# 预测银行电话营销成功的数据挖掘方法

### 摘要:在本文中我们使用数据挖掘的方法来预测电话销售银行长期存款的成功率。数据集是来源于一家葡萄牙零售银行2008年至2013年间的数据（共有41188条数据），因此包括了最近金融危机的影响。本文使用数据挖掘中的方法来对具有20个属性的数据集进行预测，预测银行客户是否会订阅定期存款。我们首先对数据集进行特征分析，对数据集中的缺失值进行处理，分析特征与结果之间的相关性，将一些不重要的特征删除，筛选出更合适的特征；然后使用逻辑回归，k近邻，朴素贝叶斯，决策树，SVM和集成模型等六种模型来进行预测。使用Accuracy来评估六个模型，然后发现LGBMClassifier呈现出最好结果（Accuracy=0.91676）。

### **关键词：**数据挖掘；特征工程；机器学习

### **摘要:** In this paper, we use data mining method to predict the success rate of telemarketing bank long-term deposits. The data set is from the data of a Portuguese retail bank from 2008 to 2013 (41188 data in total), so it includes the impact of the recent financial crisis. This paper uses the methods in data mining to predict the data set with 20 attributes and predict whether the bank customers will subscribe to fixed deposits. We first analyze the features of the dataset, process the missing values in the dataset, analyze the correlation between the features and the results, delete some unimportant features, and screen out more appropriate features; Then six models are used to forecast, including logical regression, k-nearest neighbor, naive Bayes, decision tree, SVM and integration model. Use Accuracy to evaluate the six models, and then find that LGBMClassifier presents the best results (Accuracy=0.91676).

### **Keywords：**data mining；Feature engineering；machine learning

## 1 引言

在本文中，我们使用数据挖掘方法分析与葡萄牙银行机构的直接营销活动（电话）相关的数据。目标是预测电话销售银行长期存款的成功率。分类目标是预测客户是否会认购定期存款（变量y）。对于数据集我们需要考虑特征选择，在数据集中有一些无用的特征或有太多的噪声，这些噪声会降低预测结果。数据与葡萄牙银行机构的直接营销活动有关。营销活动以电话为基础。通常，需要与同一客户进行多次联系，才能访问产品（银行定期存款）是否订阅（“是”）。在本文中我们使用机器学习中的一些模型来进行预测长期存款的成功率，使用模型的平均精度来评估不同模型的表现。本文下面由三章组成，特征工程，对数据集进行分析，找出其中有用的工程；数据整理，将数据修改为适合模型训练；模型训练，使用不同的分类模型来进行训练，找出其中表现最好的分类器。

## 2 特征工程

### 2.1 数据集介绍

数据集[1]总共有41188个样本，具有20个特征（age，job，marital，education，default，housing，loan，contact，month，day\_of\_week，duration，campaign，pdays，previous，poutcome，emp.var.rate，cons.price.idx，cons.conf.idx，euribor3m，nr.employed），20个特征的数据类型分别为整型，浮点型和字符串类型。

### 2.2 数据分布

由2.1节可知数据分为字符型和数值型，对于字符型数据我们使用饼状图和柱状图来进行分析，饼状图可以得到每个特征中不同类别的比例，柱状图可以得到特征中每一个类别对标签y中的yes和no所占比例。结果如图2.1和2.2所示：

