基于机器学习的糖尿病识别

摘要：

中国是世界上糖尿病患者最多的国家，病人达到1.1亿，每年有130万人死于糖尿病及其相关疾病。更严重的是，目前中国有超过6亿人处于糖尿病前期，处于这一阶段的人的血糖水平高于正常水平，如果不进行治疗，就有可能发展为2型糖尿病。传统的血糖监测方法需要多次抽血采样，而随着机器学习和大数据医疗的兴起，基于患者医疗体检的数据预测血糖，并判断是否有可能发展为糖尿病有了一定的可行性。

本文的研究目的在于利用机器学习预测患者的血糖值，并预测病人是否是糖尿病。为医生诊断提供参考，减少患者不必要的抽血检查。本文主要的工作包括以下几个方面：

（1）整合了多角度体检特征。如多种转氨酶，球蛋白含量，胆固醇的含量等。对患者数据进行预处理、均衡化，进行特征工程，更全面地描述患者身体指标。

（2）研究了多种回归预测模型，探究最优的参数配置。对包括传统模型如支持向量机、随机森林、XGBoost，最新模型如CatBoot 在内的多种模型及其参数进行了研究，通过多种指标比较了模型的性能，得到性能最佳的模型。

（3）模型融合进行血糖预测，通过stacking，boosting等方法进行了模型融合，进一步提高预测的准确度。

（4）特征重要性排序，在训练的过程中记录特征重要性，确定较重要的几个特征，为医生诊断提供额外的参考。

关键词： 糖尿病 血糖 机器学习

1.引言

正常人的空腹血糖值为3.89～6.1mmol/L。如大于6.1mmol/L而小于7.0mmol/L为空腹血糖受损，如两次空腹血糖大于等于7.0mmol/L考虑糖尿病，建议复查空腹血糖，糖耐量试验。如果随机血糖大于等于11.1 mmol/L可确诊糖尿病。

目前中国有超过6亿人处于糖尿病前期，处于这一阶段的人血糖水平高于正常水平，也就是上述的6.1～11.1mmol/L之间。为了防止发展为2型糖尿病，需要常期去医院抽血验血糖，给患者带来了不便。随着机器学习和大数据医疗的兴起，基于患者医疗过往的体检数据预测血糖，并判断是否有可能发展为糖尿病有了一定的可行性。这将为医生决策提供参考，专注于大概率有可能发展为糖尿病的患者，同时减少普通患者的时间和经济成本。

本文主要的工作包括以下几个方面：

（1）整合了多角度体检特征。如多种转氨酶，球蛋白含量，胆固醇的含量等。对患者数据进行预处理、均衡化，进行特征工程，更全面地描述患者身体指标。

（2）研究了多种回归预测模型，探究最优的参数配置。对包括传统模型如支持向量机、随机森林、XGBoost，最新模型如CatBoot 在内的多种模型及其参数进行了研究，通过多种指标比较了模型的性能，得到性能最佳的模型。

（3）模型融合进行血糖预测，通过stacking，boosting等方法进行了模型融合，进一步提高预测的准确度。

（4）特征重要性排序，在训练的过程中记录特征重要性，确定较重要的几个特征，为医生诊断提供额外的参考。

本文的内容安排如下，第2章阐述几个不同的算法模型原理及关键参数的意义，为训练提供理论指导。第3章对数据进行预处理，包括对查看数据分布和图形化显示，缺失值填充、重要性排序及归一化等，为模型输入做好准备。第4章开始训练、评估模型，并详细介绍了其调参的过程和意义，并对实验结果进行分析。最后通过模型融合得到最终的结果。第5章对本文进行总结和延伸思考。第6章是参考文献。

2. 模型简介

（待完善）

机器学习的算法有很多，本文这里选取了性能较好的常见模型如RandomForest、XGBoost等，也尝试了最新推出的CatBoost模型，最后尝试了CNN神经网络模型。这里对各个模型做简要介绍。

