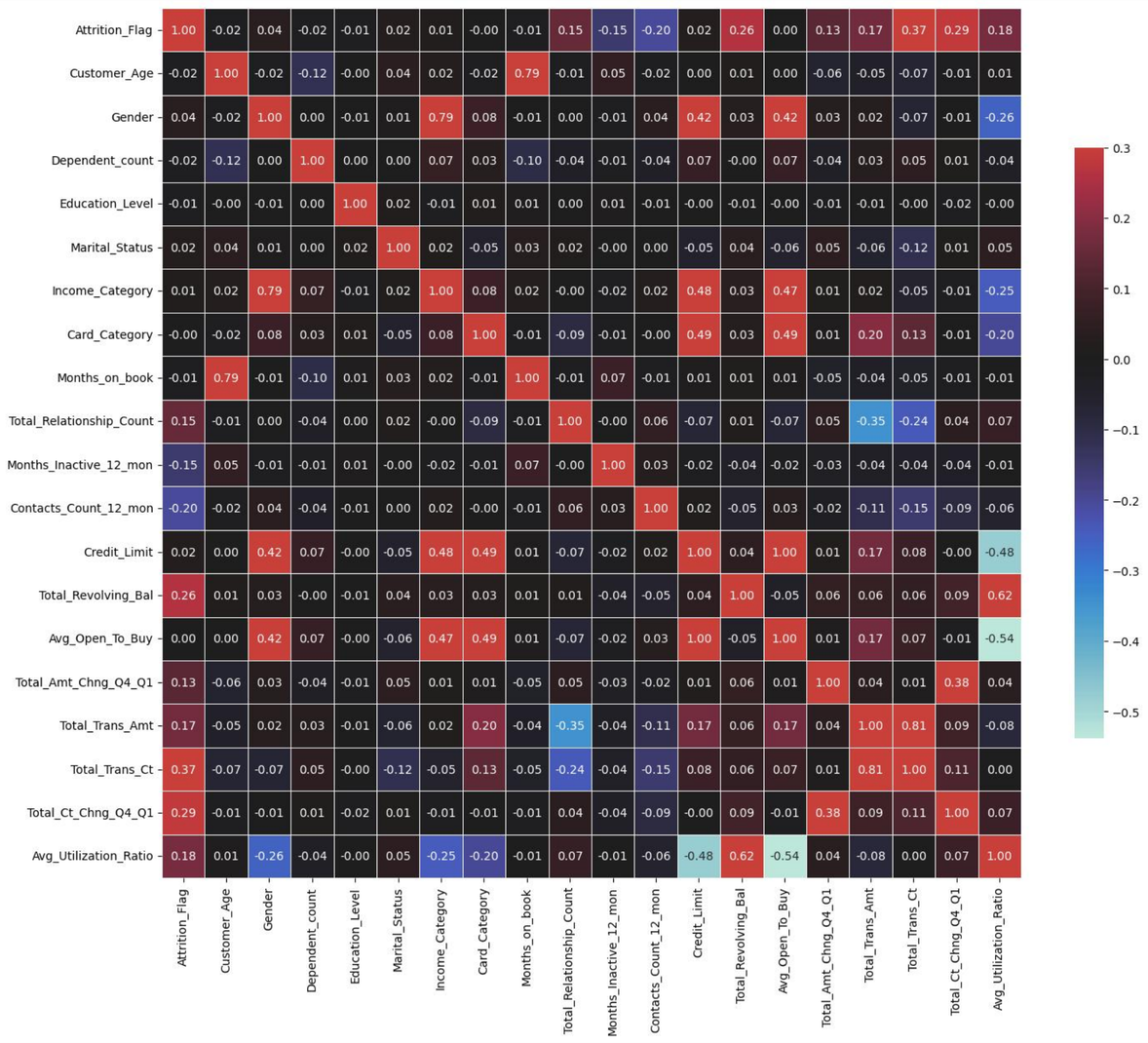


图4：各特征间的相关性系数图

左图展示了20个特征之间的相关性系数，方格的颜色越接近于红色，就表明特征的正相关程度越高；方格的颜色越接近于浅蓝色，就表明特征的负相关程度越高。

要判断“Attrition_Flag”（客户是否流失）与哪一组特征的相关性更高，只需要看第一行或者第一列。通过观察可知，总体而言，过去12个月与银行交互的数量与客户是否流失的负相关性较高；而过去12个月的总交易数和以及第4季度相比第1季度的交易数量变化，与客户是否流失的正相关性较高。

