# 摘要

本次课程设计目的是通过对Admission\_Predict数据集对研究生入学情况进行分析与预测，此过程所用到的数据集是美国加州大学洛杉矶分校毕业生的数据集，包括GRE成绩，托福成绩，学校等级，SOP，POR，CGPA成绩，是否参加科研工作和入学概率，其中入学概率即所需预测字段。这里将对各个数据可视化展示并进行深入分析。从相关系数图和多变量关系图中我们可以看到与入学概率相关性最大的三项指标分别为CGPA,GRE,TOEFL成绩。

回归分析是对客观事物数量依存关系的分析是数理统计中的一个常用的方法，是处理多个变量之间相互关系的一种数学方法。回归模型属于监督机器学习。

本次课程设计将根据数据建立将建立各种回归模型（线性回归，决策树回归，随机森林回归），这里将对各个模型的原理进行梳理，分析每个模型的区别，建立预测模型，通过调整不同参数值来建立模型，调整选择数据集的不同数据特征（删除其他关系较为薄弱的属性特征），将通过可视化将预测结果和真实数据对比展示出来，对比各个模型的准确率和均方误差，从而选择正确率最优的预测模型。通过分析对比计算我们最终得到拟合度最优的模型为线性预测模型，其次是随机森林回归模型，最差的是决策树回归模型。其中线性回归和随机森林模型的预测准确率十分接近，都达80%以上。

关键词：研究生入学预测；数据可视化；线性回归；决策回归树；随机森林回归

目录

[摘要 1](#_Toc13165669)

[一、研究背景 3](#_Toc13165670)

[二、主要技术介绍 3](#_Toc13165671)

[2.1 数据可视化 3](#_Toc13165672)

[2.2 线性回归预测模型 3](#_Toc13165673)

[2.2.1 假设函数 3](#_Toc13165674)

[2.2.2 损失函数 3](#_Toc13165675)

[2.2.3 梯度下降函数 3](#_Toc13165676)

[2.3 回归决策树模型 4](#_Toc13165677)

[2.3.1 决策树模型介绍 4](#_Toc13165678)

[2.3.2 回归树原理及步骤 4](#_Toc13165679)

[2.3.3 提高决策树泛化能力 5](#_Toc13165680)

[2.4 随机森林回归模型 5](#_Toc13165681)

[2.4.1 随机森林原理 5](#_Toc13165682)

[2.4.2 建立随机森林模型步骤 5](#_Toc13165683)

[2.4.3 随机森林调参 5](#_Toc13165684)

[2.5 模型衡量标准 6](#_Toc13165685)

[2.5.1拟合优度（R2） 6](#_Toc13165686)

[2.5.2均方误差根（RMSE） 6](#_Toc13165687)

[三、案例分析 7](#_Toc13165688)

[3.1 数据统计及分析 7](#_Toc13165689)

[3.1.1 总述 7](#_Toc13165690)

[3.1.2 各项数据分析 7](#_Toc13165691)

[3.1.3 各项指标与入学概率的相关性 11](#_Toc13165692)

[1) 各指标的相关系数热力图 11](#_Toc13165693)

[2) 多变量关系图 12](#_Toc13165694)

[3) 线性关系拟合图 13](#_Toc13165695)

[3.2 入学机会预测模型 14](#_Toc13165696)

[3.2.1 前期数据准备 14](#_Toc13165697)

[3.2.2 线性回归模型 14](#_Toc13165698)

[3.2.3 决策树回归模型 15](#_Toc13165699)

[3.2.4 随机森林回归模型 17](#_Toc13165700)

[3.2.5各回归预测模型对比 19](#_Toc13165701)

[1) 各回归模型预测值对比图 19](#_Toc13165702)

[2) 各回归模型拟合优度对比图 19](#_Toc13165703)

[3) 各回归模型误差对比图 20](#_Toc13165704)

[3.2.6 去掉个别特征后准确率对比 20](#_Toc13165705)

[3.2.7最终模型结果 21](#_Toc13165706)

[四、结论 21](#_Toc13165707)

[五、参考文献 22](#_Toc13165708)

# 一、研究背景

如今越来越多的人报考研究生，通过科学方法分析研究生入学的各项标准，可以帮助人们帮助学生能够更好的了解是否能够进入想要的大学进行研究生学习，同时也可以给到他们一些参考。

# 二、主要技术介绍

## 2.1 数据可视化

通过绘制扇形统计图，条形统计图，热力图，曲线图等图形，直观的展示数据，将数据可视化，易于分析。绘制图形所用到的python模块有matplotlib.pyplot，seaborn。

## 2.2 线性回归预测模型

### 2.2.1 假设函数

按照惯例x0 = 1。此模型的参数包括θ0~θn，我们不把他们看作n个相互独立的变量，而是将其视为一个 n+1维的向量。所以我们可以把此模型的参数看作模型自己的一个向量。



### 2.2.2 损失函数

衡量一个假设函数的“损失”，又称作“平方和误差函数”，对所有样本的假设值与真实值之差的平方求和，除以两倍的样本数量。代价函数通过误差项的平方的总和来指定，但不把J当作一个有n+1个参数的函数，而是将其视为参数为θ向量的函数。若总体的真实值与总体的假设值差别巨大，会造成损失函数的值较大，计算多个输入特征时的损失函数如下：

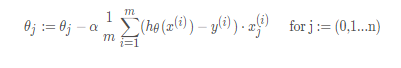


### 2.2.3 梯度下降函数

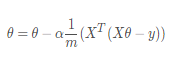
梯度下降：使得自变量沿着使下降最快的方向移动，尽快取得 的最小值，其中α是学习速率（learning rate）



最终目的就是将损失函数的值降为最低，对于梯度下降而言，不同的初始值，移动的方向可能会不同，导致最后收敛的值不同，会造成局部最优解，对于线性回归而言，其代价函数总是凸函数，局部最优解就是全局最优解。



向量化后：



## 2.3 回归决策树模型

### 2.3.1 决策树模型介绍

一颗决策树包含一个根节点、若干个内部节点、若干个叶子节点，叶子节点对应于决策结果，其他每个节点对应于一个属性测试。树模型可以解决非线性特征的问题，树模型不要求对特征标准化和统一量化(即数值型和类目型特征都可以直接被用到树模型的构建和预测过程)，树模型可以直观地输出决策过程，使得预测结果具有可解释性。

### 2.3.2 回归树原理及步骤

回归树是可以用于回归的决策树模型，一个回归树对应着输入空间（即特征空间）的一个划分以及在划分单元上的输出值。回归树对输入空间的划分采用一种启发式的方法，会遍历所有输入变量，找到最优的切分变量j和最优的切分点s，即选择第j个特征xj和它的取值s将输入空间划分为两部分，然后重复这个操作。

确定下一步分割的普遍原则是，最小化均方误差



步骤：

1. 选择最优的切分变量j和最优的切分点s，求解



遍历所有特征，对固定的特征扫描所有取值，找到使上式达到最小值的对(j, s)

1. 用选定的对 (j, s)划分区域，并确定该区域的预测值；
2. 继续对两个子区域调用上述步骤，直至满足停止条件；
3. 生成回归决策树



回归树叶节点是一个个具体的值；回归树的叶节点返回的是“一团”训练数据的均值，而不是具体的、连续的预测值。

### 2.3.3 提高决策树泛化能力

sklearn中我们可以用来提高决策树泛化能力的超参数主要有

- max\_depth:树的最大深度,也就是说当树的深度到达max\_depth的时候无论还有多少可以分支的特征,决策树都会停止运算.