2.1随机森林（RandomForest）

模型介绍

重要参数意义

2.2极限梯度提升（XGBoost）

模型介绍

重要参数意义

2.3 CatBoost

模型介绍

重要参数意义

2.4卷积神经网络（CNN）

模型介绍

重要参数意义

3.数据预处理及特征工程

3.1 数据集介绍

本数据集采用某数据平台提供的6663例患者的体检数据，已经过脱敏处理。数据集包括41个特征，如下图所示。

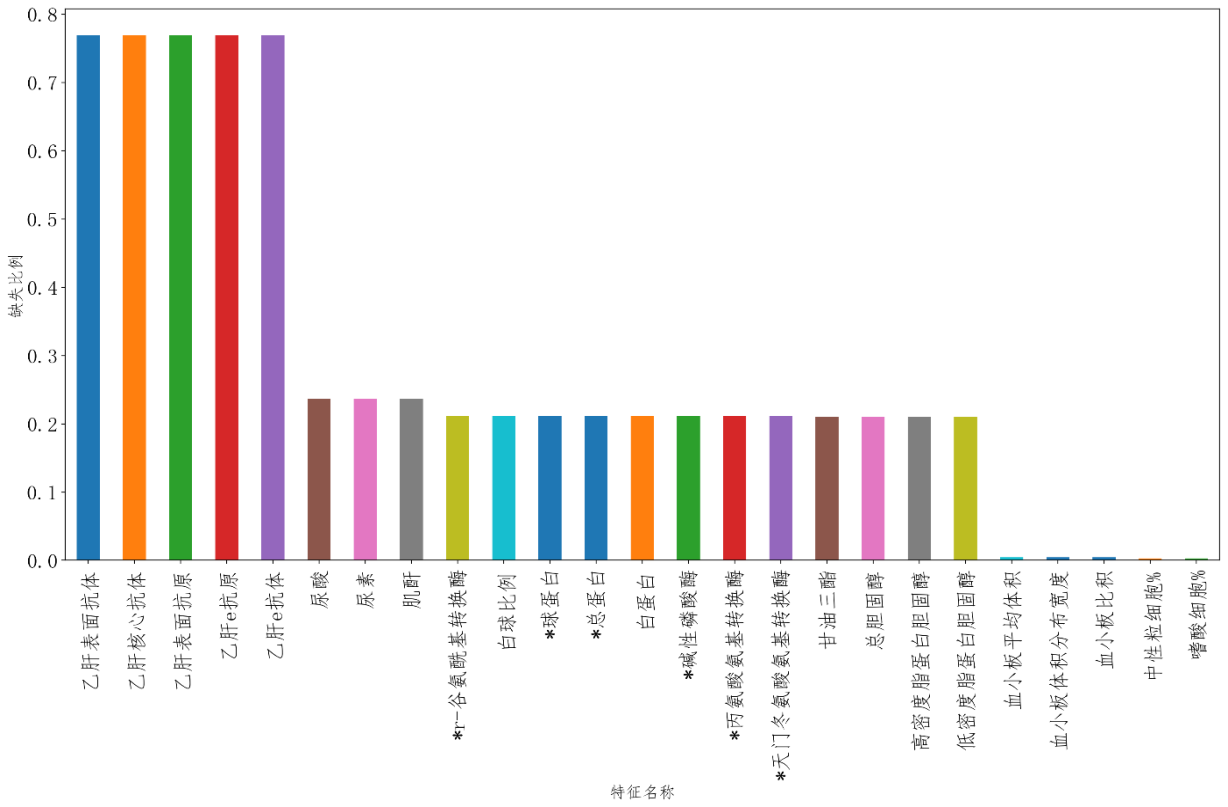
数据包含患者的基本信息如性别、年龄等，也包括详尽的生理病理学指标如氨基酸转换酶、球蛋白含量、淋巴细胞含量等。其中男性3268例，女性3395例，糖尿病患者1411例，占比21%。最后一列是要预测的血糖值。



3.2 数据及特征分析

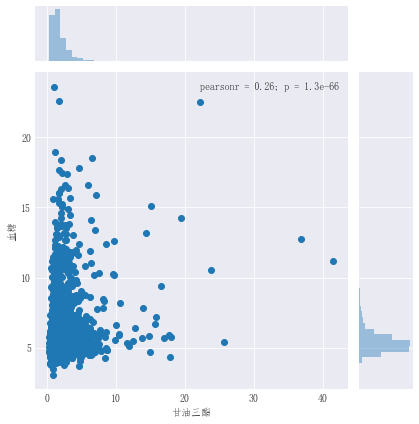
3.2.1 缺失值填充

首先查看数据的缺失值分布情况，对数据集每列中的缺失值占比进行统计，排序后将前25项特征进行可视化显示如下。由图可见，乙肝表面抗体, 乙肝核心抗体, 乙肝表面抗原, 乙肝e抗原, 乙肝e抗体等5个特征缺失值在70%以上，可以直接做删除处理。剩下的缺失值采用-999进行填充。



