电信用户深度挖掘与多类型预测模型构建：从总收入回归到行为分类

数据挖掘原理与技术第3次个人作业

汇报人：钟忠权

时间：2025.11.17

目 录

CONTENTS

01

项目背景与数据介绍

数据基本情况

记录数：8999，字段数：23

缺失值概览

	缺失数量	缺失比例(%)
最近使用操作系统偏好	1010	11.22
终端品牌	465	5.17
年龄	219	2.43
终端类型	4	0.04
换机频率	1	0.01
入网渠道类型	1	0.01
渠道类型描述	1	0.01
上网流量使用	1	0.01
漫游流量使用	1	0.01
总收入	1	0.01
增值收入	1	0.01
流量收入	1	0.01
短信收入	1	0.01
彩信收入	1	0.01
语音收入	1	0.01
是否欠费	1	0.01
产品大类	1	0.01
产品分类	1	0.01

	A	B	C	D	E	F	G	H	I	J
1	用户ID	性别	年	归属	客户类型	在网时	换机频率	终端品牌	终端类	最近使用操作系统偏
2	00891A97F	女	20	石家庄	公众客户	15	频繁换机型	LG	4G	ANDROID 5.1.0
3	00F84124D	女	63	北京	公众客户	70	频繁换机型	苹果	3G	IOS 6.0
4	0129B973E	男	21	北京	公众客户	33	频繁换机型	三星	4G	ANDROID 5.1.1
5	015B82EFB	无法区分		北京	集团-中小企	1	频繁换机型	华为	4G	ANDROID 5.0.2
6	01EA77A21	男	32	北京	公众客户	58	频繁换机型	黑莓	3G	BLACKBERRY
7	01EBB91E4	男	20	天津	公众客户	45	频繁换机型	苹果	4G	IOS 9.0
8	02771333B	男	23	北京	公众客户	45	频繁换机型	摩托罗拉	4G	ANDROID 5.0.1
9	02D556F6A	男	47	天津	公众客户	47	频繁换机型	华为	4G	ANDROID 5.1.1
10	02E00D5F4	男	34	北京	公众客户	24	频繁换机型	苹果	4G	IOS 8.0
11	03C44E8DE	女	18	北京	公众客户	1	频繁换机型	魅族	4G	
12	03FB7FC1E	男	26	邯郸	公众客户	2	频繁换机型		0	
13	041A728C4	女	45	承德	公众客户	65	频繁换机型	小米	4G	ANDROID 4.4.2
14	042D41E3E	男	19	天津	公众客户	1	频繁换机型	三星	4G	ANDROID 4.3
15	0486E0CC7	男	28	北京	公众客户	33	频繁换机型	苹果	4G	IOS 9.0
16	055E74FA1	男	38	邯郸	公众客户	1	频繁换机型		0	
17	056BC202E	女	19	北京	公众客户	36	频繁换机型	苹果	3G	IOS 6.0
18	06BA8104F	男	25	天津	公众客户	32	频繁换机型	苹果	4G	IOS 8.0
19	07B8F8288	男	21	承德	公众客户	7	频繁换机型	三星	4G	ANDROID 4.0.4

公众客户样本数：8846 (占全部用户 98.30%)。

（后续任务统一使用这个子集）

目标：多任务建模（1 回归 + 4 分类），尽量提高预测的准确度。

数据基本情况

连续变量（10个）

变量名	说明
年龄	用户年龄
在网时长	用户入网时长，对收入与欠费均有影响
上网流量使用	明显右偏，存在极端值
漫游流量使用	右偏，大部分为 0
总收入	每月总收入**（任务 1 回归目标）**
增值收入	子项
流量收入	子项
短信收入	子项
彩信收入	子项
语音收入	子项

将明显的数值列转为 numeric（错误转为 NaN）

分类变量（13个）

变量名	说明
用户ID	用户唯一标识（不参与建模）
性别	男 / 女 / 无法区分
归属地	用户所属地区
客户类型	客户类型**（公众客户为主要分析对象）**
换机频率	频繁换机/偶尔换机/从不换机
终端品牌	手机品牌（高基数类别）
终端类型	2G / 3G / 4G**（任务 3 预测目标）**
最近使用操作系统偏好	ANDROID / IOS**（任务 5 目标）**
入网渠道类型	用户入网方式：1/2/99
渠道类型描述	社会渠道 / 自有渠道**（任务 4 目标）**
是否欠费	是 / 否**（二分类，严重不平衡，任务 2 目标）**
产品大类	电信产品的大类
产品分类	产品细分类型（高基数类别）

