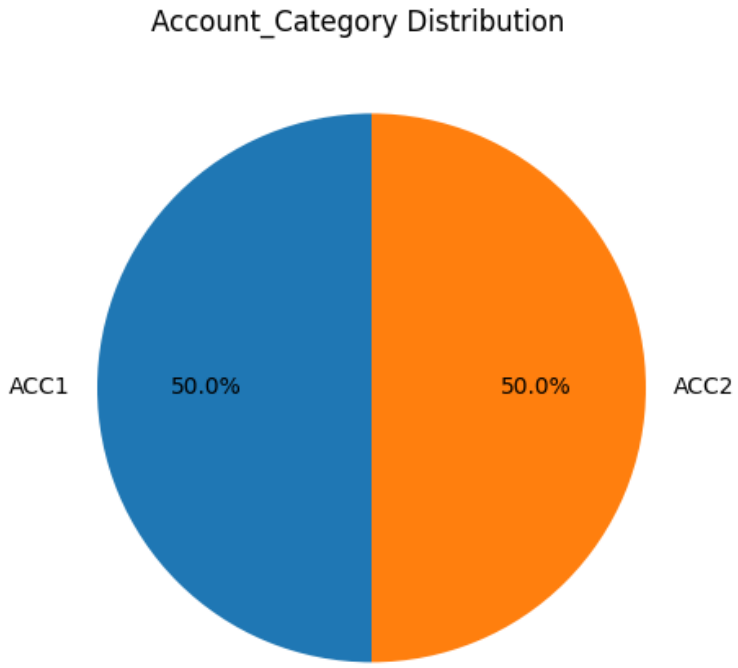
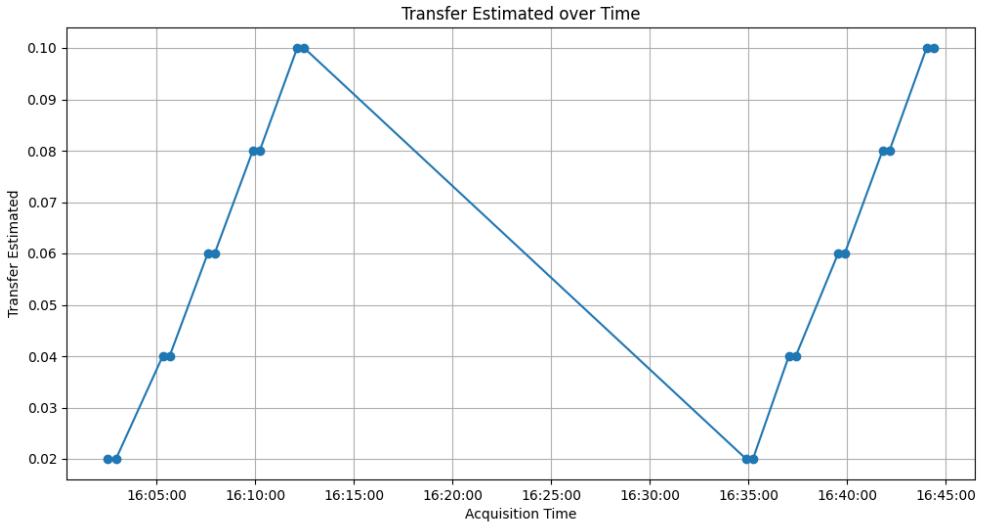
项目总结

