成绩	

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

数据挖掘

项目结课报告

题　　目： 利用上市公司财务报表决策投资

学　　院： 计算机学院

专业名称： 1 网络空间安全、2 软件工程

姓　　名： 李斌斌(3120181094)[1]

王文杰(3120181095)[1]

赵鹏飞(3220180904)[2]

罗妹秋(3220180726)[2]

任课教师： 汤世平

项目地址： https://github.com/mijeff/datamining_homework/tree/master/team_project

摘要：

自 2015 年以来，炒股走进人们的视野并大火，但绝大部分股民不具备相关专业知识背景，投资决策盲目跟风，亏损连连，因此如何通过分析各大企业公开发布的财务信息，帮助股民合理运用资金，做出正确投资决策具有重大意义。本项目对东方财经网 2012 年至 2018 年的年度业绩快报进行分析，通过净资产收益率来评估企业是否值得投资。不同于传统财务报表的分析方法所采用的比率分析、比较分析法、趋势分析法等的单一，本项目以往年的营业收入同比增长、净利润同比增长、净资产收益率为特征，采用 Logistic 回归模型和决策树模型来预测当年的净资产收益率大于 15%的可能性。本项目将 2012 年至 2017 年的财务信息合并为训练数据集，将 2013 年至 2018 年的财务信息合并为测试数据集，实验结果显示，使用 Logistc 回归模型的准确率为 90.82%，使用决策树模型的准确率为 71.44%。

关键词：投资决策，上市公司财务报表，Logistc 回归模型，决策树模型

项目分工

姓名	学号	任务
李斌斌	3120181094	Logistic 回归模型，撰写开题、结题报告
王文杰	3120181095	决策树算法模型，撰写中期、结题报告
赵鹏飞	3220180904	数据获取和清洗，撰写中期、结题报告
罗妹秋	3220180726	分析预测结果，撰写结题报告

一、问题描述

1.1 问题背景分析

互联网的迅猛发展也带动着电子商务的广泛普及，产生了海量数据。截止 2019 年 4 月 2 日，我国上海、深圳两大证券交易所的上市公司数量在 3621 家。这些公司每年发布了大量财务信息和数据，而在市场经济时代，公司财务信息不再是枯燥乏味的数据，而是可为投资行为提供决策的有用数据。无论投资者对于投资风险持何种态度，对于能有助于评估投资收益和风险的信息都是十分渴望的。

如何对大量财务信息进行整理归纳，抽取出可运用和分析的数据，进而挖掘其中隐含的有价值的信息，使之变为可以利用的资源，意义非凡。本项目希望能

根据从已有的财务信息中找出对投资行为有用的特征，并应用于实际的投资决策中。

1.2 问题描述

证券行业作为一个需要高度信息化的行业，每年产生的数据量都在急剧的增大，上市公司每个季度都会发布相应的财务报告，向所有者、债权人、政府及其他有关各方及社会公众等外部使用者披露企业当前的生产经营结果和财务活动的状况。其中财务报表作为财务报告的主要组成部分，它所提供的企业财务信息具有极其重要的价值。然而我们得到的财务报告往往都是不可编辑的电子格式，当我们想要对相同公司不同时期的财务项目和不同公司间的相同财务项目进行比较时，往往只能手工翻阅查看。并且财务报表所反映的财务内容往往是高度概括和抽象的，一项财务内容往往被分开为不同的报表项目来进行展现。这不利于相关的财务报表使用者清晰的了解到数据后更深层次的意义。

二、实验数据

2.1 数据获取

上市公司的财务数据均来自于上市公司对外发布的财务报表等信息，我们希望根据这些信息预测下一年的企业净资产收益率。首先需要把这些财务报表中的财务数据抓取出来，具体过程为：从证券交易所抓取上市公司公布的财务报告文件，然后对其进行格式转换，变为可以用代码提取的结构化文件，自动提取其中的财务报表数据保存到数据库中。

本项目数据爬取自东方财经网 2012 年至 2018 年的年度业绩快报，共 12828 条记录，包含了股票代码、营业收入、净利润等 21 条财务信息，数据获取结果如图 2-1 所示。

```
Int64Index: 12750 entries, 0 to 1953
Data columns (total 21 columns):
scode                12750 non-null object
sname                12750 non-null object
securitytype         12750 non-null object
trademarket          12750 non-null object
ldate                12750 non-null object
rdate                12750 non-null object
basiceps             12750 non-null object
yysr                 12750 non-null object
ys                   12750 non-null object
yshz                 12750 non-null object
qntqys               12750 non-null object
jlr                  12750 non-null object
lr                   12750 non-null object
sjlhz                12750 non-null object
qntqjlr              12750 non-null object
parentbvps           12750 non-null object
roeweighted          12750 non-null object
publishname          12750 non-null object
securitytypecode     12750 non-null object
trademarketcode      12750 non-null object
firstnoticedate      12750 non-null object
```