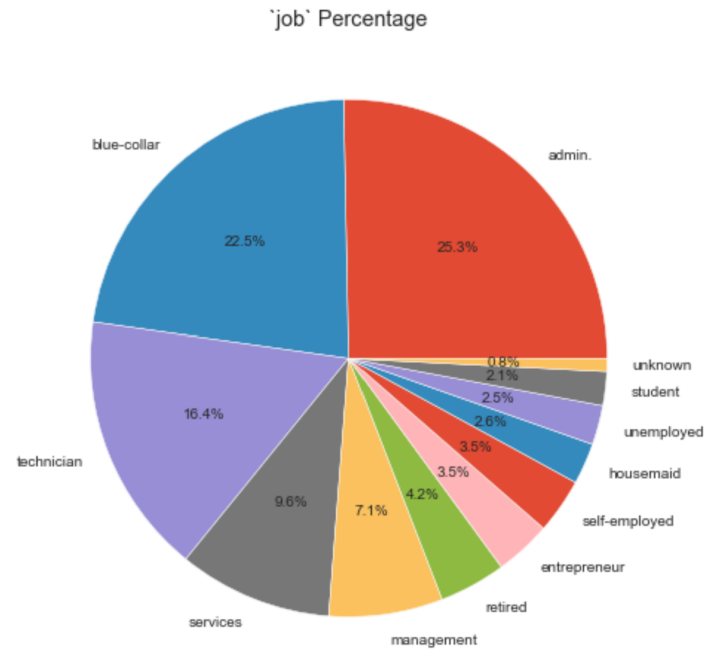


图2.1: job特征比例分布

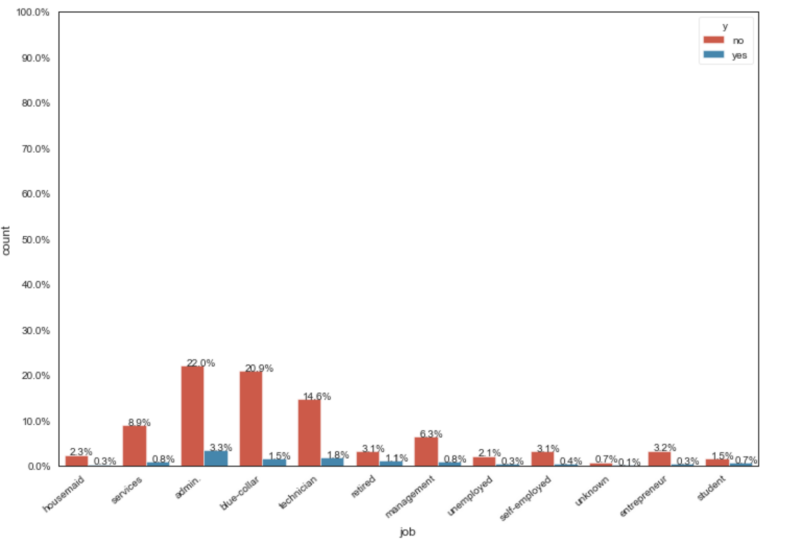


图2.2 不同类别在标签中所占比例

对于每个特征的比例分布在代码中都有详细展示，在此不做赘述。对于数值型数据我们使用饼状图，箱线图和柱状图来查看比例分布，对于数值型数据中不同特征之间的相关性我们使用热力图来表示，如图2.3：