3.2.2 查看单个特征与血糖的相关性

既然有这么多个特征，一个比较自然的想法就是先看看单个特征和血糖之间的关系。最理想的情况下，如果存在线性的关系则说明该特征对血糖有非常直接的影响。随机选取了10个特征进行可视化展示，如下图所示，发现并没有比较明显的函数关系。但之后我们可以通过机器学习模型找出较为重要的特征。

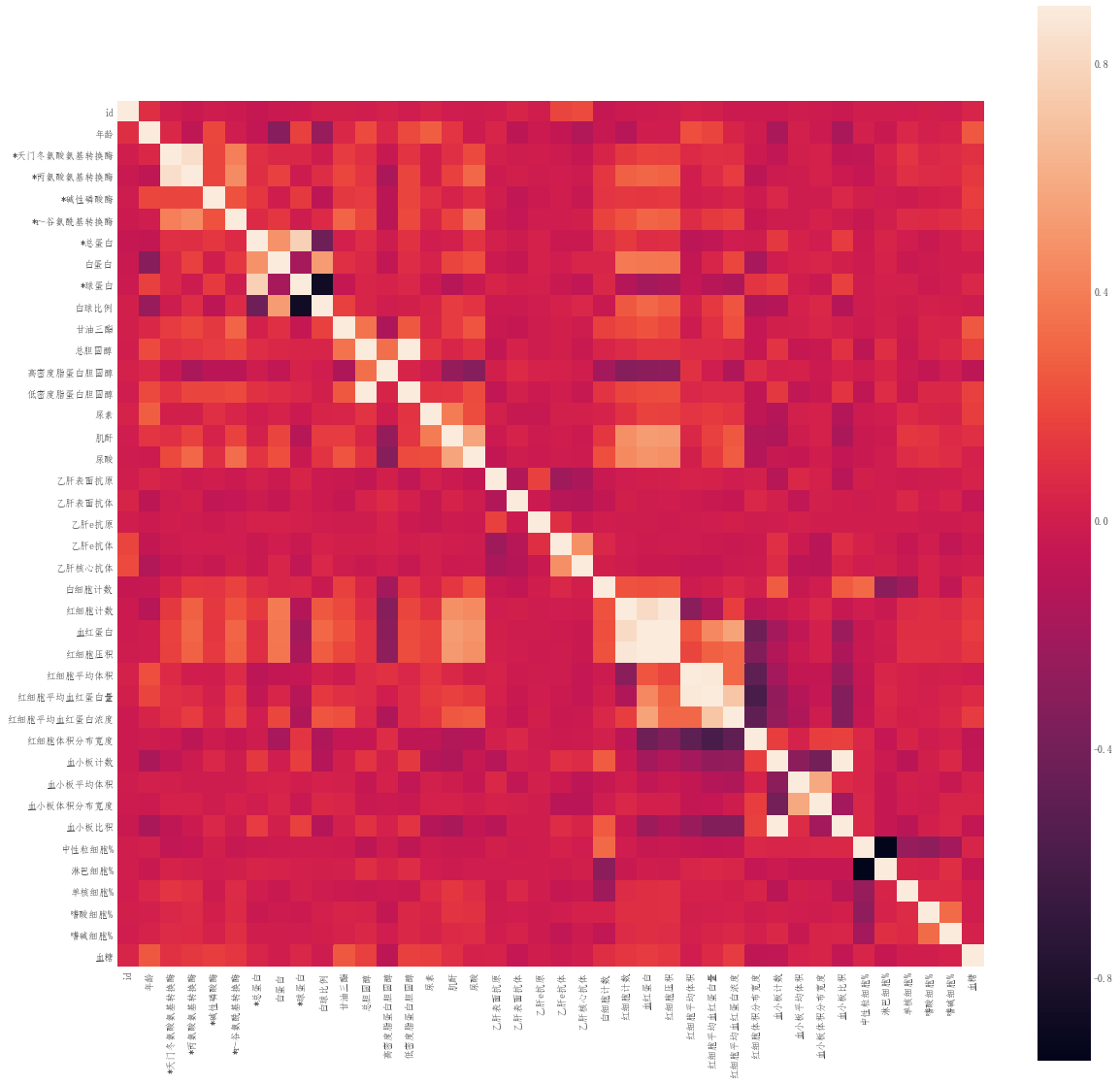


3.2.3 查看各个特征之间的相关性

既然单个特征与血糖之间没有明显的相关性，我们还需要查看两两特征之间是否存在，明显的相关性。如果两个特征之间高度相关，说明其中一个特征是不必要的，可以把它给删除掉，有利于降低输入的维度，减少拟合参数并提高模型的鲁棒性。