图 2-1 数据获取结果

2.2 数据预处理

数据预处理部分分为训练数据预处理和测试数据预处理，两部分步骤相同，不同之处在于年份的划分。本项目将 2012 年至 2017 年的财务信息合并为训练数据集 train.csv，2013 年至 2018 年的财务信息合并为测试数据集 test.csv。

2.3 数据分析与可视化

本项目在分析财务数据时主要采用趋势分析法。通过比较企业连续几年财务报表中的财务项目，通过统计学方法，计算存储财务项目的发展趋势函数，来预测财务状况和经营成果的变化和发展趋势。具体是通过比较企业财务数据来分析企业的财务状况，可以分为按金额分析和按百分比分析两种方式。

根据上述爬虫获得的 2012 年至 2018 年的企业年报业绩快报数据，主要从企业盈利能力、盈利质量、偿债能力、营运能力和发展能力五个方面选择特征，找到特征之间的依赖关系。经过分析，我们选取每年的营业收入同比增长、净利润同比增长、净资产收益率作为决定企业净资产收益率的关键特征，将企业的净资产收益率区间作为分类结果。净资产收益率是反映上市公司盈利能力及经营管理水平的核心指标，指标值越高，说明投资带来的收益越高。所以本系统以净资产收益率为指标，来评估企业的投资价值，当企业的净资产收益率低于当年的存款利率时，说明企业是不值得投资的。净资产收益率公式如式（2-1）所示。

$$\text{净资产收益率} = \frac{\textit{税后利润}}{\textit{所有者权益}} \qquad \text{式(2-1)}$$

经过对数据集的分析和处理，我们可以得到 2012 年净资产收益率的直方图(如图 2-2)，从直方图可以看出 2012 年净资产收益率大部分属于[0,20]的区间，而(-∞，0)和(20，+∞)占比很小；另外本文还构建了 2012 年净资产收益率的 QQ 图(如图 2-3)、12 年净资产收益率的盒图(如图 2-4)。