关键信息分布可视化：

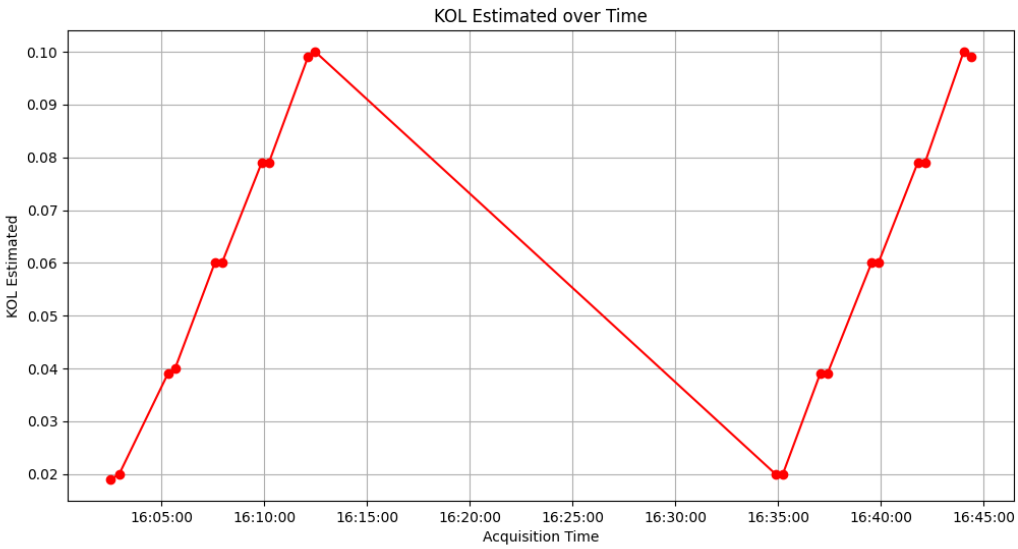
1. Account_Category 分布饼图：



2. Transfer_Estimated 时序图:



3. KOL_Estimated 时序图:



数据处理：

1. 缺失值处理：

删去缺失值列 Note

2. 数据整理：

2.1 新建两列分别用列表储存 x, Account_Corrected 列的数据

2.2 对账户类型数据进行编码

3. 确定特征向量与目标向量：

3.1 特征向量:

每个向量由 46 个特征组成：x(22)+Account_Corrected(22)+编码化账户类型(2)

3.2 目标向量:

每个向量由 2 个特征组成：KOL_Estimated(1)+Transfer_Estimated(1)

4. 数据集划分：

训练集(16) + 测试集(4)

模型建立：

1. 支持多输出的模型：输出带有 KOL_Estimated 和 Transfer_Estimated 值的 np,array
2. 不支持多输出的模型：建立两个分别训练，输出 KOL_Estimated 和 Transfer_Estimated 的值

模型评估：

方法一：

1. 根据测试集的特征向量获取模型对应的预测结果
2. 利用 mean_squared_error()函数对比测试集的目标变量与模型预测结果，得到模型的**均方误差（MSE）**。

方法二： 评估模型在交叉验证中的表现

1. 利用 cross_validate()函数执行交叉验证得到负均方误差，再取负得到均方误差（MSE）
2. 计算并输出模型在交叉验证中的**平均均方误差（RMSE）**。

方法三：模型评分

1. 利用 score()函数得到模型在测试集上的**表现评分**。

模型超参数调整：

方法一：GridSearchCV

1. 通过 get_params()函数获取到模型的所有超参数，再结合模型相关文档了解这些参数的设置原理，作用与合适的范围。

2. 利用 GridSearchCV，初始化以原模型为基础的 GridSearchCV，设定 param_grid，合理范围内进行参数的取值尝试。
3. 训练 GridSearchCV，利用 best_estimator_函数获取最佳模型，再对最佳模型利用 get_params()函数了解超参数最优取值的倾向。同时分别对原模型与所获得的最佳模型进行模型评估了解优化效果。根据最优取值的倾向和优化效果再次调节 param_grid 中超参数的取值。
4. 重复步骤三直到模型评估效果达到最好，同时最优模型超参数的取值已不再根据 param_grid 的调整而变化。

方法二：RandomizedSearchCV 与 GridSearchCV 结合

1. 对需要调整的全部超参数都分别设定一个合理的取值范围。
2. 利用 RandomizedSearchCV，将设定的超参数搜索范围作为参数代入初始化 RandomizedSearchCV。
3. 训练该 RandomizedSearchCV 进行超参数随机匹配择优，通过 best_params_得到目前的最佳超参数组合。
4. 在上面步骤得到的超参数组合的基础上对其中的每个超参数进行进一步的微调，做一些取值尝试构成新的取值范围字典。
5. 利用新的取值范围字典初始化一个新的 GridSearchCV 并训练进行超参数遍历匹配择优。
6. 参照方法一中的 3，4 步骤确定最优参数。

方法三：Pipeline 与 GridSearchCV 结合

1. 创建 Pipeline，利用其中的 SelectKBest()函数来选择最好的特征，或是 StandardScaler()函数来标准化特征。
2. 定义超参数范围，设定 param_grid，带入初始化和训练 GridSearchCV。
3. 利用 best_params_函数和 best_estimator_函数得到最佳参数和模型。
4. 参照方法一中的 3，4 步骤确定最优参数。

单个模型训练与调整过程：

1. 建立原始模型，使用默认参数

2. 利用训练集进行训练
3. 对模型进行评估，保留各项评估数据
4. 对模型进行超参数调整，找到最优参数
5. 对最优参数模型进行评估，保留各项评估数据
6. 将调参前与调参后的各项评估数据进行对比