通过查看特征的协方差矩阵可以做到这一点。协方差是反映的变量之间的二阶统计特性，如果向量的不同分量之间的相关性很小，则所得的协方差矩阵几乎是一个对角矩阵。对于一些特殊的应用场合，为了使随机向量的长度较小，可以采用主成分分析（PCA）的方法，使变换之后的变量的协方差矩阵完全是一个对角矩阵，之后就可以舍弃一些能量较小的分量了。

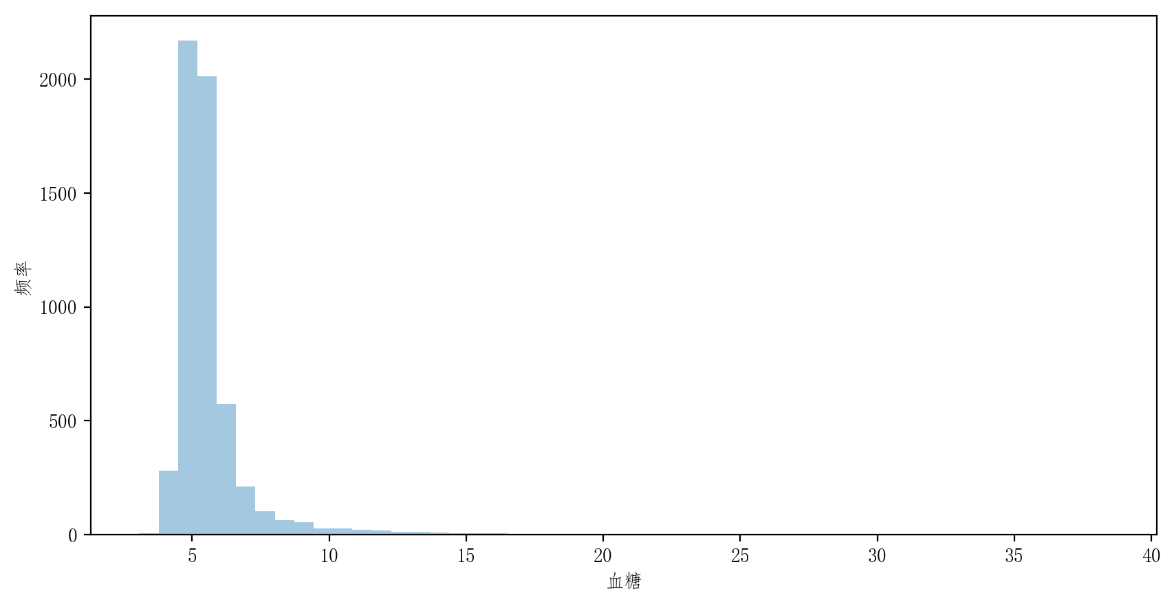
对协方差矩阵进行可视化显示如下图，颜色接近浅色说明正相关程度越高，颜色接近深色说明负相关程度越高，居中的颜色说明相关性不明显。可见协方差矩阵除对角线外的相关系数都非常小，说明两两特征直接不存在比较明显的相关关系，已经是比较独立的了，无需进一步约简。