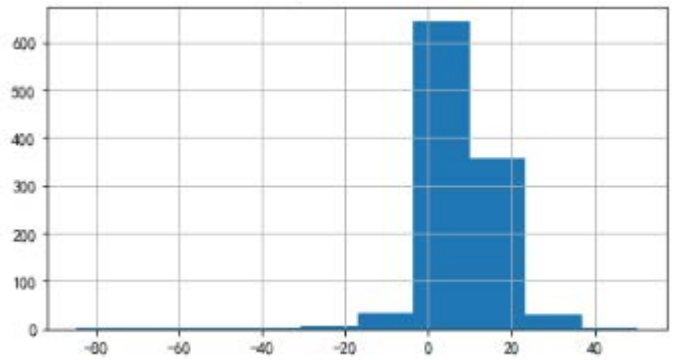

图 2-2　2012 年净资产收益率的直方图

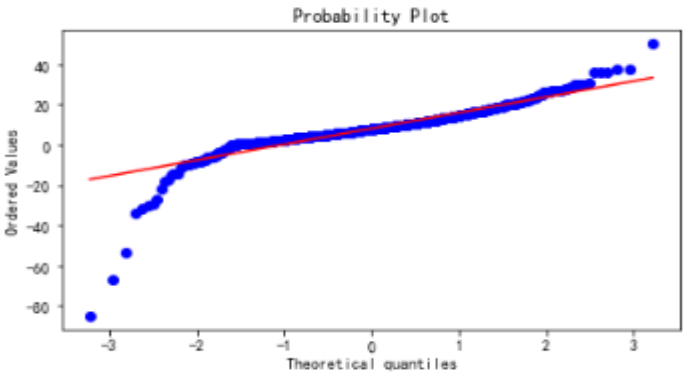

图 2-3 2012 年净资产收益率的 QQ 图

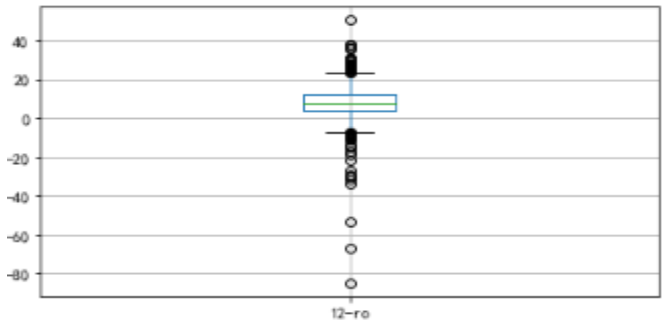

图 2-4 2012 年净资产收益率的盒图

三、模型构建

整个模型包括财务数据的提取模块、数据挖掘模块和数据展现模块。在数据挖掘模块主要从企业盈利能力、盈利质量、偿债能力、营运能力和发展能力五个方面选择变量。模型选择方面使用 Logistic 回归模型和决策树算法模型，进行对比实验。

3.1 Logistc 回归模型

Logistic 回归是一种特殊的回归模型，也可以用来进行分类预测，主要用于二分类分析。Logistic 回归模型是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法。其基本原理是利用一组数据拟合一个 Logistic 回归模型，然后借助这个模型揭示总体若干个自变量与一个因变量取某个值的概率之间的关系。

在本项目实现中使用了 sklearn.preprocessing 库中的 StandardScaler 函数来标准化待训练和待测试的数据，保证每个维度的特征数据方差为 1，均值为 0，使得预测结果不会被某些维度过大的特征值而主导；之后使用库 sklearn.linear_model.logistic 中的 LogisticRegression 函数来生成模型，并使用其分类器中的评分函数（score 函数）来计算预测结果的准确性；每次输入模型的数

据必须经过标准化，才能使用；相应代码如图 3-1 所示。

```
##模型训练
#标准化数据，保证每个维度的特征数据方差为1，均值为0。使得预测结果不会被某些维度过大的特征值而主导。
X_train_raw = trainData.iloc[:,1:-1]
y_train = trainData.iloc[:,-1]
X_train = ss.fit_transform(X_train_raw)
classifier=LogisticRegression(solver='liblinear')
#调用LogisticRegression中的fit函数/模块用来训练模型参数
classifier.fit(X_train,y_train)

##测试模型
X_test_raw = testData.iloc[:,1:-1]
y_test = testData.iloc[:,-1]
X_test = ss.transform(X_test_raw)
#使用训练好的模型classifier对X_test进行预测，结果储存在变量predictions
predictions=classifier.predict(X_test)
print ("Accuracy of LR Classifier:", classifier.score(X_test, y_test))
```

图 3-1 Logistc 回归模型

3.2 决策树模型

决策树可以用于数据的分析和分类，同样也可以用来做预测。决策树依据已有的数据生成决策树模型，用模型来预测分类未来的技术。决策树优点是分类精度高、生成的模型简单并且对噪声有很好的容错性等。决策树分类算法一般分为两个步骤，决策树的生成和决策树的修剪。决策树模型的实现代码如图 3-2 所示。

```
##模型训练
target = train_dataset.iloc[:,-1]
target = pd.DataFrame(target)
#class_names = ['verylow','low', 'middle', 'high', 'veryhigh']
class_names = ['low', 'middle', 'high']
clf = tree.DecisionTreeClassifier()
clf = clf.fit(feature, target)


##测试模型
test_feature = test_dataset.iloc[:,1:16]
clf.predict(test_feature)
test_target = test_dataset.iloc[:,-1]
clf.score(test_feature,test_target)
```

图 3-2 决策树模型

四、系统评估

4.1 数据获取与清洗评估

本项目中获取数据条数为 12750 条，清洗后数据条数为 12085 条，在东方财经网的业绩快报中，空值用横线代替，因此本项目数据不包含空值，只包含异常值，含有异常值的数据条数为 665 条。本项目获取了近 13000 条数据，含有 20 余条财务属性，处理后的数据有效率为 94.8%，这是一个较高的数字，说明东方财经网提供的数据全面而规范，对于模型构建、净资产收益率预测、股民或投资者投资是非常参考价值的。

4.2 模型评估

本项目中对模型评估参数是预测的准确度（accuracy）。对于给定的测试数据集，准确度即分类器正确分类的样本数与总样本数之比，公式如式（4-1）所示。

$$准确率 = \frac{分类器正确分类的的样本数}{总样本数} \qquad 式(4\text{-}1)$$

本项目中 Logistc 回归模型对于净资产收益率大于 15%的预测准确率为 90.82%，决策树模型实验结果的准确率为 71.44%。

在 Logistc 回归模型中，本项目为了更加清晰地展示预测的准确性，使用了混淆矩阵，如图 4-1 所示，在该混淆矩阵中 TP 与 TN 的数量大，同时 FP 与 FN 的数量小，意味着模型预测更准确。所以当我们得到了模型的混淆矩阵后，就需要去看有多少观测值在第二、四象限对应的位置，这里的数值越多越好；反之，在第一、三四象限对应位置出现的观测值肯定是越少越好。本次验证中得到的混淆矩阵为$\begin{bmatrix}1022 & 19\\ 88 & 37\end{bmatrix}$，同时回执了混淆矩阵图形，如图 4-2 所示。