- min\_samples\_split: 分裂所需的最小数量的节点数.当叶节点的样本数量小于该参数后,则不再生成分支.该分支的标签分类以该分支下标签最多的类别为准

- min\_samples\_leaf; 一个分支所需要的最少样本数，如果在分支之后，某一个新增叶节点的特征样本数小于该超参数，则退回，不再进行剪枝。退回后的叶节点的标签以该叶节点中最多的标签你为准。

这次课程设计中我们采用的减少最大深度和调试最小样本叶片大小来提高决策树的泛化能力，以防产生过拟合状况。

## 2.4 随机森林回归模型

### 2.4.1 随机森林原理

随机森林是决策树的集成算法。随机森林包含多个决策树来降低过拟合的风险。随机森林同样具有易解释性、可处理类别特征、易扩展到多分类问题、不需特征缩放等性质。它建立多个这样的决策树，然后将他们合并在一起以获得更准确和稳定的预测。 这样做最直接的事实是，在这一组独立的预测结果中，用投票方式得到一个最高投票结果，这个比单独使用最好模型预测的结果要好。

### 2.4.2 建立随机森林模型步骤

1. 从原始数据随机有放回的抽取N个样本单元，生成决策或者回归树；
2. 在每一个节点随机抽取m；
3. 最终对每一颗决策或者回归树的结果进行整合，生成预测值。

### 2.4.3 随机森林调参

1. max\_features：

随机森林允许单个决策树使用特征的最大数量。 Python为最大特征数提供了多个可选项。 下面是其中的几个：

Auto/None ：简单地选取所有特征，每颗树都可以利用他们。这种情况下，每颗树都没有任何的限制。

sqrt ：此选项是每颗子树可以利用总特征数的平方根个。 例如，如果变量（特征）的总数是100，所以每颗子树只能取其中的10个。“log2”是另一种相似类型的选项。

0.2：此选项允许每个随机森林的子树可以利用变量（特征）数的20％。如果想考察的特征x％的作用， 我们可以使用“0.X”的格式

1. n\_estimators：

建立子树的数量。 较多的子树可以让模型有更好的性能，但同时让你的代码变慢。 虽然子树量越高，能够改善预测的质量，但在实践中证明，并非越高越好，在到达一定值时最优。

1. min\_sample\_leaf：

最小样本叶片大小。 叶是决策树的末端节点。 较小的叶子使模型更容易捕捉训练数据中的噪声。 一般来说，尽量尝试多种叶子大小种类，以找到最优的那个。

1. n\_jobs：

这个参数告诉引擎有多少处理器是它可以使用。 “-1”意味着没有限制，而“1”值意味着它只能使用一个处理器。

1. random\_state：

此参数让结果容易复现。 一个确定的随机值将会产生相同的结果，在参数和训练数据不变的情况下。

1. oob\_score：

随机森林交叉验证方法。 它和留一验证方法非常相似，但这快很多。这种方法只是简单的标记在每颗子树中用的观察数据。 然后对每一个观察样本找出一个最大投票得分，是由那些没有使用该观察样本进行训练的子树投票得到。