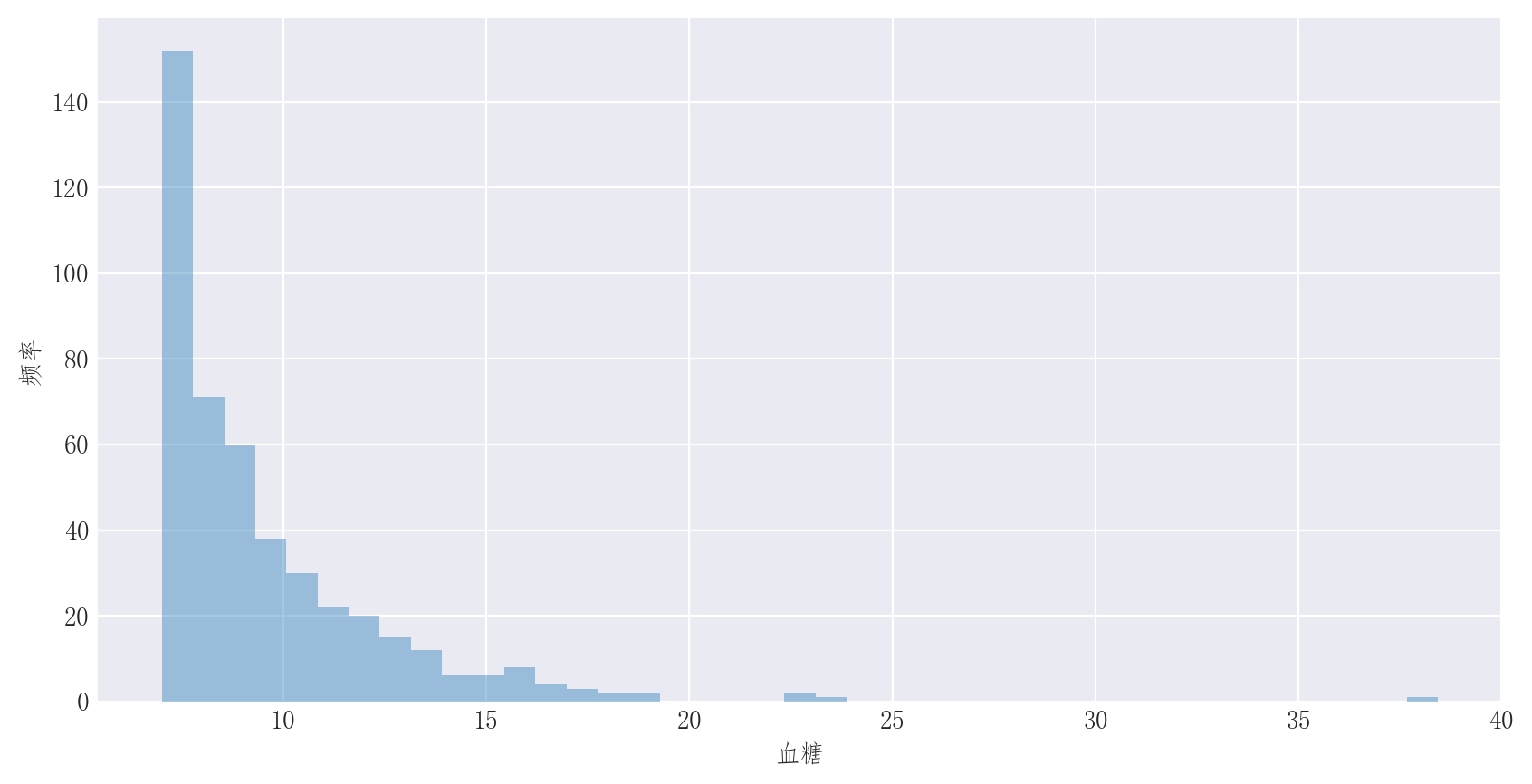


3.2.4 查看血糖分布

因为血糖是要预测的值，因此我们首先得知道训练的样本中的血糖分布大概是怎样的，是均匀分布、正态分布还是呈现其他特征。如果训练集呈现明显的不均衡的情况，我们就需要对其进行过采样或者欠采样等操作使得样本比较均衡，如此才能保证模型的预测比较全面准确。

对训练集中的血糖值进行频率统计并绘制直方图，如下图所示。横轴是血糖值，纵轴是出现的频次，可见血糖高度集中在4-10之间，但也有极少数偏差很大的点。对高血糖部分（7以上）进行放大显示如下图。可见随着高血糖的样本还是少数，需要对其进行过采样处理，尽量保证样本的均衡。