在初步了解特征与客户是否流失的相关性后，就可以大致知晓特征的重要程度。因此在之后的模型训练中，可以给较为重要的特征更大的权重。

03

模型的评价指标

EVALUATION INDEX OF MODELS

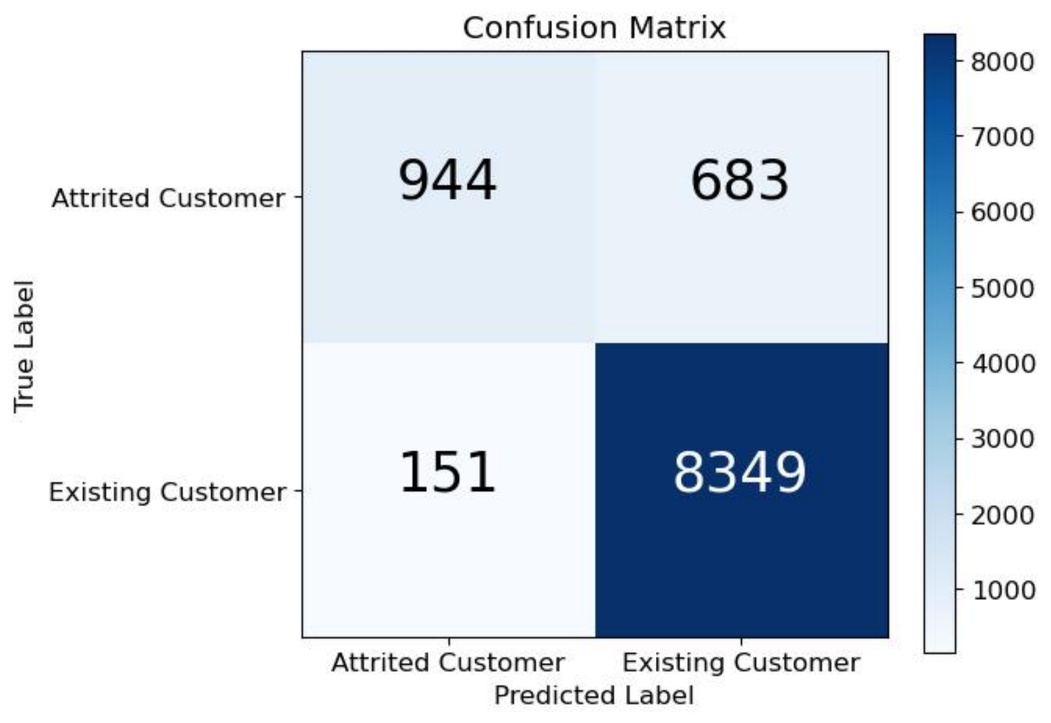


图5：混淆矩阵示例（SVM模型的混淆矩阵）

混淆矩阵

混淆矩阵是用来评价模型的一个重要指标，它分为TP（真阳性）、FP（假阳性）、FN（假阴性）、TN（真阴性）。应用到这一数据集上：

左上角：预测是流失的客户，结果也是流失的客户

右上角：预测是未流失的客户，结果是流失的客户

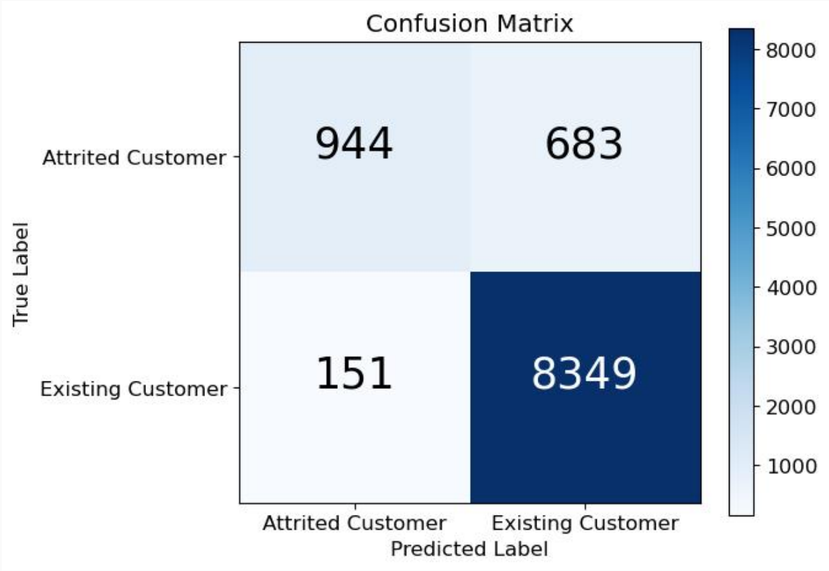
左下角：预测是流失的客户，结果是未流失的客户

右下角：预测是未流失的客户，结果也是未流失的客户

这一矩阵可以帮助我们清晰地看到模型的预测结果，方便我们对模型进行判断。

03 模型的评价指标——评价指标之精确率和召回率

EVALUATION INDEX OF MODELS



	Attrition_Flag	precision	recall	accuracy	F1-score
0	Attrited Customer	86.210	58.0209	91.7646	0.693608
1	Existing Customer	92.438	98.2235	91.7646	0.952430