前三种为改善特征，从而提高预测准确率；后三种为提高模型的训练速度。

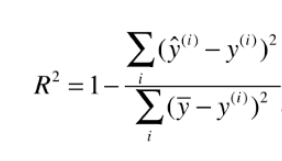
本次实验通过调整子树数量，最小样本叶片大小，最大深度和随机数来建立到最优模型

## 2.5 模型衡量标准

### 2.5.1拟合优度（R2）

准确度：[0, 1]，即使分类的问题不同，也可以比较模型应用在不同问题上所体现的优劣；

是将预测结果转换为准确度，结果都在[0, 1]之间，针对不同问题的预测准确度，可以比较并来判断哪个模型更适合预测此问题

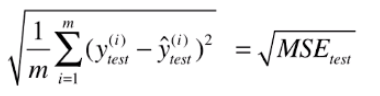
计算公式如下：  


公式上面部分计算的是使用我们的模型预测得到的错误，下面为预测y-y的平均值得到的错误。

### 2.5.2均方误差根（RMSE）

测试数据集中的点，距离模型的平均距离越小，该模型越精确

计算公式如下：



# 三、案例分析

## 3.1 数据统计及分析

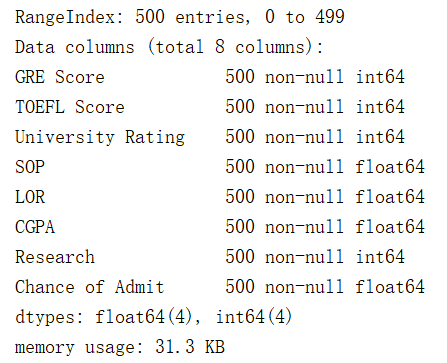
### 3.1.1 总述

本次数据是研究生入学数据，我们的目标是分析数据建立合适的入学该概率预测模型

我们先对各项数据进行分析和统计

该数据共有9列，分别为序号，GRE成绩，托福成绩，学校等级，SOP，POR，CGPA成绩，是否参加科研工作和入学概率，其中

下图为各项数据类型及文件大小

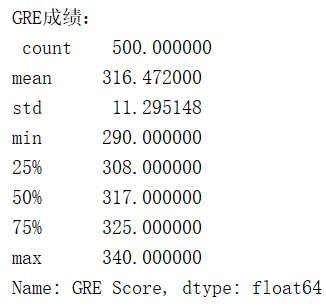


### 3.1.2 各项数据分析

下面三张图均为正态分布图，分别为GRE,TOFEL，CGPA成绩

1. GRE score：

GRE成绩平均分为316.47，标准方差为11.29，最高分为340，最低分为290。



GRE为美国大学生研究生考试，主要分数集中在310-330之间。

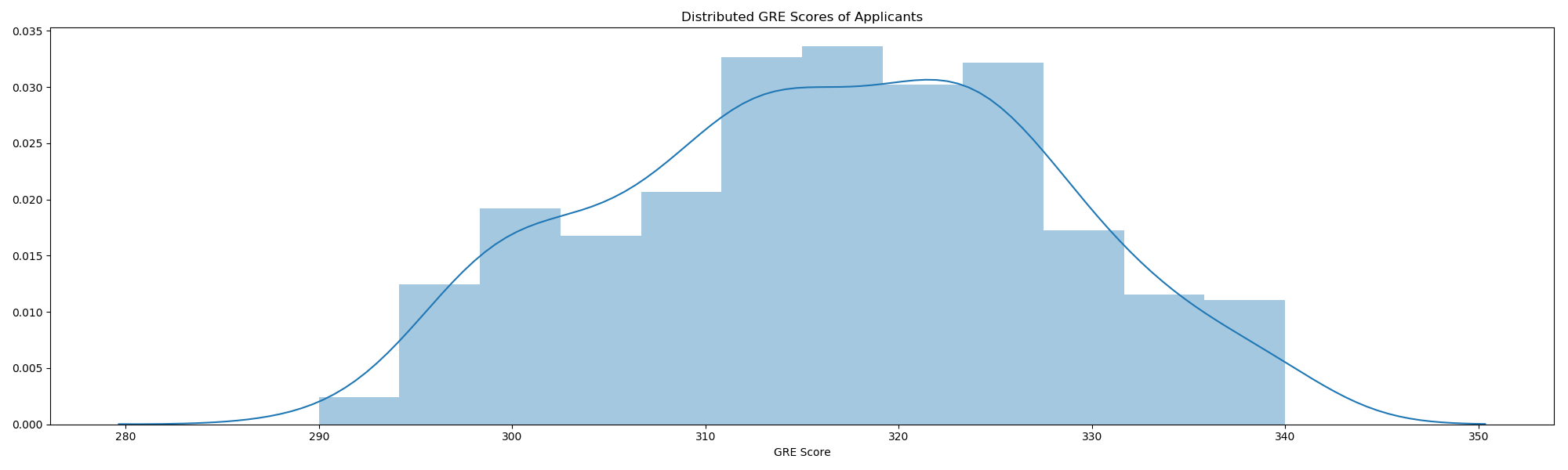
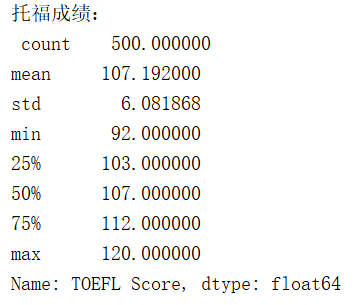


图 1 GRE成绩分布图

1. TOEFL SCORE

托福成绩的平均分为107，最高分为120，最低分为92，标准差为6



OFEL为是由美国教育测验服务社(ETS)举办的英语能力考试成绩，主要集中在100-115之间，呈正态分布

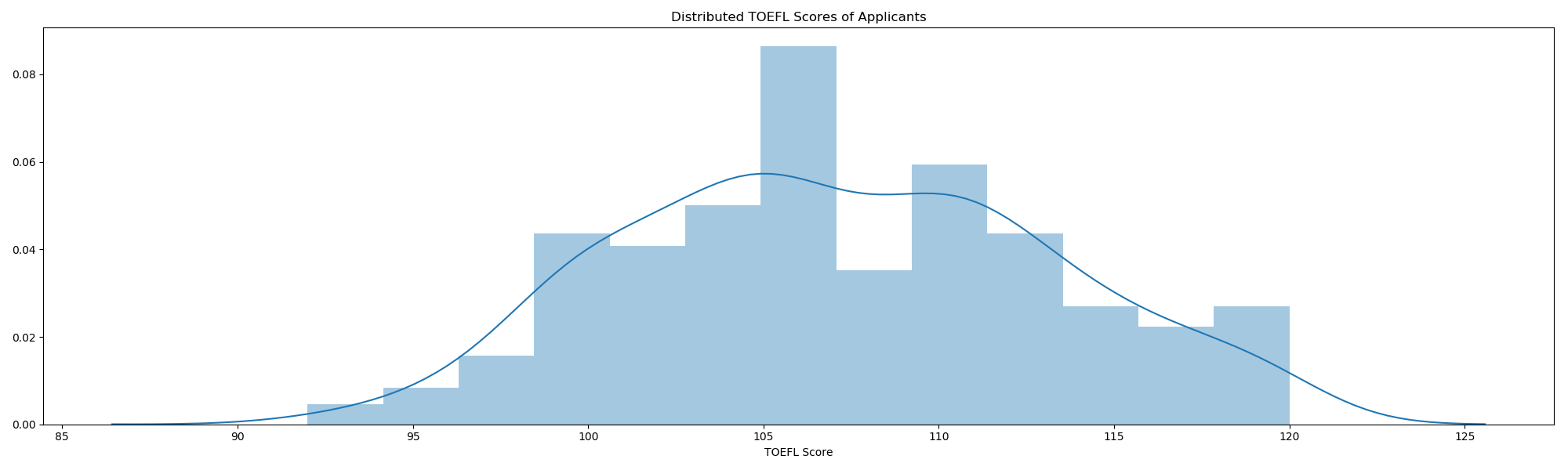
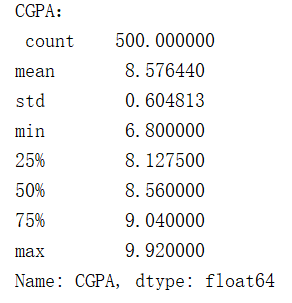