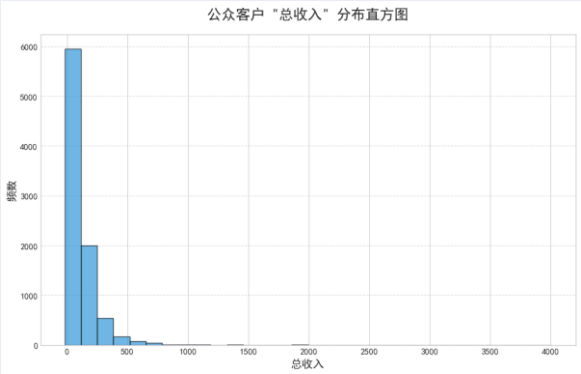
02

探索性数据分析与特征工程

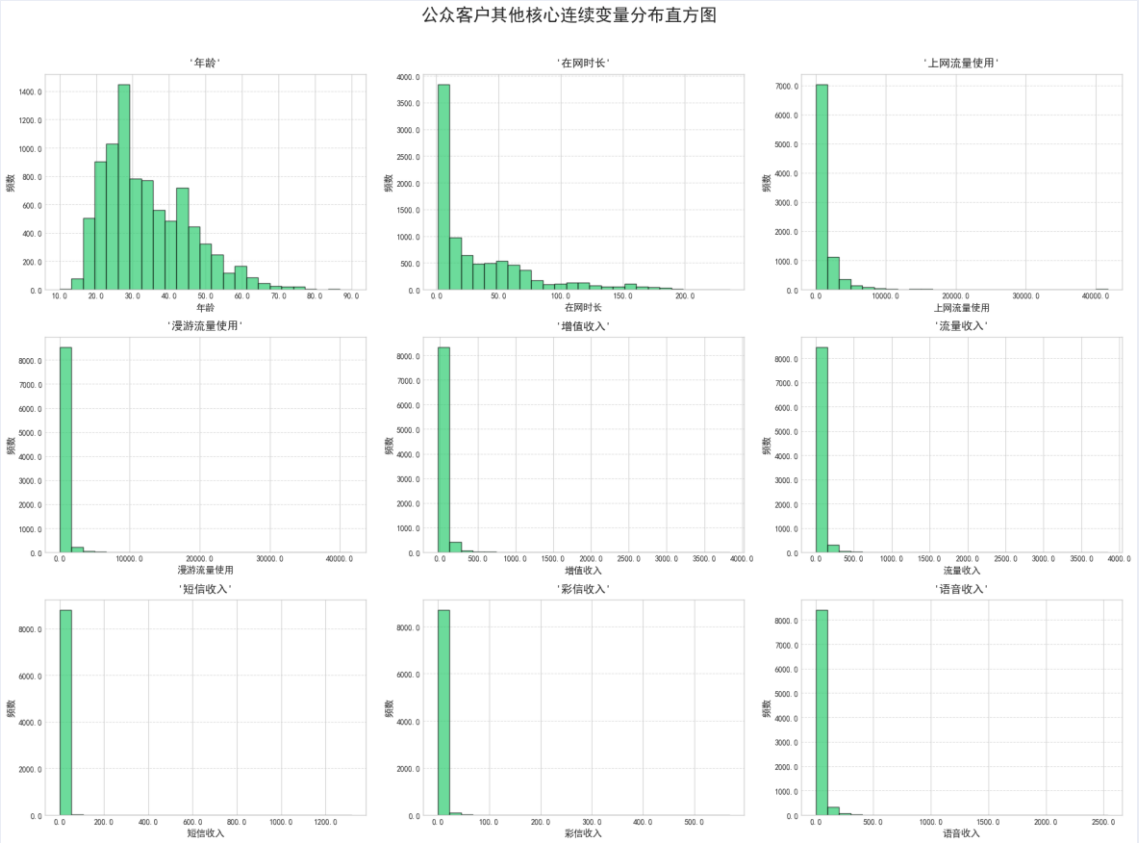
基于公众客户样本

连续变量

公众客户样本数：8846 (占全部用户 98.30%)

