9.26 第三次会议记录

* 答辩技巧
* “田忌赛马”
* 准备答辩材料时要留气口，不要太全面

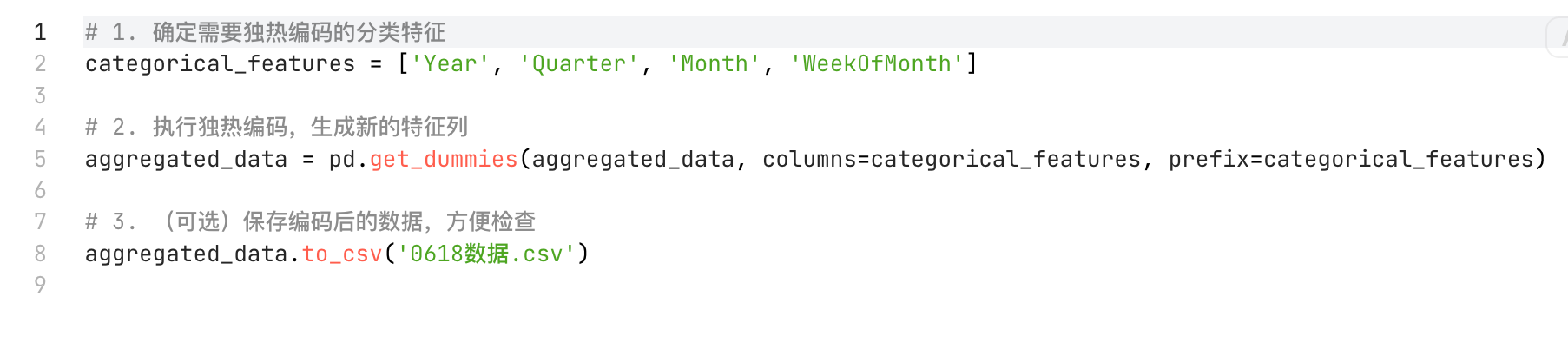
1.LSTM模型代码解读

**（1）【独热编码】**

* 将分类特征转换为数值特征

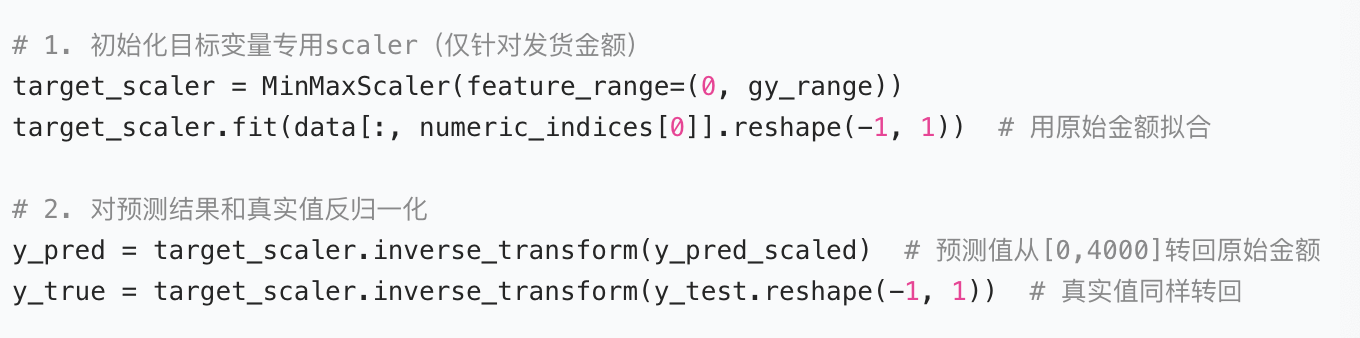


* 原始时间特征的数值是“类别”而非“大小”；如果直接输入模型，则会让其误以为时间数值存在大小关系，学习错误的规律。
* LSTM等神经网络只能处理数值型输入，无法直接识别“类别标签”。独热编码会将分类特征转换为0、1的数值矩阵，让模型能正确捕捉“不同类别对发货金额的影响”。
* 具体作用：①精准区分不同时间类别的差异；②学习到“类别组合”的影响（年末&季度末&月末组合时，权重叠加，预测更高金额）
* 注意：独热编码（0/1值）不需要归一化



**（2）【scaler】**

* 核心作用：消除特征量级差异，让所有输入特征对模型的影响处于“同一量级水平”，同时保留特征内部的相对大小关系
* 示例代码中使用MinMaxScaler，针对“发货金额”进行缩放
* 「feature\_range=(0, gy\_range)（gy\_range=4000）」，表示将数据映射到[0, 4000]区间。gy\_range默认为1
* 【反归一化】：缩放只是为了方便模型训练，最终的预测结果需要转换回原始发货金额的量级才有实际意义。反归一化是缩放的逆运算。





**（3）softplus函数**

* 输出层的激活函数，用于将模型的原始输出转换为符合业务场景的预测值
* 具体作用：输出的值始终为非负数；输出范围可随x无限增大，覆盖不同量级的金额；函数平滑可导，确保模型训练时梯度稳定传递。

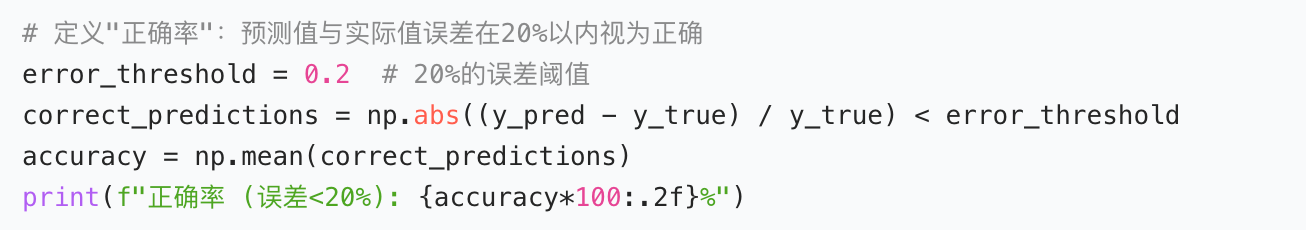


**（4）MSE：均方误差**

* 先计算每个样本的预测误差，再对误差进行平方，最后求所有样本的平均值
* 评估模型预测精度：MSE值越小，预测精度越高
* 间接辅助模型训练：虽然使用了 Huber 损失函数（对异常值更稳健），但 MSE 的“平方放大误差”特性对模型训练有参考意义：
  + 训练过程中，若 MSE 持续下降，说明模型正在学习减少预测误差（尤其是对大额误差的优化）；
  + 若 MSE 在训练后期波动或上升，可能提示模型过拟合（对训练数据过度拟合，泛化能力差）。

**（5）设定阈值**

* 使用“误差阈值”计算“预测准确率”



（6）每一次测试都要留记录

（7）假设一些数据具有时间序列，构建基础模型。把模型工具训练好后再进行测试。

2.OCR模型优化

（1）标注特征值【只需要特征，暂时不需要数值】

* 把突出的特征值提取出来并定义
* 比较容易提取的数据，精确度提高；提高识别精确度的平均值

（2）标注票据中不同区域坐标，按布局进行识别

3.读文献

（1）袁老师推荐文献：先找需要的部分，单独摘出来

（2）整理后会议分享

4.建立github代码仓，添加会议纪要&代码（日期，谁写的，什么内容）

5.尝试copilot