数据分析集合

这里记录了我的一些数据分析的项目，主要展示了对于澳洲 Sprocket Central Pty Ltd 的客户价值预测和一些其他的成果。

基于机器学习技术预测新客户价值

Abstract

本次项目基于 SCPL 公司的 4000 名老客户的 20000 条消费记录，训练了决策树回归模型，模型输入新客户的性别，职业，地址等特征信息，模型输出该用户的消费价值预测。模型实验结果对于训练集的数据相关性达到了 0.8 左右，验证集达到了 0.7 左右。

Method

首先根据消费记录，计算出每个客户的消费价值，之后去除掉无效字段，空字段，接着判断各个特征(比如性别)和价值的相关性，筛选出相关性较高的字段，之后将行业等字符性进行离散处理，接着通过 Sklearn 的预处理库对于数据进行预处理，之后在决策树模型上进行训练。多次训练测试，保存训练集效果最好的模型，对新客户表的数据进行同样处理后，获取预测结果。

Results and discussions

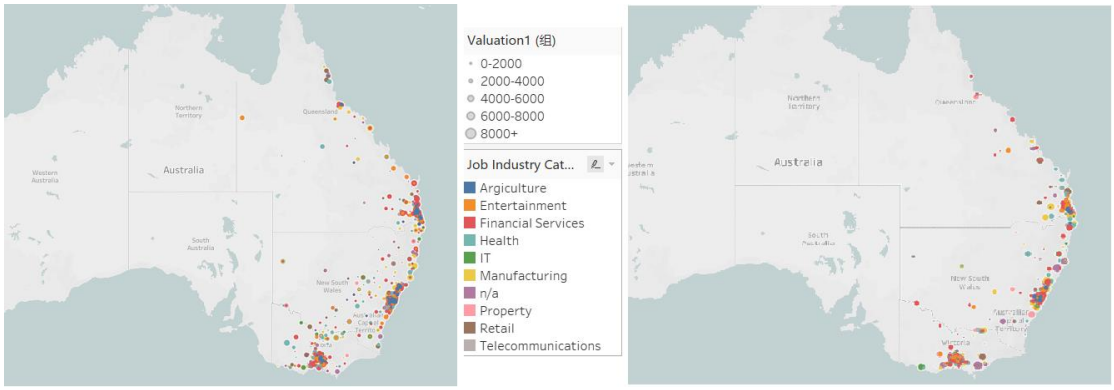

Figure 1: 左图为 4000 名老顾客分布，右图为 1000 名新顾客预测分布.

如图 1 所示，原始数据的老顾客分布情况和预测数据的新顾客预测情况大体一致，除此之外，各个行业的价值分布，以及用户的消费占比，新顾客的情况和老用户也几乎一致，因此本次决策树回归模型对于新用户的价值预测拥有一定的准确度。然而由于没有后续数据，因此很难对模型进行评估，实际应用中，模型在上线前，仍可

以对于新数据进行一段时间的测试迭代，最终获得一个成熟的新顾客价值预测模型。

推文真伪识别

Abstract

本次研究提出了 5 种机器学习技术基于中世纪的上万条推文进行建模训练，首先分析了数据集的数据质量，数据偏向等数据集基本特征，之后对于数据进行了去除常用词，去除停用词等预处理手段，之后采用词频提取等特征提取方法提取文本数据特征，之后在五种**机器学习模型中进行测试记录。**

Method

首先对于推文数据进行预处理，包括对于不同类型语言进行机器翻译，网址等冗余信息进行数据清理，之后采用 TF-IDF 等特征提取技术对文本数据进行特征提取，随后在多层感知机，文本 CNN，SVM，KNN 和朴素贝叶斯五种模型上进行测试，通过测试集的表现以及识别速度，成本等对五种模型进行排序。

Results and discussions

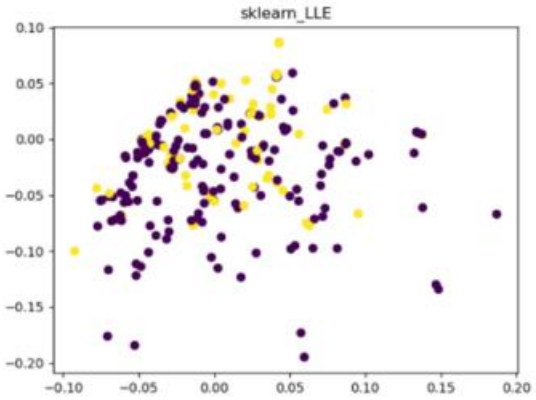

Figure2:LLE 降维

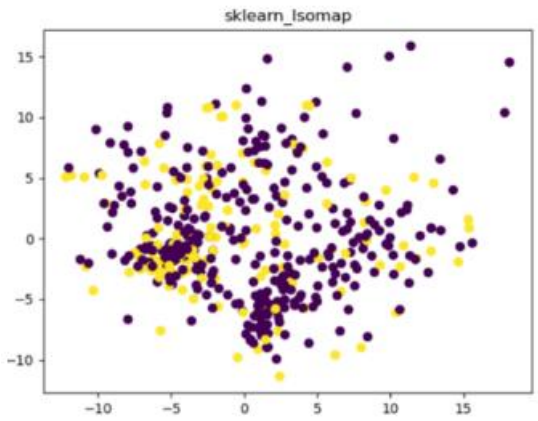

Figure3:LSOmap 降维

五种机器学习模型中，识别准确率最高的是文本 CNN，然后文本 CNN 的训练成本最高，KNN 的识别准确率最差，而且计算量很大，很次 KNN 的整体效果最差，SVM 的识别准确率略好于 KNN，同时训练过程也比较简单，五种模型中整体效果最好的是朴素贝叶斯，该模型被广泛认为适合文本分类任务，实际表现中不断训练速度快，而且准确率高。图 2，3 显示了两种特征降维技术降维后的数据情况，考虑到较多的特征数据会增加模型训练成本，因此合适的特征降维技术也应该应用到本次应用中来。