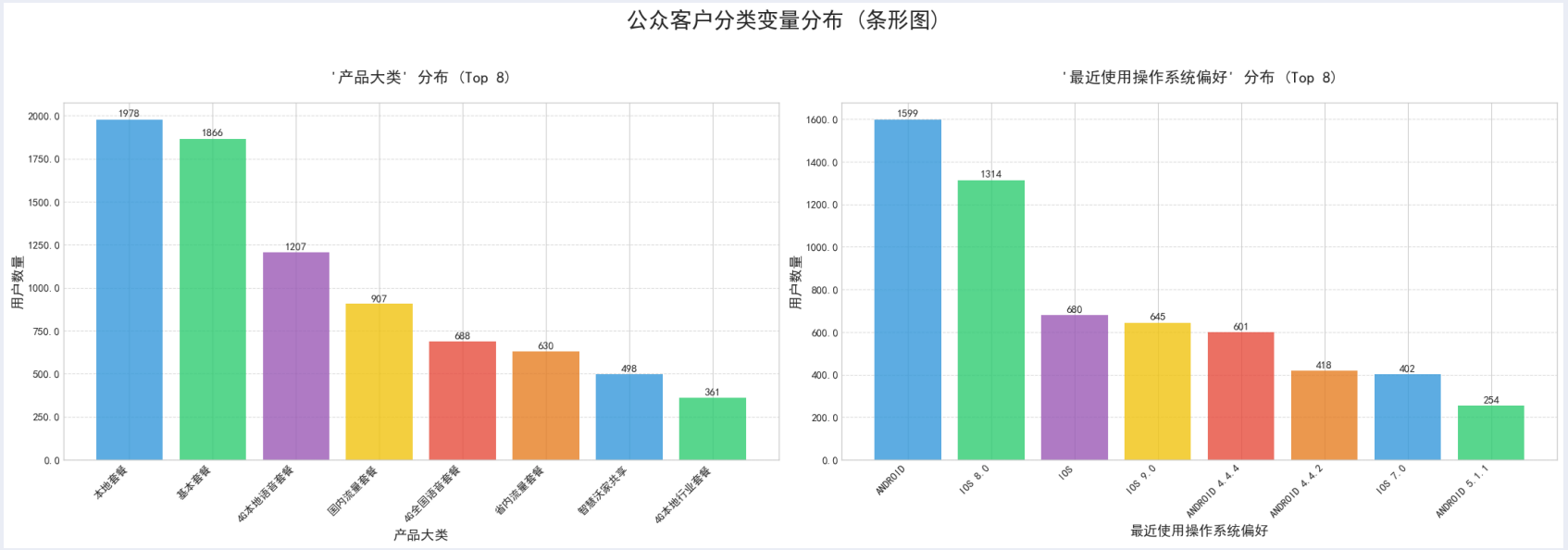


除年龄外，基本严重右偏



分类变量情况

公众客户样本数：8846 (占全部用户 98.30%)