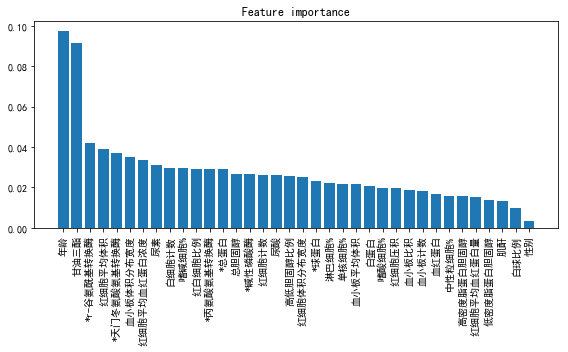


3.2.5 特征选择及重要性排序

作为特征分析的最后一步我们对特征的重要性进行一个排序，将重要性较低的特征抛弃可以起到降低模型复杂度的作用，这也称作特征选择。特征选择在维数很高时是非常有必要的。常用的选择方法有PCA, RFE, 基于模型的特征选择方法等。这样采用基于随机森林模型的特征选择方法。输出的特征重要性排序如下：

根据特征重要性排序可以看出，年龄和甘三油脂对血糖的影响比较明显，这也比较符合常理。年龄不必多说，甘油三酯据查是由于胰岛素的敏感性不够，所以需要更多的胰岛素来维持血糖正常水平，同时这个时候甘油会结合脂肪酸变成甘油三酯。甘油三酯>2.8mmol/L的成年人差不多八成在5-8年后成为糖尿病患者或者糖尿病高危人群。

由于只有30多个特征，因此我们暂时不对特征进行约减，将全部特征都用于学习。



3.3 数据预处理

3.3.1 缺失值处理

一般来说缺失值处理是数据处理的第一步。处理可以分为直接删除和填充两种。对于缺失值较多的特征或样本可以直接删除，如前面看到的乙肝表面抗体等5个特征。对缺失值较少的可以采用填充。填充又可以分为均值、中值、众数、负值填充等几种。经过比较发现，这里采用-999值作为填充效果较好。

3.3.2 类别特征处理

由于机器学习只能处理数值型数据，因此要对字符串型的类别数据进行编码。在此处表现为性别特征的“男”“””女“等两种类别。类别特征常用的编码方法有独热编码（One-Hot）和二值化编码（Binarizer）等。这里由于性别只有两种，我们可以直接通过映射将”“男”设为1，“女”设为0达到编码的效果。

3.3.3 特征转换处理

在进行训练之前需要对特征进行特征转换处理，防止过拟合。特征转换包括标准化、归一化、区间缩放等。标准化（StandardScale）是基于特征矩阵的列，将特征值转换至服从标准正态分布。归一化（Normalizer）是基于矩阵的行，将样本向量转化为单位向量。区间缩放（MinMaxScale）是基于特征矩阵的列，将特征值按比例缩放到[0,1]区间上。

本文中采用区间缩放作进行特征转换。

3.3.4 数据集拆分

交叉验证（Cross Validation）是用来验证分类器的性能一种统计分析方法，基本思想是把在某种意义下将原始数据进行分组，一部分做为训练集，另一部分做为验证集，首先用训练集对分类器进行训练，在利用验证集来测试训练得到的模型，以此来做为评价分类器的性能指标。常见的交叉验证方法如下：

（1）Hold-Out Method

将原始数据随机分为两组，一组做为训练集，一组做为验证集，利用训练集训练分类器，然后利用验证集验证模型，记录最后的分类准确率为此分类器的性能指标。此种方法的好处的处理简单，只需随机把原始数据分为两组即可，其实严格意义来说Hold-Out Method并不能算是CV，因为这种方法没有达到交叉的思想，由于是随机的将原始数据分组，所以最后验证集分类准确率的高低与原始数据的分组有很大的关系，所以这种方法得到的结果其实并不具有说服性。

（2）Double Cross Validation

做法是将数据集分成两个相等大小的子集，进行两回合的分类器训练。在第一回合中，一个子集作为training set，另一个便作为testing set；在第二回合中，则将training set与testing set对换后，再次训练分类器，而其中我们比较关心的是两次testing sets的辨识率。不过在实务上2-CV并不常用，主要原因是training set样本数太少，通常不足以代表母体样本的分布，导致testing阶段辨识率容易出现明显落差。此外，2-CV中分子集的变异度大，往往无法达到“实验过程必须可以被复制”的要求。