图 2 TOEFL成绩分布图

1. CGPA

平均分为8.58，最低分为6.8，最高分为9.92。



CGPA为所有学期的加权平均分，主要集中在8-9分，呈正态分布

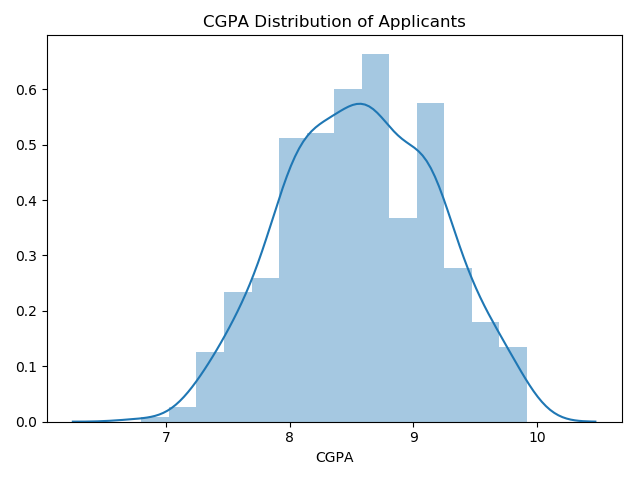


图 3 CGPA成绩分布图

1. University Rating

大学学校等级，其中等级为3的最多

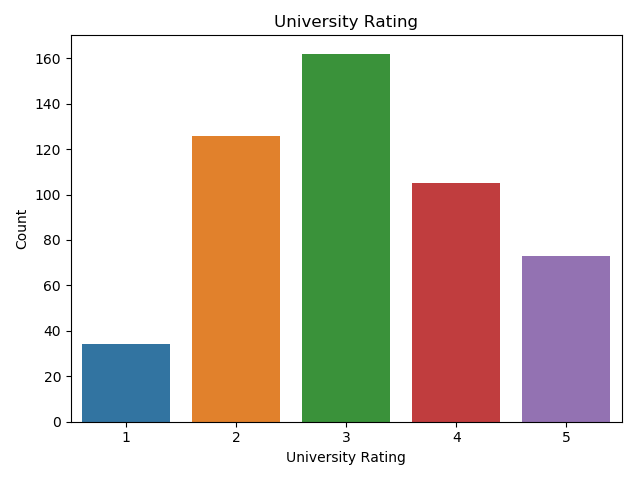


图 4 大学等级统计图

1. SOP

全称为statement of purpose，即自我评估，包括学生本人的研究背景和未来规划等

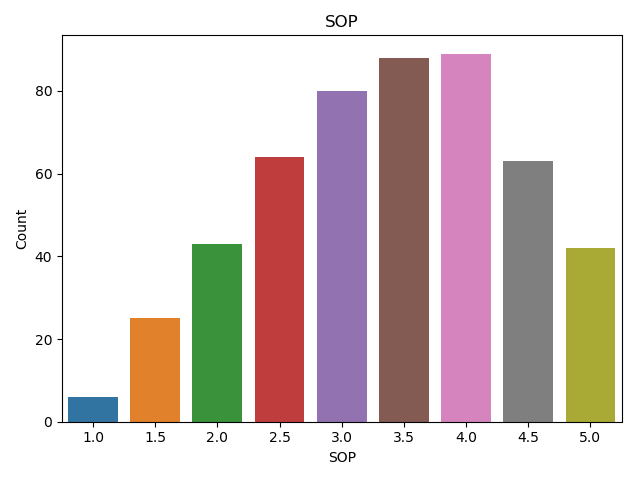


图 5 SOP统计图

1. LOR

LOR为推荐信分值

下图为LOR的分散图，LOR 是推荐信的分数，其中主要集中在3.0-4.5中

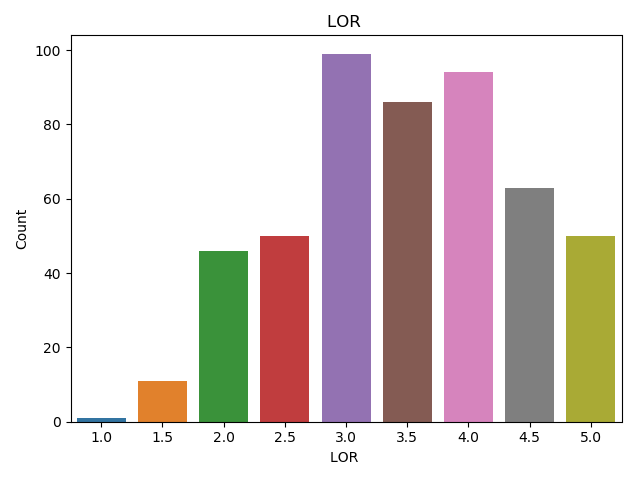


图 6 LOR统计图

1. Research

参加科研人数比如下：

蓝色为没有参加科研工作的占44%，橙色参加科研工作的占56%，参加人数多于没参加的人数，但相差不大。

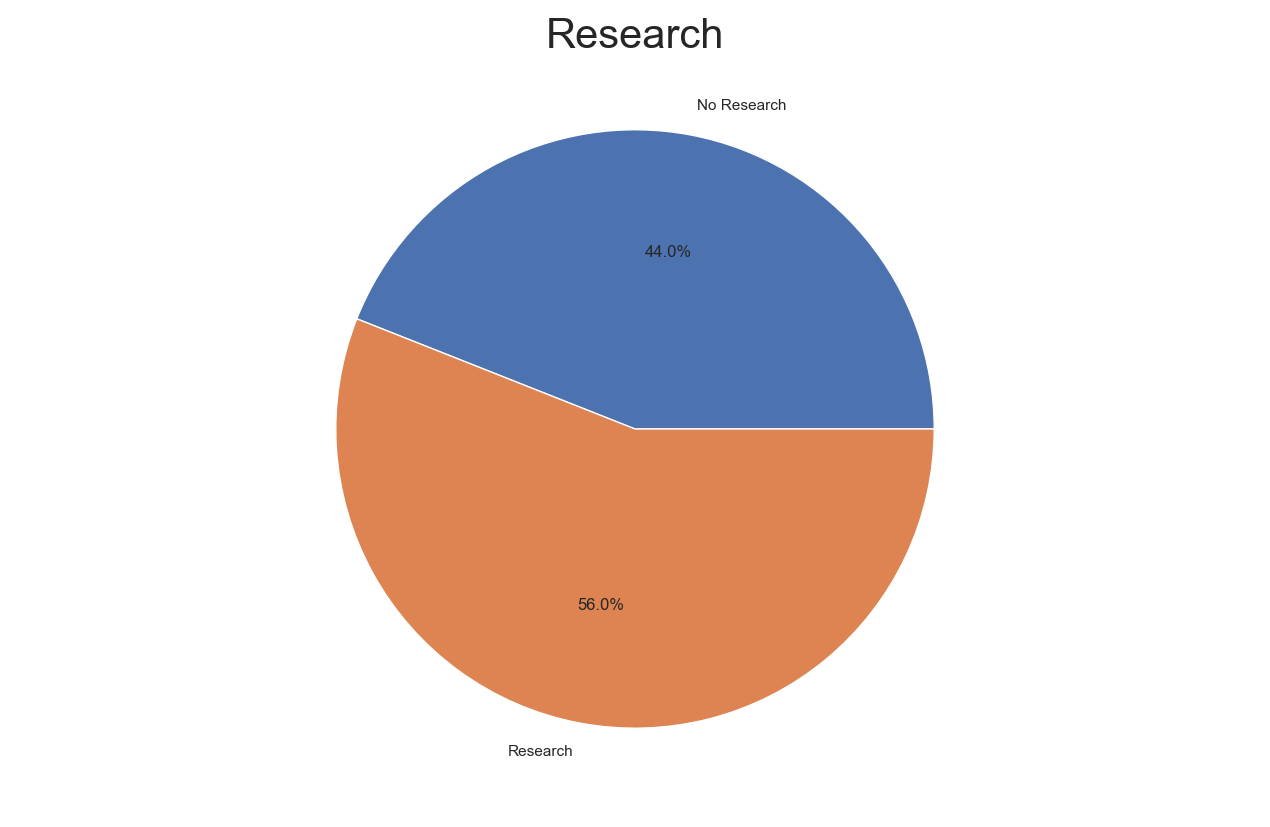


图 7 是否参加科研比例图

### 3.1.3 各项指标与入学概率的相关性

#### 各指标的相关系数热力图

下图为各指标的相关系数热力图，我们可以清楚的看到各指标之间的相关程度，其中与入学概率关系最大的三个指标为CGPA，GRE Score和TOEFL Score，且GRE Score和CGAP、TOEFL Score之间关系密切，由此可以推出大部分平时成绩好的同学往往GRE和托福成绩也会比较好。