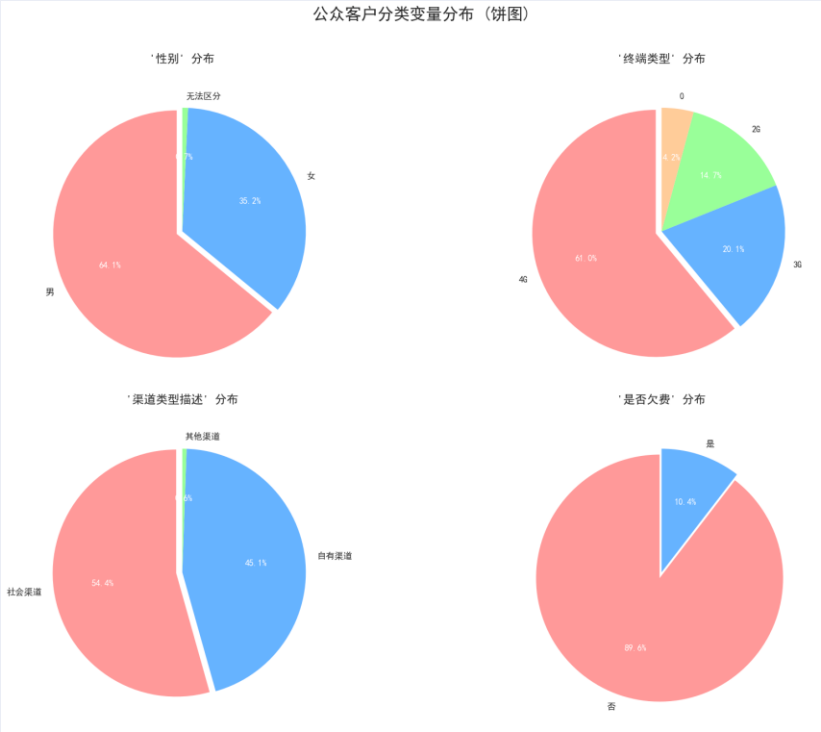


“产品大类”中，本地套餐最多；

“最近使用操作系统偏好”中有许多相同系统的具体版本数据，后续需要合并。

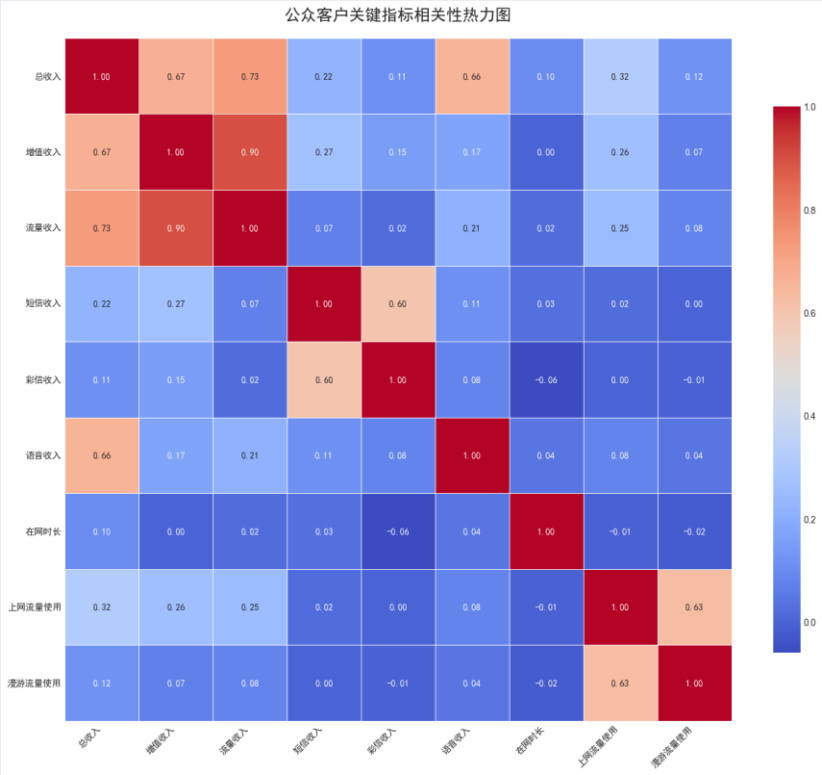
分类变量情况

公众客户样本数：8846 (占全部用户 98.30%)



“是否欠费”类别严重不平衡

连续变量相关性分析



收入数据之间、上网流量使用和漫游流量使用之间相关性强，符合直观

- **强正相关：**

增值收入与流量收入（0.90）：用户流量使用越多，增值服务消费也越高

总收入与流量收入（0.73）：流量收入是总收入的重要组成部分

短信收入与彩信收入（0.66）：使用短信的用户更可能同时使用彩信服务

- **弱相关 / 无显著关联：**

在网时长与多数收入指标相关性微弱，说明用户忠诚度对收入影响不明显

漫游流量使用与其他指标关联度低，表明漫游业务目前贡献有限

- **唯一弱负相关：**

在网时长与彩信收入（-0.06）、漫游流量（-0.02），相关性几乎可忽略

特征工程

统一特征工程：

- 删除指定列
- 划分 X, y
- 数值缺失：中位数填补
- 类别缺失：填 'Unknown’
- 类别：One-hot 编码

```python
# 1. 划分 X, y
if target_col:
    y = df[target_col]
    X = df.drop(columns=[target_col] + drop_cols)
else:
    y = None
    X = df.drop(columns=drop_cols)

# 2. 缺失值处理
num_cols = X.select_dtypes(include=[np.number]).columns
cat_cols = X.select_dtypes(include=["object"]).columns

# 数值：中位数填补
X[num_cols] = X[num_cols].fillna(X[num_cols].median())

# 类别：填 "Unknown"
X[cat_cols] = X[cat_cols].fillna("Unknown")

# 3. One-hot 编码
X = pd.get_dummies(X, drop_first=True)
ohe_cols = X.columns.tolist()
```

03

任务1

使用线性回归模型和随机森林预测总收入，并汇报RMSE和MAE

下列变量除外：[增值收入],[流量收入],[短信收入],[彩信收入],[语音收入]