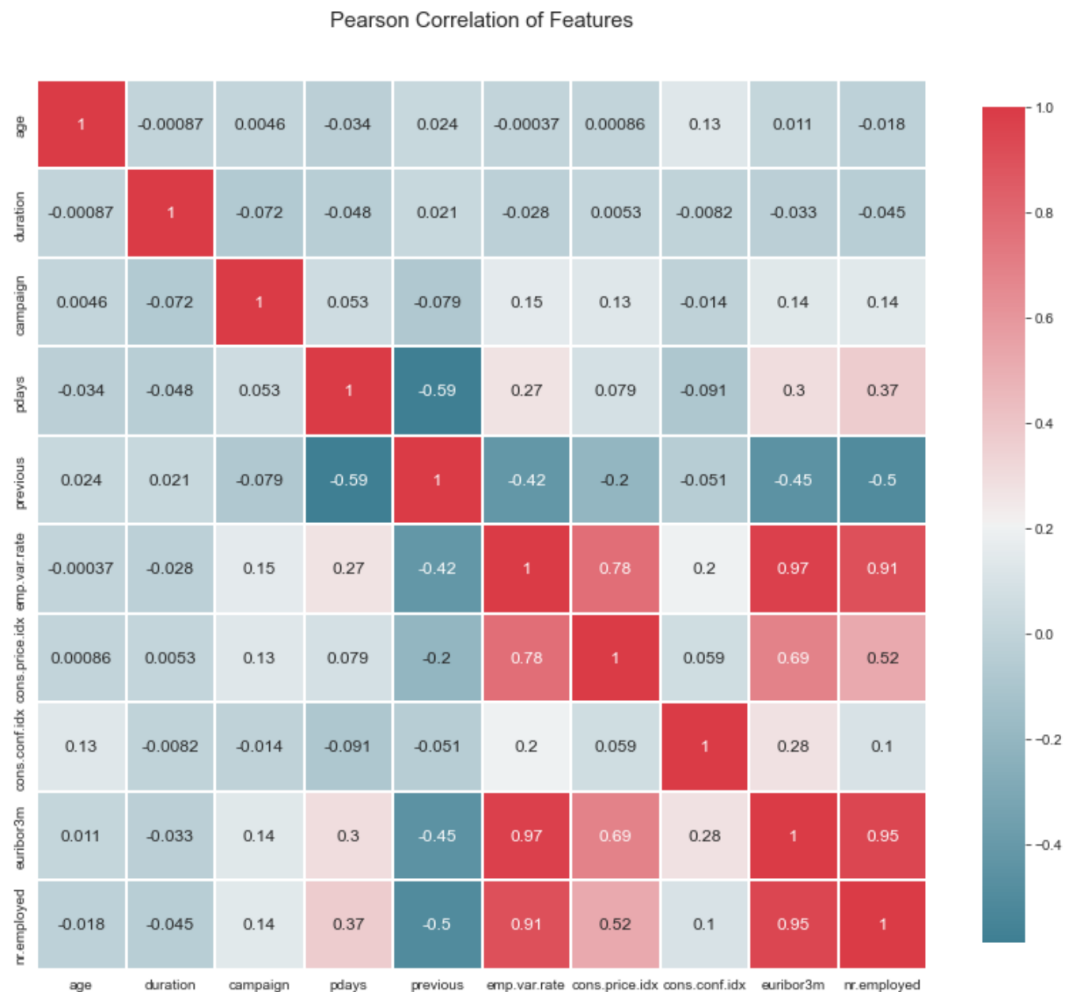


图2.3：特征之间的皮尔逊相关系数

### 2.3 缺失值分析

使用饼状图对数据集进行分析，我们发现某些特征具有空值unknown。并且在不同特征中unknown所占比例不同，为了减少这些空值对预测结果的影响，我们具有两种解决办法，如果unknown所占比例少于5%的话，我们将具有unknown的样本删除，如果大于5%的话，我们将这个特征删除。并且经过分析我们发现将unknown比例大于5%的特征删除之后的预测结果比未删除之前要好。经过画图发现特征job，marital，education，housing，loan中unknown比例少于5%，default中unknown比例为20.9%，所以删除default这个特征。

## 3 数据整理

本节主要由三部分组成，数据校正，数据补全，数据转换。

### 3.1 数据校正

数据校正是为了修改异常数据。通过第二节的分析与参考其他人[2][3]的做法，’material’，’day\_of\_week’，’month’特征对于预测结果作用很小,所以我们将其删除；由第2节可知，’default’中缺失值比例高达20.9%，所以我们将此特征也删除。

### 3.2 数据补全

对于许多特征没有显性缺失，但是具有隐性缺失（unknown）。如job：’unknown’ 0.8%；marital: 'unknown': 0.2%；education: 'unknown' 4.2%；housing: 'unknown' 2.4%；loan: 'unknown' 2.4%。如果我们直接将这些特征的unknown使用每个特征中出现最多的值代替的话会产生噪声，影响预测结果，所以我们选择将出现unknown的样本进行删除。