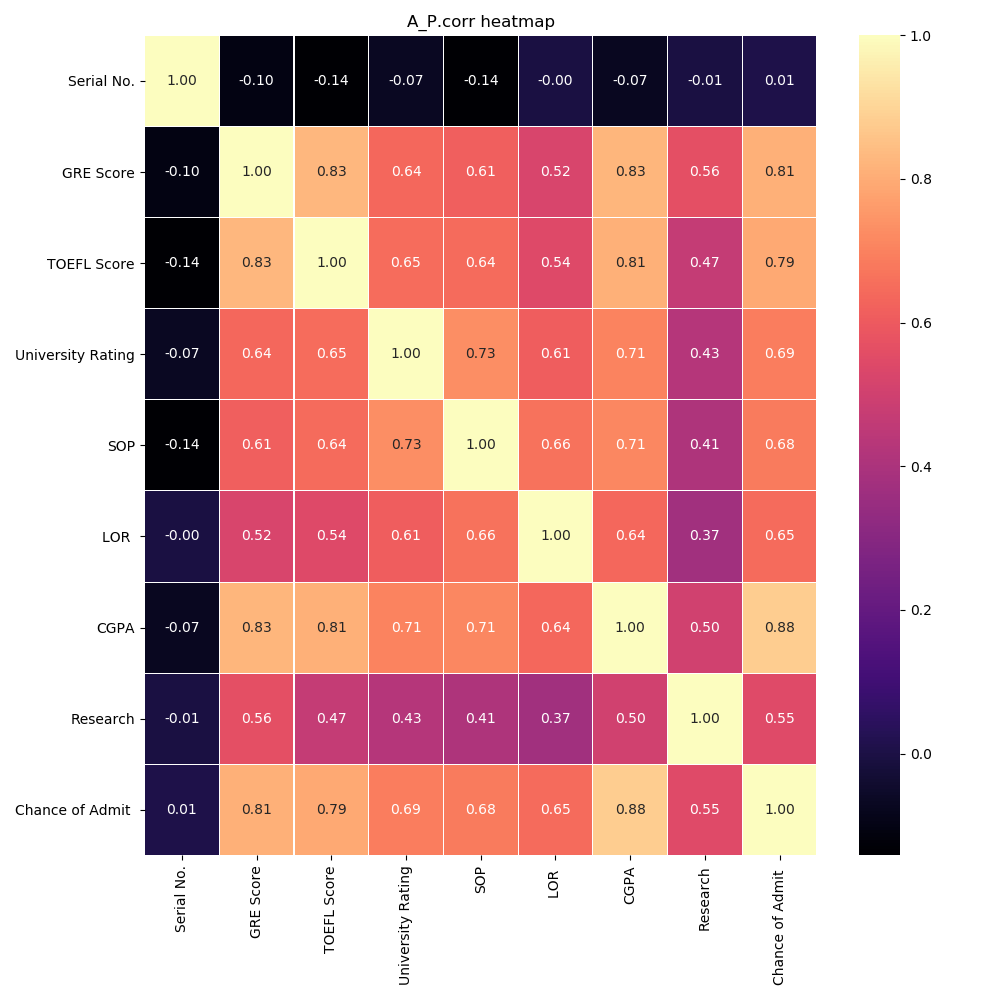


图 8 相关系数热力图

#### 多变量关系图

通过sns.pairplot()方法将各个指标两两比较，画出 下图为各关系之间的散点图，其中CGPA,TOEFL,GRE与chance关系呈线性关系，且三者之间也呈较明显的线性关系