线性回归模型

目标变量“总收入”存在1个缺失值，已用中位数（74.20）填充

数值变量多重共线性（VIF）检测：

	变量	VIF
0	年龄	1.985794
1	在网时长	1.854073
2	上网流量使用	1.904400
3	漫游流量使用	1.701312

均小于10

模型：

总收入 ~ 年龄 + 在网时长 + 上网流量使用 + 漫游流量使用 + 性别_女+ 性别_男……

训练集 RMSE: 138.63，MAE: 61.50

测试集 RMSE: 117.40，MAE: 64.85

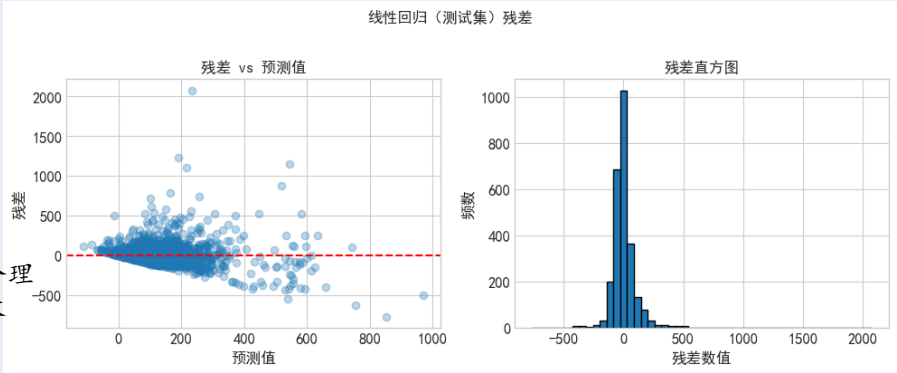
模型基本假设满足：残差无明显趋势，说明线性模型对数据的整体拟合形式合理

预测精度特征：对中低收入样本预测较稳定，对高收入样本预测误差波动更大

```
rf_exclude_cols = ["用户ID", "客户类型", "增值收入", "流量收入", "短信收入", "彩信收入", "语音收入"]

y_rf = df["总收入"]
X_rf = df.drop(columns=["总收入"] + rf_exclude_cols)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   总收入   R-squared:                       0.373
Model:                            OLS   Adj. R-squared:                  0.338
Method:                 Least Squares   F-statistic:                     10.79
Date:                Sat, 15 Nov 2025   Prob (F-statistic):               0.00
Time:                        18:06:37   Log-Likelihood:                -39324.
No. Observations:                6192   AIC:                         7.930e+04
Df Residuals:                    5868   BIC:                         8.148e+04
Df Model:                         323
Covariance Type:            nonrobust
```



随机森林

```
rf_exclude_cols = ["用户ID", "客户类型", "增值收入", "流量收入", "短信收入", "彩信收入", "语音收入"]

y_rf = df["总收入"]
X_rf = df.drop(columns=["总收入"] + rf_exclude_cols)
```

GridSearchCV调参
最佳参数： {'max_depth': 12, 'max_features': 0.5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300}

训练 RMSE: 81.70, MAE: 39.75
测试 RMSE: 104.89, MAE: 52.58

随机森林特征重要性 TOP5：

	特征	重要性
3	上网流量使用	0.283217
4	漫游流量使用	0.132042
357	产品分类_596元基本套餐（含国内语音3000分钟，国内流量11GB）	0.097006
1	在网时长	0.068570
0	年龄	0.065395

最终结论：随机森林 在训练集（拟合能力）和测试集（预测能力）均表现更好

04

任务2-5

任选3个模型进行分类

模型选择思路

- 鉴于本项目数据集特征包含大量类别型变量（如终端品牌、套餐类型）及右偏的连续变量（如流量使用、收入等），**随机森林、XGBoost、逻辑回归普遍表现优于SVM、神经网络、朴素贝叶斯、 k 近邻算法。**
- **数据特点：高维 one-hot 特征与强特征相关性。**

- **朴素贝叶斯**：基于特征条件独立假设
- **SVM**：更适合特征维度低、样本量小、数据分布相对均匀
- **神经网络**：需要**大规模数据、强非线性关系建模**，且特征需经过精细预处理
- **k近邻**：适合样本量小、特征维度低、局部模式明显的场景

- 因此，本项目**重点考虑**逻辑回归、随机森林、XGBoost，进行深入比较和调参。

分别汇报：**a)混淆矩阵，b)预测准确度/精度，**
c)Precision（查准率）, d)Recall（查全率）。
对3 个模型中的至少一个模型，汇报：ROC 曲线。

统一建模流程：

任务2：是否欠费

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

有效样本数量： 8845

---- LogisticRegression ----
混淆矩阵：
[[1060 524]
[70 115]]

---- RandomForest ----
混淆矩阵：
[[1472 112]
[122 63]]