精确率(presicion)：TP/(TP+FP)

如：流失用户的精准率=（预测流失真实也流失944）/（预测流失真实也流失944+预测流失真实未流失151）=86.210%

含义：在模型预测为流失的所有用户中，真正流失的用户所占的比例。

在这里可以看到，模型预测为流失的用户中，真正流失的用户占86.210%。

召回率(recall)：TP/(TP+FN)

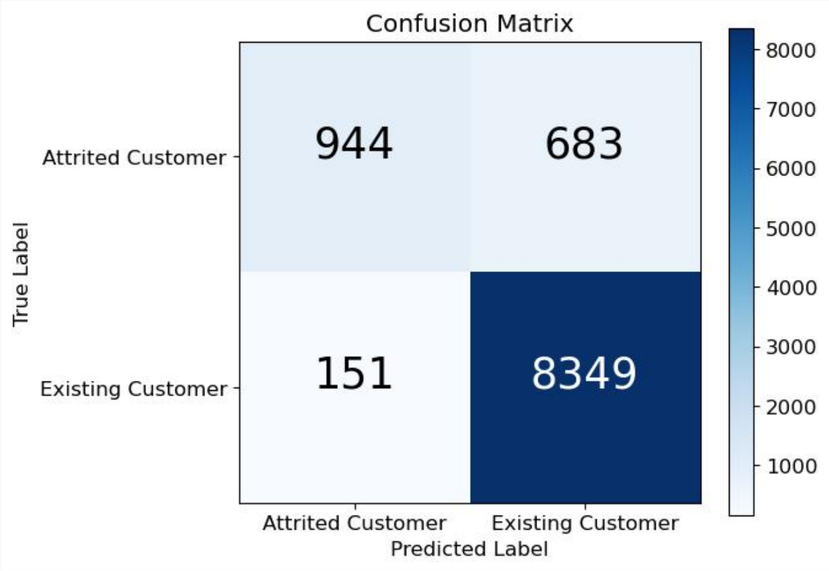
如：未流失用户的召回率=（预测未流失真实也未流失8349）/（预测未流失真实也未流失8349+预测流失真实未流失151）=98.2235%

含义：在所有真正未流失的用户中，被预测为未流失的用户所占的比例。这一概率越高，代表错失的个数越少。

在这里，在所有没有流失的用户中，被准确预测到的用户占98.2235%。

03 模型的评价指标——评价指标之正确率和F1 Score

EVALUATION INDEX OF MODELS



	Attrition_Flag	precision	recall	accuracy	F1-score
0	Attrited Customer	86.210	58.0209	91.7646	0.693608
1	Existing Customer	92.438	98.2235	91.7646	0.952430