（3）K-fold Cross Validation

将原始数据分成K组（一般是均分），将每个子集数据分别做一次验证集，其余的K-1组子集数据作为训练集，这样会得到K个模型，用这K个模型最终的验证集的分类准确率的平均数作为此K-CV下分类器的性能指标。K一般大于等于2，实际操作时一般从3开始取，只有在原始数据集合数据量小的时候才会尝试取2。K-CV可以有效的避免过学习以及欠学习状态的发生，最后得到的结果也比较具有说服性。

综上所述，本文采用折交叉验证的方式进行数据集划分，用平均的方均方差（RMSE）对模型整体性能做评估。RMSE的计算公式如下,其中分别是样本预测值和真实值，n为样本个数。

4. 模型训练及评估

（待做）

选用随机森林、XGBoost、CatBoost、CNN四种模型进行训练。下面依次针对各个模型中的重要参数展开实验，并通过交叉验证的方法进行模型评估。最后通过模型融合得到最终的模型。

4.1 随机森林

最大深度（max\_depth）和决策树数量（n\_estimators）决定了随机森林模型的性能，本文设计了相关的参数评估实验来寻找最佳的模型参数。

4.1.1最大深度（max\_depth）

决策树的最大深度会影响模型运行速度及模型准确性。如果决策树的最大深度太小， 模型准确性较差，但决策树的最大深度达到某个阈值，模型准确性的提升不明显。在其他参数固定的情况下，设置不同的深度得到的RMSE值及运行时间如下图：

4.1.2决策树数量（n\_estimators）

决策树的数量也会影响模型运行速度及模型准确性。在设定了最大深度的基础上，从一系列决策树数量中寻找最优值的结果如下：

4.2 XGBoost

XGBoost主要是调节其超参数，即开始训练之前就已经设置好的参数，直接影响模型的性能。主要参数有：控制学习速率和迭代次数的学习速率和迭代次数的learning\_rate和 n\_estimators；控制 CART生长的 max\_depth 、min\_child\_weight和 gamma；控制行列采样的subsample和colsample\_bytree；正则项系数reg\_lambda。

4.2.1 learning\_rate和 n\_estimators

4.2.2 max\_depth 、min\_child\_weight和 gamma

4.2.3 subsample和 colsample\_bytree

4.2.4 reg\_lambda

4.3 CatBoost

CatBoost的主要参数和XGBoost类似，但是经过的优化，需要调节的参数较少。如下所示。

4.3.1 learning\_rate和 n\_estimators

4.3.2 depth 、l2\_leaf\_reg

4.4 CNN

卷积神经网络是近年来非常热门的神经网络，主要用于对图像的处理。但经过一定的变换，我们也可以用于数据挖掘领域。卷积神经网络的优点在于可以自动选择特征，有非常良好的预测效果。缺点在于参数过多，需要很大的训练样本。我们主要需要调整的是learningrate和dropout。

4.4.1 learningrate

调节learningrate， 查看RMSE。

4.4.2 dropout

调节dropout比例， 查看RMSE。

4.5模型评估

经过上面的实验后，可以得到各个模型比较好的参数。以此为基础用6折交叉验证进行模型评估。得到的各模型评估结果如下图所示：

4.6模型融合

单个模型的预测能力有限，我们希望充分利用多个模型得到更准确的预测结果。常用的模型融合的方法有Voting/ Averaging, Bagging, Boosting, Stacking等。Bagging 和Boosting已经集成到了前面选用的随机森林和XGBoosting, CatBoost模型中了。我们现在只需要采用Averaging 进行模型融合。

Averaging是取平均，稍稍改进的方法是进行加权平均。权值可以用排序的方法确定，比如A、B、C三种基本模型，模型效果进行排名，假设排名分别是1，2，3，那么给这三个模型赋予的权值分别是3/6、2/6、1/6，这便是一种把许多弱分类器这样融合成强分类器的思想。

根据前面实验得到的模型效果排名进行模型融合的结果如下，这也就是我们最终的模型。

5. 总结展望

6. 参考文献