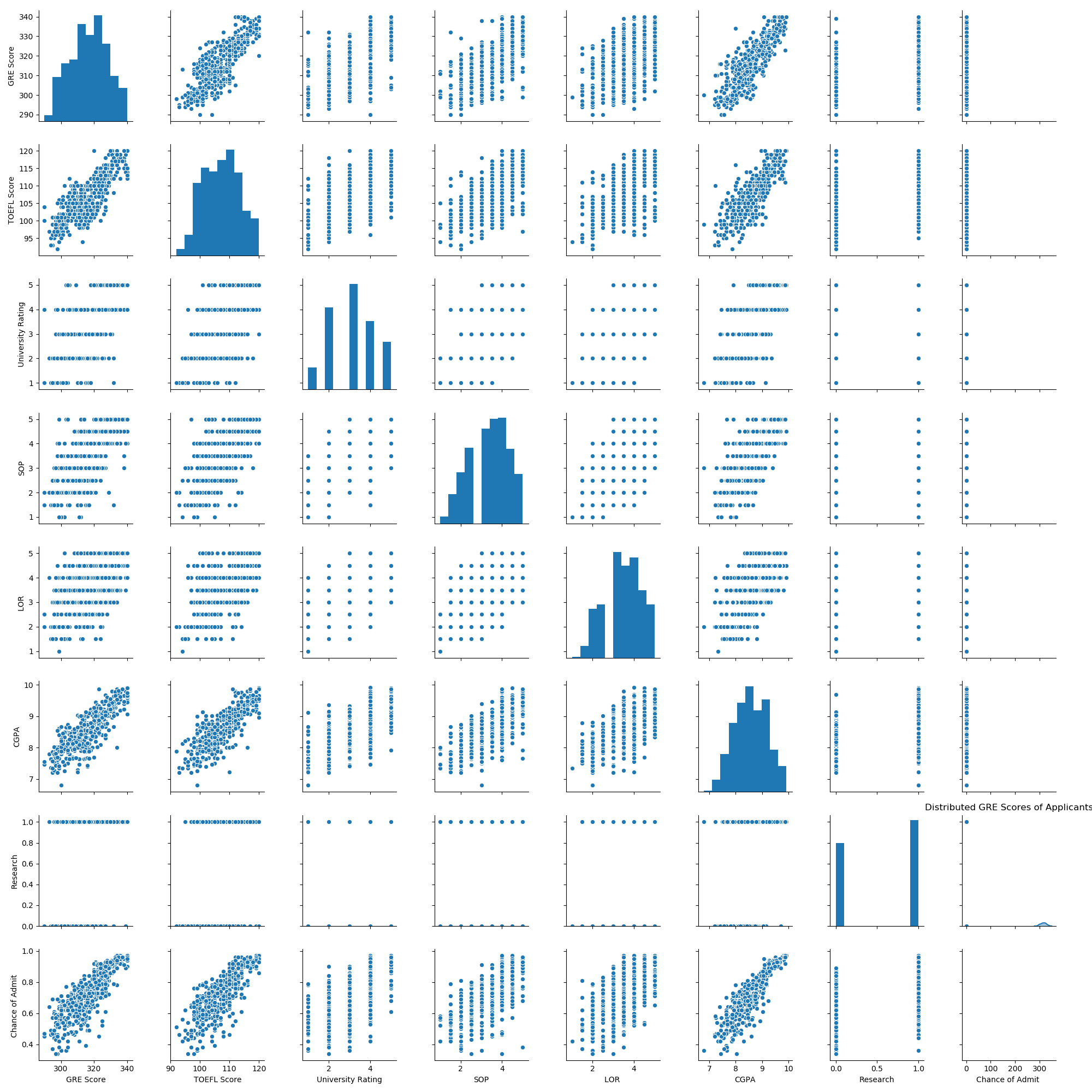


图 9 多变量关系图

#### 线性关系拟合图

CGPA, TOEFL Score，GRE Score和chance of admit的线性关系拟合图通过seaborn中regplot()用于可视化通过回归确定的线性关系。我们从上图相关系数矩阵可看出对CGPA, TOEFL Score，GRE Score对入学机会进行线性拟合，从散点图我们可以看出CGPA的拟合程度大于其他两个指标，从下图还能看出三者的线性关系明显。

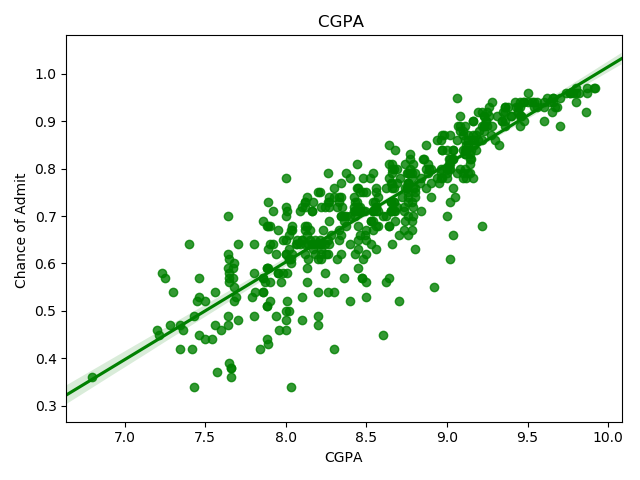


图 10 CGPA线性关系拟合图

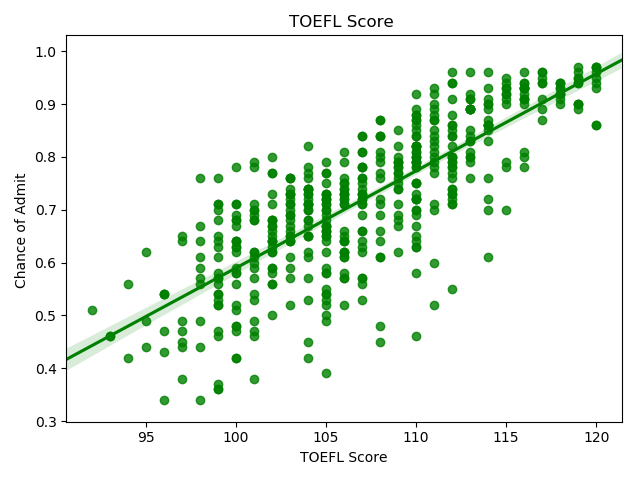


图 11 TOEFL score线性关系拟合图

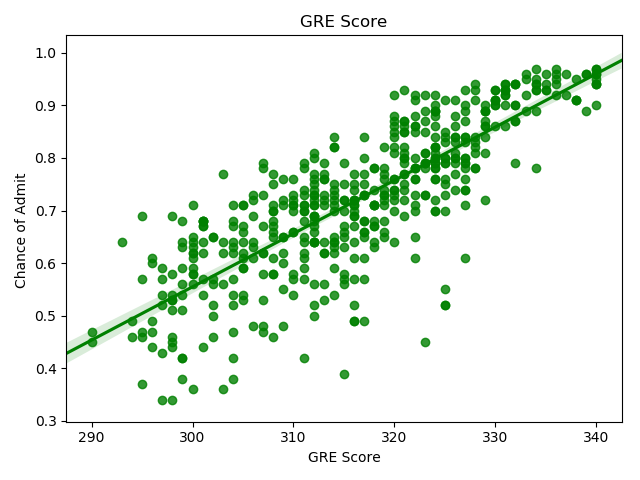


图 12 GRE线性关系拟合图

## 3.2 入学机会预测模型

### 3.2.1 前期数据准备

先读入入学数据，将数据分为建立预测模型所用的训练数据（400条）和测试所用的测试数据（100条），再将数据进行归一化。

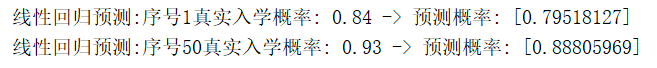
#第一列序号与入学概率之间没有相关性，所以删除掉  
serialNo = df[**"Serial No."**].values  
df.drop([**"Serial No."**],axis=1,inplace = True)  
# 入学机会为预测字段  
y = df[**"Chance of Admit "**].values  
x = df.drop([**"Chance of Admit "**],axis=1)  
# 准备训练数据  
x\_train, x\_test,y\_train, y\_test = train\_test\_split(x,y,test\_size = 0.20,random\_state = 42)  
#数据归一化  
from sklearn.preprocessing import MinMaxScaler  
scalerX = MinMaxScaler(feature\_range=(0, 1))  
x\_train[x\_train.columns] = scalerX.fit\_transform(x\_train[x\_train.columns])  
x\_test[x\_test.columns] = scalerX.transform(x\_test[x\_test.columns])

### 3.2.2 线性回归模型

1. 训练数据创建线性回归模型

lr = LinearRegression()  
lr.fit(x\_train,y\_train)