正确率(accuracy)：(TP+TN)/ALL

如：正确率=（预测流失真实也流失944+预测未流失真实也未流失8349）/（总样本10127）=91.7646%

含义：在所有样本中，预测值和真实值一致的样本所占的概率。

在这里可以看到，被准确预测的用户占总数的91.7646%。

F1 Score：2 (precision * recall) / (precision + recall)

含义：也就是调和平均数（即P和R的倒数之和的1/2的倒数）。只有在P和R都比较好的时候，F1 Score才有可能比较高；如果P和R中仅有一者较好，而另一者与其相差较大，那么F1 Score也不会很高。

在这里可以看到，相比于流失的客户，该模型（SVM）对未流失的客户的预测效果是更好的。

03 模型的评价指标——评价指标之ROC曲线

EVALUATION INDEX OF MODELS

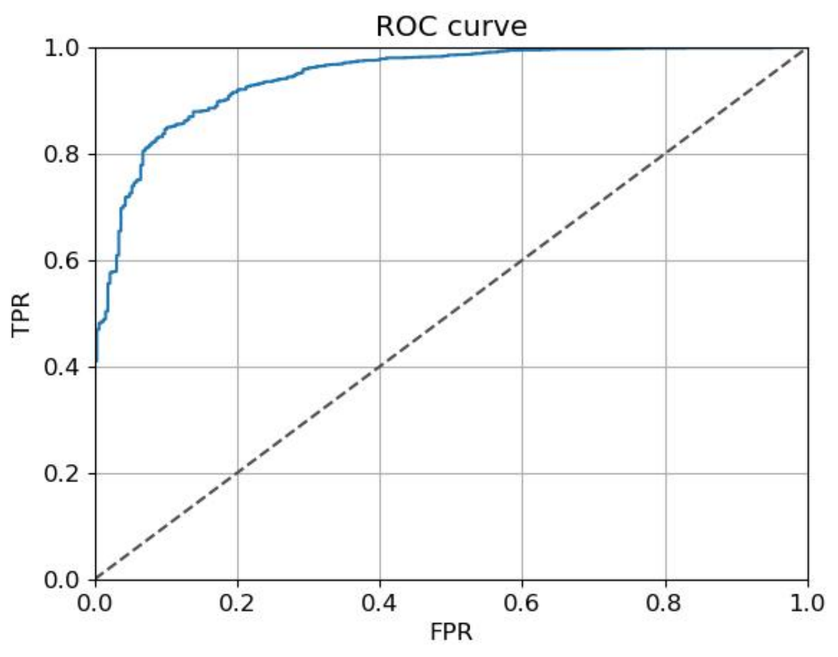


图6：ROC曲线示例（SVM模型的混淆矩阵）

ROC曲线（Receiver Operating Characteristic curve）是一种用于评估分类模型性能的图形工具。

它的横坐标为假阳性率（False Positive Rate），表示错误地判断为正例的概率（错误地预测为正的数量/原本为负的数量）；纵坐标为真阳性率（True Positive Rate，即precision）（正确地预测为正的数量/原本为正的数量），表示正确地判断为正例的概率。

通常，我们认为曲线的凸起程度越高，模型准确率越好。图中的虚线是对角线，表示随即猜测，因此ROC曲线越接近对角线，则模型的预测率越低。

ROC曲线下方的面积称为AUC，一般来说，AUC越大、分类器越好。AUC为0.5表示随机猜测。

04

后续研究计划安排

FOLLOW-UP RESEARCH PLAN ARRANGEMENT

04 后续研究计划安排

FOLLOW-UP RESEARCH PLAN ARRANGEMENT

第13周（11月28日）：中期汇报

- 完成数据预处理
- 完成模型评价指标制定
- 完成一个模型的训练（SVM）

第14周（12月初）

- 完成其它模型的训练（决策树、随机森林、朴素贝叶斯、KNN、逻辑回归、神经网络）
- 开始撰写研究报告

第15周（12月中旬）

- 对完成训练的模型做评估和分析，找到性能最佳的模型
- 对研究进行总结
- 完成70%研究报告撰写

第16周（12月下旬）：

- 完成最终课程论文撰写并提交报告

请老师和同学们批评指正

THANK THE TEACHER AND STUDENTS FOR THEIR CRITICISM AND CORRECTION

徐朱玮

刘琮璟