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP （Type II）
	Negative	FN （Type I）	TN

图 4-1 混淆矩阵表

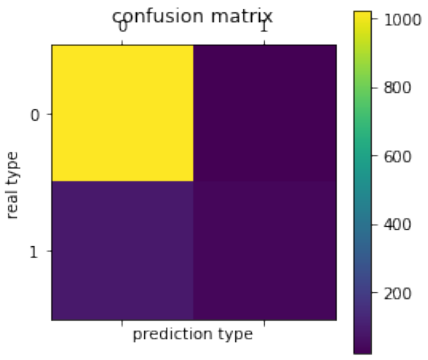

图 4-2 模型验证得到的混淆矩阵

在决策树模型中，为了便于进行分类预测，本项目将净资产收益率预先进行如下分类:

分类条件	分类结果
净资产收益率 ＜0	low
净资产收益率 ≥0 & 净资产收益率 ≤20	middle
净资产收益率 ＞20	high

使用决策树模型得到了如图 4-3 所示的决策树。

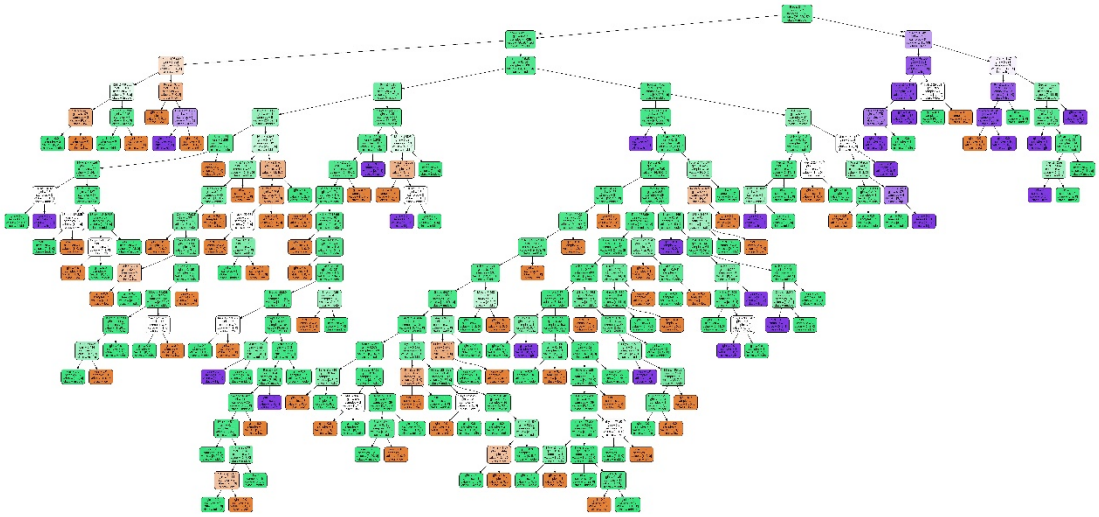

图 4-3 决策树

五、总结

本项目对东方财经网 2012 年至 2018 年的年度业绩快报进行数据分析，通过净资产收益率来评估企业是否值得投资，使用 Logistic 回归模型和决策树算法模型进行对比试验。实验结果表明，使用 Logistc 回归模型预测准确率要优于决策树模型。

实验过程中发现决策树算法模型的准确率不高的原因是获取到的数据的数量和质量较低。后期实验中分别在数据获取阶段爬取了更多的报表和在数据预处理阶段去除不规范文本、处理缺省值，实验结果表明准确率有了一定的提高。

本项目利用上市公司财务报表来决策投资，主要使用了往年的营业收入同比增长、净利润同比增长、净资产收益率三个特征来预测接下来某一季度或年份的净资产收益率，有助于投资者或股民的高效投资。本项目中团队成员一起合作，将数据挖掘课程中学习的理论知识应用到实际项目中，加深了知识点的理解，熟练了数据获取、数据预处理、模型构建、结果分析、数据可视化等技术。