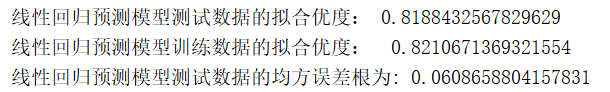
1. 取样进行预测：



1. 使用建立好的模型对于训练数据和测试数据的拟合优度及均方误差根分别如下：







1. 下图为测试数据中线性回归模型预测值与真实值对比：

我们可以看到红色预测线和绿色实际线大致重合，有个别没有吻合

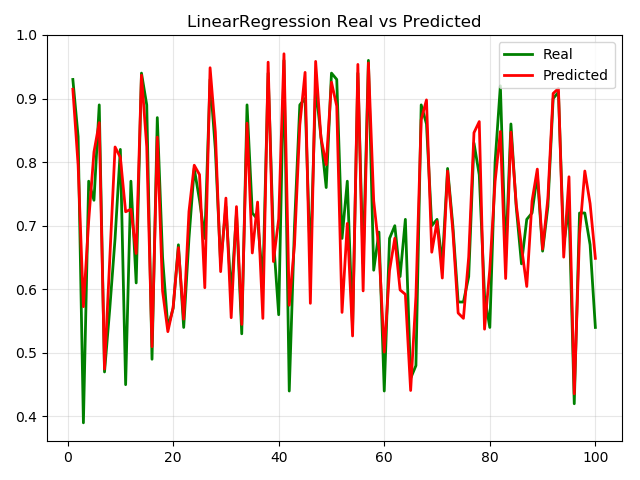


图 13 线性回归模型预测值与真实值对比

### 3.2.3 决策树回归模型

1. 训练数据创建决策树回归模型

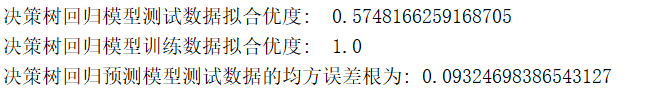
dtr = DecisionTreeRegressor()  
dtr.fit(x\_train,y\_train)

1. 取样进行预测：

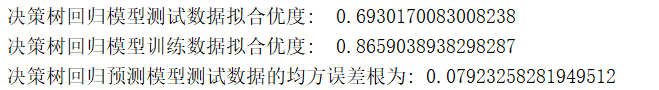


1. 使用建立好的模型对于训练数据和测试数据的拟合优度及误差分别如下：

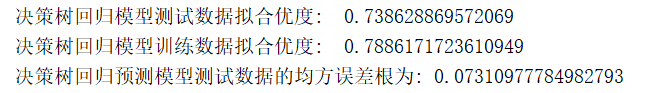
我们看到对于训练数据，回归决策树模型的拟合优度为1，而对于测试数据的拟合优度才0.57，此时出现了过拟合状态。



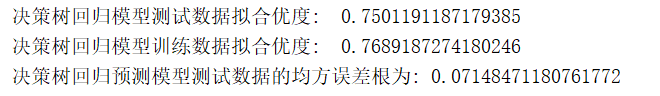
将最大深度改为5时，



将最大深度改为3时 训练数据和测试数据的拟合优度逐渐接近 此时达到决策树模型的最优点。



经过调试最终确定将最小样本叶片大小改成40，决策树回归模型测试数据的拟合优度最高



1. 修改各参数对于误差及拟合优度的比较

我们可以看到当最大深度为3，最小叶片大小为40时，拟合优度最高，均方误差根最小。虽然训练数据的拟合优度下降，但测试数据的拟合优度提升，调整最大深度可提高模型的泛化能力，防止数据过拟合。



表格 1决策树修改各参数结果对比

1. 下图为最终模型测试数据中线性回归模型预测值与真实值对比：

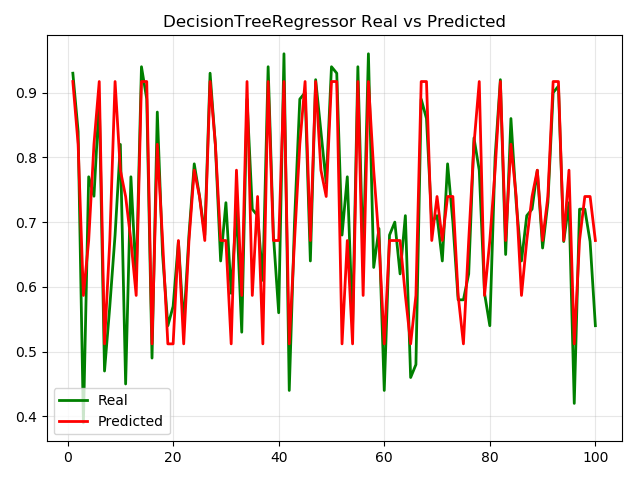


图 14决策树回归模型预测值与真实值对比

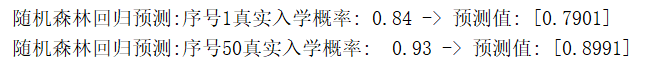
### 3.2.4 随机森林回归模型

1. 训练数据创建决策树回归模型

rfr = RandomForestRegressor(n\_estimators = 100, random\_state = 42)  
rfr.fit(x\_train,y\_train)

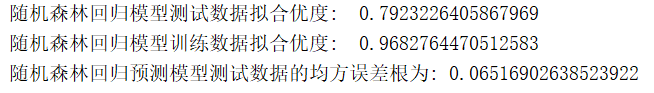
1. 取样进行预测：

两个抽样预测概率均略低于真实入学概率

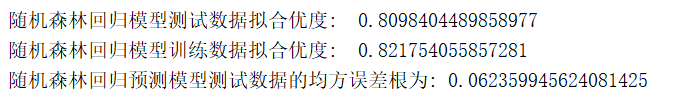


1. 使用建立好的模型对于训练数据和测试数据的拟合优度及误差分别如下：

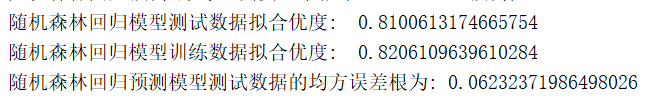
随机森林模型对于训练数据的拟合度高达96%，虽然低于回归决策树模型，但仍然存在过拟合状态，对于测试数据的拟合度为79%，高于决策树预测模型。



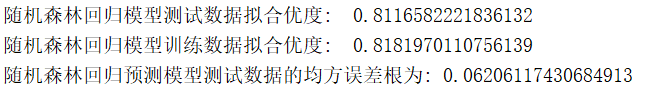
将最大深度改为3后：



再将树的数量改成80后 此时的拟合优度变高。



经过调试，再将最小样本叶片大小改为15，误差略下降，此时拟合优度最高



1. 修改各参数对于误差及拟合优度的比较

我们可以看到当最大深度为3，最大子树个数为80，最小叶片大小为15时，拟合优度最高，均方误差根小。训练数据的拟合优度下降，但测试数据的拟合优度提升，调整最大深度可提高模型的泛化能力，防止数据过拟合。随机数提高模型的训练速度，最小子叶大小通过改善特征来提高模型准确度。