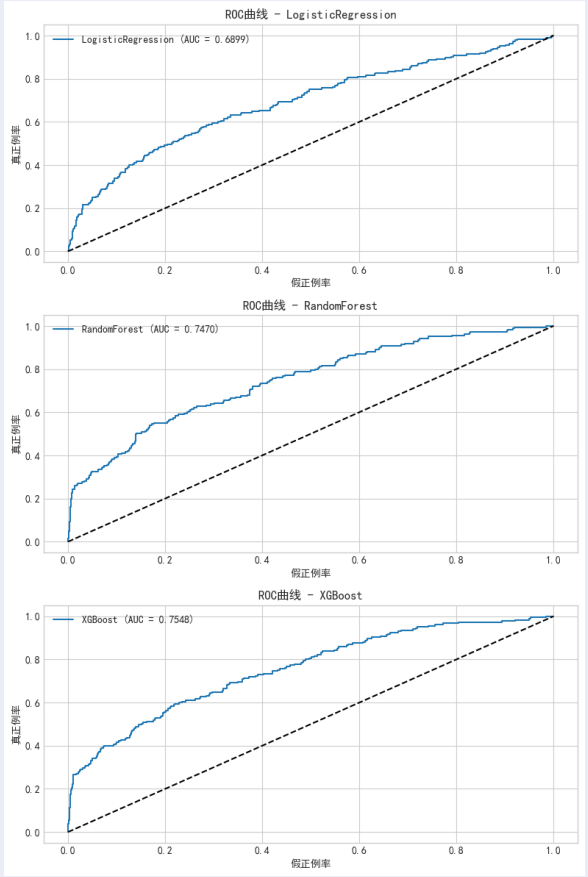
---- XGBoost ----
混淆矩阵：
[[1382 202]
[102 83]]

===== 三模型对比（任务2）=====

	Accuracy	Precision	Recall	F1	ROC-AUC
XGBoost	0.828151	0.291228	0.448649	0.353191	0.754784
RandomForest	0.867722	0.360000	0.340541	0.350000	0.747048
LogisticRegression	0.664217	0.179969	0.621622	0.279126	0.689882

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

最佳模型：XGBoost
核心原因：
任务 2 是严重不平衡的二分类问题（欠费样本占比低），此时ROC-AUC（反映整体分类能力）和 F1 分数（平衡精确率和召回率）是更重要的指标。
XGBoost 的 ROC-AUC（0.7548）和 F1 分数（0.3532）均优于其他模型，尤其在召回率（0.4486）上明显高于随机森林（0.3405），说明其能更好地识别出“欠费”这一少数类（对不平衡数据更敏感）。
随机森林虽然准确率更高（0.8677），但主要是通过“偏向多数类（不欠费）”实现的，对少数类的识别能力较弱，不适合不平衡场景。



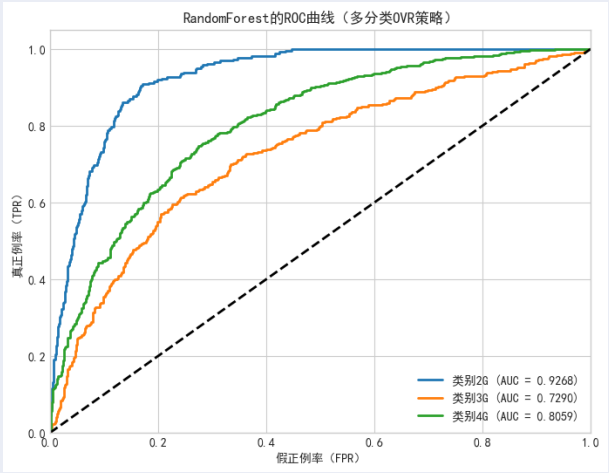
任务3：终端类型=2G、3G、4G

有效样本数： 8472

任务3：模型对比

	Accuracy	Macro-F1
RandomForest	0.694985	0.631291
XGBoost	0.722124	0.593926
MultinomialLogit	0.686726	0.446029

模型	混淆矩阵（实际-预测）	准确度	查准率（类别 0/1/2）	查全率（类别 0/1/2）
MultinomialLogit	[[122,0,139],[42,0,313],[37,0,1042]]	0.6867	0.61/0.00/0.70	0.47/0.00/0.97
RandomForest	[[218,7,36],[72,150,133],[110,159,810]]	0.6950	0.55/0.47/0.83	0.84/0.42/0.75
XGBoost	[[171,10,80],[39,74,242],[50,50,979]]	0.7221	0.66/0.55/0.75	0.66/0.21/0.91



- **最佳模型**：随机森林
- **核心原因**：
- 多分类任务中，**宏 F1（Macro-F1）** 是关键指标（反映模型对**所有类别的平均表现**，不受样本不平衡影响）。
- 随机森林的宏 F1（0.6313）显著高于 XGBoost（0.5939）和逻辑回归（0.4460），尤其在“3G”这一中间类别上表现更均衡（召回率 0.42，F1=0.45），而 XGBoost 对“3G”的识别能力极差（召回率仅 0.21）。
- 随机森林的集成特性使其在处理多类别、复杂特征交互时更稳健，不易受某一占比高的类别（如 4G，样本量最大）主导。