各项评估数据对比结果：

多模型 1（Models1.ipynb）

	调参前平均均方误差	调参后平均均方误差	调参前均方误差 KOL_Estimated	调参后均方误差 KOL_Estimated	调参前均方误差 Transfer_Estimated	调参后均方误差 Transfer_Estimated
线性回归	7.52340369e-06	8.16179358e-06	4.23516259e-06	3.64170088e-06	3.12221869e-06	2.15056134e-06
决策树回归_KOL_Estimated	5.70550000e-04	2.00000000e-07	2.10500000e-04	5.00000000e-07	nan	nan
决策树回归_Transfer_Estimated	5.20000000e-04	0.00000000e+00	nan	nan	2.00000000e-04	0.00000000e+00
岭回归	7.52340432e-06	8.16171113e-06	4.23515493e-06	3.63587744e-06	3.12220989e-06	2.14381477e-06
拉索回归	4.15462703e-06	3.96182674e-06	6.68106003e-06	4.74810700e-06	4.31030920e-06	2.93595534e-06
SVM回归_KOL_Estimated	8.00300000e-04	1.20653512e-04	9.05250000e-04	1.82073944e-04	nan	nan
SVM回归_Transfer_Estimated	8.00000000e-04	1.21861295e-04	nan	nan	9.00000000e-04	1.75530936e-04

	训练前评分	训练后评分
线性回归	9.94574300e-01	9.95730254e-01
决策树回归_KOL_Estimated	6.90554943e-01	9.99264976e-01
决策树回归_Transfer_Estimated	7.03703704e-01	1.00000000e+00
岭回归	9.94574312e-01	9.95739532e-01
拉索回归	9.91896440e-01	9.94335246e-01
SVM回归_KOL_Estimated	-3.30760750e-01	7.32342603e-01
SVM回归_Transfer_Estimated	-3.33333333e-01	7.39954168e-01

结合起来看，决策树回归模型是最适合的模型。

多模型 2（Models2.ipynb）

	调参前平均均方误差	调参后平均均方误差	调参前均方误差 KOL_Estimated	调参后均方误差 KOL_Estimated	调参前均方误差 Transfer_Estimated	调参后均方误差 Transfer_Estimated
KNN 回归	2.01082000e-04	1.87014675e-04	3.98260000e-04	2.70907502e-04	3.96000000e-04	2.65393904e-04
随机森林回归	1.69363250e-04	1.60899686e-04	1.29681850e-04	1.28215397e-04	1.24310000e-04	1.23266133e-04
Adaboost回归_KOL_Estimated	2.34550000e-04	2.74600000e-04	2.10500000e-04	2.10500000e-04	nan	nan
Adaboost回归_Transfer_Estimated	5.60000000e-04	4.60000000e-04	nan	nan	2.00000000e-04	2.00000000e-04
GBRT回归_KOL_Estimated	1.47088031e-04	1.34997491e-04	9.30017721e-05	8.05463085e-05	nan	nan
GBRT回归_Transfer_Estimated	1.06809344e-04	1.25917124e-04	nan	nan	1.20159311e-04	8.66514027e-05
Bagging回归	2.37263250e-04	1.85892568e-04	1.21755000e-04	1.04650850e-04	1.15000000e-04	9.80000000e-05
ExtraTree回归	1.04306660e-04	9.52786500e-05	8.54437250e-05	8.21942000e-05	8.00800000e-05	7.71600000e-05

	训练前评分	训练后评分
KNN 回归	4.13936053e-01	6.04288438e-01
随机森林回归	8.12599261e-01	8.14217035e-01
Adaboost回归_KOL_Estimated	6.90554943e-01	6.90554943e-01
Adaboost回归_Transfer_Estimated	7.03703704e-01	7.03703704e-01
GBRT回归_KOL_Estimated	8.63282952e-01	8.81593078e-01
GBRT回归_Transfer_Estimated	8.21986206e-01	8.71627552e-01
Bagging回归	8.25321981e-01	8.54968679e-01
ExtraTree回归	8.77878303e-01	9.04341757e-01

结合起来看，ExtraTree 回归模型是最适合的模型。

重要信息对比可视化：（以线性回归模型为例）

散点图 + 理想线

x 轴为测试集实际值，y 轴为模型测试值

图例表示：

绿色圆点：KOL_Estimated 的实际值与模型测试值的关系

蓝色×点：Transfer_Estimated 的实际值与模型测试值的关系

红色实线：KOL_Estimated 的实际值与模型测试值相一致的理想线

蓝色虚线：Transfer_Estimated 的实际值与模型测试值相一致的理想线