### 3.3 数据转换

因为’age’特征数值范围太大，所以我们将此特征值进行分段处理，将其转变为类别特征（age\_period）。经过前面的处理后，类别特征有age\_period: (0-16.919, 16.919-33.2, 33.2-49.4, 49.4-65.6, 65.6-98.0)；job: ('admin', 'blue-collar', 'technician', 'services', 'management', 'retired', 'entrepreneur', 'self-employed', 'housemaid', 'unemployed', 'student')；marital: ('married', 'single', 'divorced'；education: ('university.degree', 'high.school', 'basic.9y', 'professional.course', 'basic.4y', 'basic.6y', 'illiterate')；housing ('yes', 'no')；loan ('yes', 'no')；contact ('cellular', 'telephone')；target ('yes', 'no')；month ('jan', 'feb', 'mar', 'apr', 'may', 'jun', 'jul', 'aug', 'sep', 'dec', 'nov', 'dec')；day\_of\_week ('mon', 'tue', 'wed', 'thu', 'fri')。将housing，loan,y中的’yes’=>1,no=>0;其他的类别特征使用fit\_transform()转变为0-1之间的数值。

## 4 模型训练

### 4.1 特征和目标分离与模型参数确定

首先将数据集中’y’与其他特征进行分离；其次使用的模型分别有LogisticRegressionCV()，SGDClassifier()，Perceptron(),KNeighborsClassifier()，BernoulliNB()，GaussianNB()，DecisionTreeClassifier()，LinearSVC()，SVC(probability=True)，BaggingClassifier()，RandomForestClassifier()，AdaBoostClassifier()，XGBClassifier()，LGBMClassifier()，CatBoostClassifier(verbose=False)，对模型中的参数使用的都是默认参数

数据集分为训练集和测试集，训练集占数据集的60%，测试集占数据集的30%。

### 4.2 模型训练结果

各个模型的预测结果如表4.1所示：

表4.1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型名字(MLA Name) | MLA Train  Accuracy Mean | MLA Test  Accuracy Mean | MLA Test Accuracy 3\*STD | MLA Time |
| LGBMClassifier | 0.942789 | 0.91724 | 0.004936 | 0.425527 |
| CatBoostClassifier | 0.952903 | 0.916786 | 0.005364 | 7.177264 |
| XGBClassifier | 0.963736 | 0.914036 | 0.008474 | 0.449606 |
| RandomForestClassifier | 0.999987 | 0.913382 | 0.006828 | 1.669706 |
| AdaBoostClassifier | 0.909651 | 0.911243 | 0.007564 | 0.76998 |
| LogisticRegressionCV | 0.909467 | 0.91037 | 0.003942 | 1.379639 |
| LinearSVC | 0.907652 | 0.908703 | 0.005764 | 0.273751 |
| BaggingClassifier | 0.993134 | 0.907516 | 0.009617 | 0.506717 |
| SGDClassifier | 0.905775 | 0.905971 | 0.007112 | 0.037652 |
| SVC | 0.9044 | 0.903029 | 0.008198 | 33.005436 |
| KNeighborsClassifier | 0.917281 | 0.895243 | 0.007765 | 0.007261 |
| BernoulliNB | 0.887953 | 0.889045 | 0.007768 | 0.016743 |
| DecisionTreeClassifier | 1.0 | 0.888897 | 0.011145 | 0.084303 |
| Perceptron | 0.864692 | 0.86545 | 0.127614 | 0.020237 |
| GaussianNB | 0.833444 | 0.833869 | 0.009151 | 0.013714 |

**参考文献**

1. Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing[J]. Decision Support Systems, 2014, 62: 22-31.
2. <https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c>
3. https://medium.com/nerd-for-tech/analysis-of-bank-marketing-dataset-by-using-support-vector-machine-svm-1ccae6eaa782