任务4和5中的数据泄露问题

```python
# 操作系统类型清洗
def simplify_os(x):
    if isinstance(x, str):
        xu = x.upper()
        if "ANDROID" in xu:
            return "ANDROID"
        if "IOS" in xu:
            return "IOS"
    return None

df5 = df_public.copy()
df5["最近使用操作系统偏好"] = df5["最近使用操作系统偏好"].apply(simplify_os)
df5 = df5[df5["最近使用操作系统偏好"].isin(["ANDROID", "IOS"])].copy()

# 1. 统计终端品牌和最近使用操作系统偏好的组合数据
brand_os_counts = df5.groupby(['终端品牌', '最近使用操作系统偏好']).size().sort_values(ascending=False)

# 2. 统计入网渠道类型和渠道类型描述的组合数据
channel_desc_counts = df5.groupby(['入网渠道类型', '渠道类型描述']).size().sort_values(ascending=False)
```

```
终端品牌和最近使用操作系统偏好组合统计：
终端品牌      最近使用操作系统偏好
苹果          IOS              3289
小米          ANDROID           866
三星          ANDROID           851
华为          ANDROID           803
维沃          ANDROID           431
                               ...
锤子（作废）    ANDROID             1
飞利浦        ANDROID             1
骋娱传媒      ANDROID             1
鸿锦          ANDROID             1
黑莓          ANDROID             1
Length: 135, dtype: int64

入网渠道类型和渠道类型描述组合统计：
入网渠道类型   渠道类型描述
2.0        社会渠道        4210
1.0        自有渠道        3580
99.0       其他渠道          45
dtype: int64
```

“终端品牌”和“最近使用操作系统偏好”实际的信息是相同的。
“入网渠道类型”和“渠道类型描述”实际的信息是相同的。
两者不能同时出现。

任务4：渠道类型描述=社会渠道和自有渠道

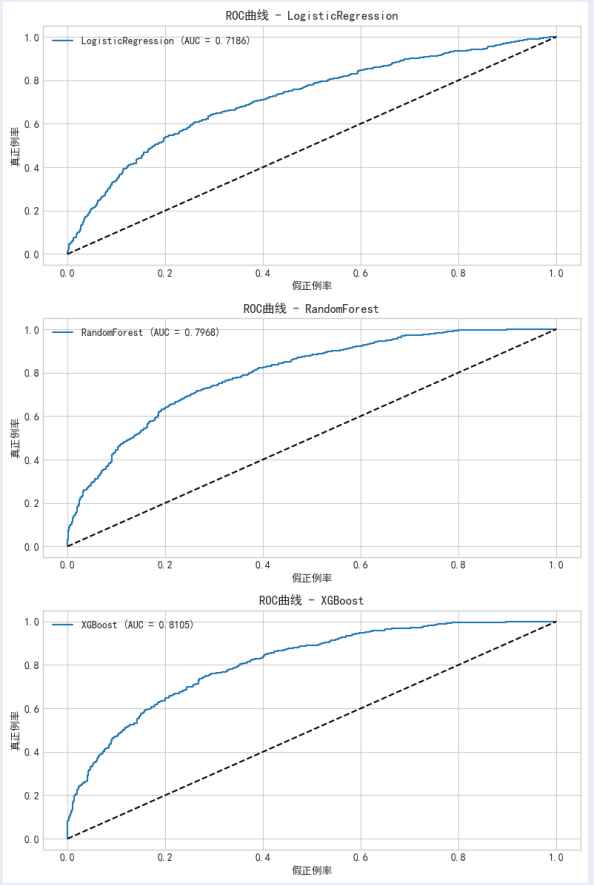
有效样本数： 8796

```
drop_cols = ["用户ID", "客户类型", "入网渠道类型"]
X, y, _ = preprocess_features(df4, target_col="渠道类型描述", drop_cols=drop_cols)
```

删除“入网渠道类型”字段以解决数据泄露

```
===== 三模型对比（任务4）=====
                    Accuracy  Precision    Recall        F1   ROC-AUC
XGBoost             0.722727   0.774306  0.695426  0.732749  0.810502
RandomForest        0.723864   0.741870  0.758836  0.750257  0.796843
LogisticRegression  0.656250   0.663014  0.754678  0.705882  0.718604
```

- **最佳模型**：XGBoost
- **核心原因**：
- 该任务为相对平衡的二分类问题，**ROC-AUC（0.8105）** 最高，说明其对“社会渠道”和“自有渠道”的区分能力最强。
- 虽然随机森林的 F1 分数（0.7503）略高于 XGBoost（0.7327），但 XGBoost 的精确率（0.7743）更高。



任务5：最近使用操作系统偏好

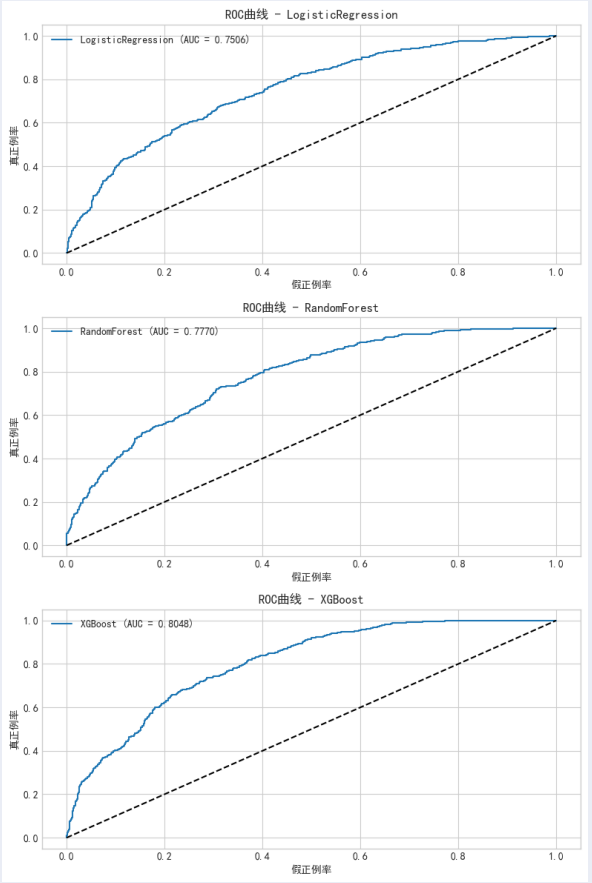
有效样本数： 7836

```
drop_cols = ["用户ID", "客户类型", "终端品牌"]
X, y, _ = preprocess_features(df5, target_col="最近使用操作系统偏好", drop_cols=drop_cols)
```

删除“终端品牌”字段以解决数据泄露

```
===== 三模型对比（任务5）=====
                    Accuracy  Precision    Recall        F1   ROC-AUC
XGBoost             0.727679   0.680685  0.663126  0.671791  0.804778
RandomForest        0.703444   0.627297  0.725341  0.672766  0.776970
LogisticRegression  0.688776   0.684665  0.481032  0.565062  0.750607
```

- **最佳模型**：XGBoost
- **核心原因**：
- XGBoost 在**准确率（0.7277）**和**ROC-AUC（0.8048）**上均为最高，综合表现最优。
- 其精确率（0.6807）和召回率（0.6631）更均衡，说明对“ANDROID”和“IOS”的识别能力更稳定。



05

模型整体对比与总结

总结

任务	模型	最佳指标	原因
任务 2：预测是否欠费（严重不平衡二分类）	XGBoost	ROC-AUC=0.7548，F1=0.3532	**该任务核心需求是识别少数“欠费”样本**，XGBoost 的 ROC-AUC（反映整体区分能力）和 F1 分数（平衡精确率与召回率）均最优。其召回率（0.4486）高于随机森林，能更好捕捉欠费样本，且通过梯度提升策略缓解了数据不平衡带来的偏置，表现优于偏向多数类的随机森林和线性拟合的逻辑回归。
任务 3：预测终端类型（2G/3G/4G，多分类）	RandomForest	Macro-F1=0.6313	多分类任务需关注对所有类别的均衡表现，RandomForest 的宏 F1 显著高于其他模型，对样本量较少的“3G”类别（召回率 0.42）识别能力更优，未出现完全无法预测的类别。其集成学习特性能有效处理多类别间的复杂关系，避免被占比高的“4G”类别主导，均衡性优于 XGBoost 和逻辑回归。
任务 4：预测渠道类型（二分类）	XGBoost	ROC-AUC=0.8105，精确率 = 0.7743	任务为相对平衡的二分类，XGBoost 的 ROC-AUC 最高，说明类别区分能力最强；精确率（0.7743）优于其他模型，意味着预测“社会渠道”的样本中实际准确率更高，误判更少。其**梯度提升机制**能精准捕捉渠道类型与用户消费、流量特征的非线性关联，拟合能力强于线性模型和随机森林。
任务 5：预测操作系统偏好（二分类）	XGBoost	ROC-AUC=0.8048，准确率 = 0.7277	XGBoost 综合表现最优，准确率和 ROC-AUC 均为最高，且精确率（0.6807）与召回率（0.6631）更均衡，无明显偏置。操作系统偏好与用户行为特征（如流量使用、收入）的关系呈非线性，**XGBoost 通过树结构迭代优化，能高效学习复杂模式，优于逻辑回归的线性拟合和随机森林的并行集成效果。**

感谢观看！

THANKS FOR WATCHING

汇报人：钟忠权

时间：2025.11.17