表格 2随机森林模型修改各参数结果对比

1. 下图为最终的随机森林模型测试数据中线性回归模型预测值与真实值对比：

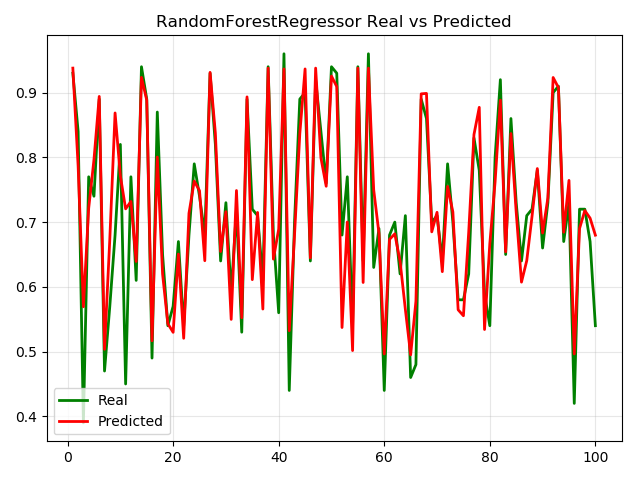


图 15回归模型预测值与真实值对比

### 3.2.5各回归预测模型对比

#### 各回归模型预测值对比图

下图为各回归模型针对测试数据所预测的值，总共有100条数据，这边按每五个取一个预测值，所红点为线性回归模型的预测概率，绿点为随机森林回归模型的预测概率，蓝点为决策树回归模型的预测概率，黑点为真实入学概率。我们从图中看到在入学概率较低的部分，各个模型都不能准确的预测出入学概率，对于入学概率高的准确率较高。这是由于数据不够多，对于低概率部分不能很好的进行预测。

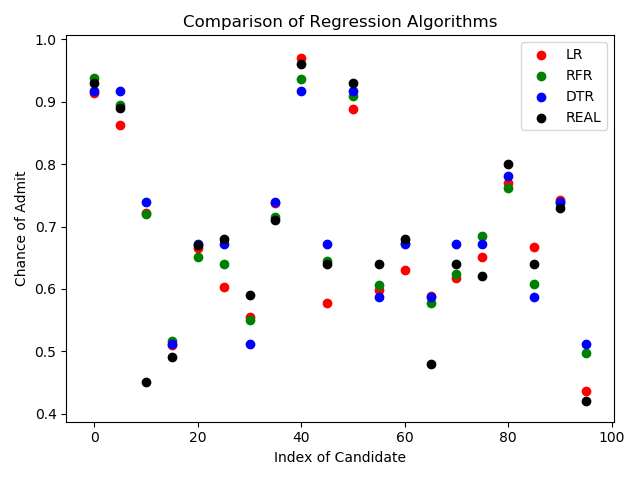


图 16 各回归模型预测值对比

#### 各回归模型拟合优度对比图

我们可以看到线性回归模型的拟合优度最高，决策树拟合优度最低。

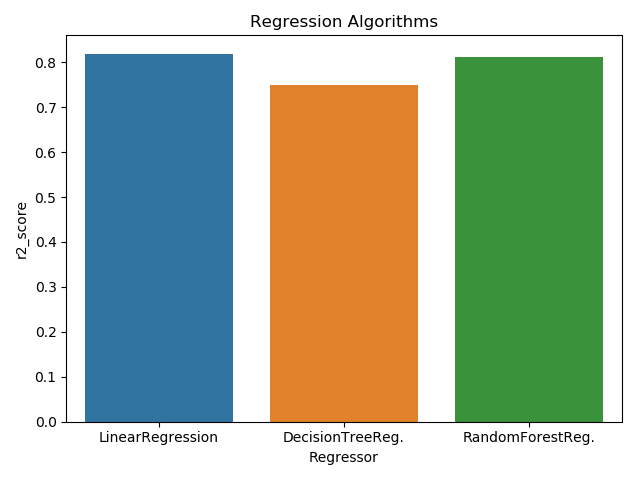


图 17 各回归模型拟合优度对比图

#### 各回归模型误差对比图

下图为各模型的均方误差根，其中误差程度大小：决策树>随机森林>线性回归

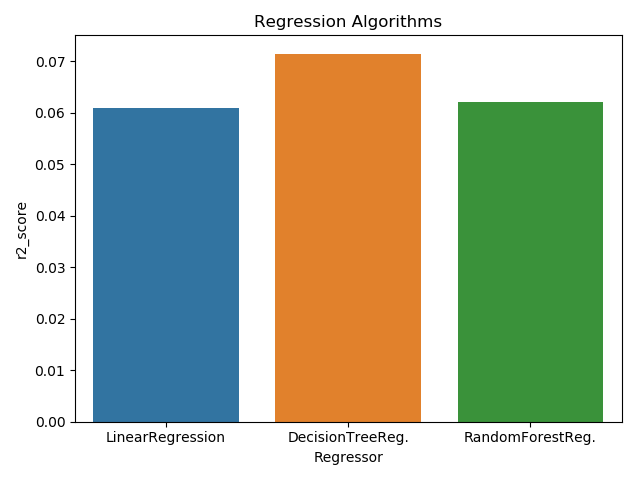
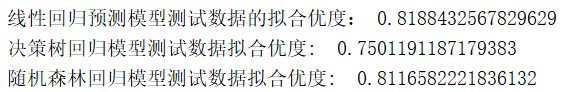


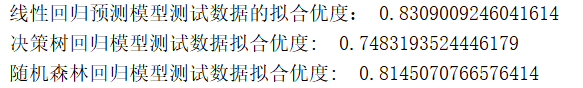
图 18各回归模型误差对比图

### 3.2.6 去掉个别特征后准确率对比

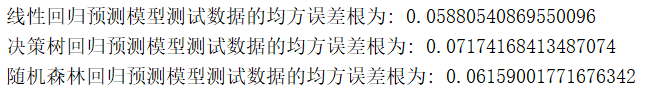
1. 将序号去掉，保留其他所有项得到的拟合优度如下：



1. 由前面图我们可以得到相关性最大的三列为CGPA，GRE，TOEFL成绩，那么我们保留这三列合并成一项除进行预测，我们得到下面的准确率线性回归和随机森林回归高于保留其他项的预测准确率，但决策树略低于回归模型：



1. 最终模型的均方误差根如下图：



### 3.2.7最终模型结果

|  |  |  |
| --- | --- | --- |
|  | R2 | RMSE |
| 线性回归 | 0.830 | 0.058 |
| 决策树回归 | 0.748 | 0.071 |
| 随机森林回归 | 0.814 | 0.061 |

表格 3 最终三种模型评价指标对比

# 四、结论

大学生研究生入学概率关系最大的三个指标为CGPA，GRE，TOEFL成绩，而往往CGPA即平时加权成绩优秀的学生，GRE和TOEFL成绩也很不会差，这和学生本人的女努力勤奋分不开，所以我们可以看出，是否能够进入研究生学习，取决于学生自身的勤奋。因为其三者呈线性关系我们将CGPA，GRE，TOEFL合并成一下在建立模型，可发现效果要优于前面使用全部数据的模型。

通过建立各个模型进行对比分析，其中线性回归预测模型的正确率最高为83.4%，其次是随机森林回归模型81.4%，最差的是决策树回归模型74.8%。

线性回归和随机森林模型的预测准确率十分接近。

适当增加树的数量，可以提高随机森林模型的准确度，但不是越高越好。当最大深度为3，最小样本叶片大小为40时，决策树的对于测试数据拟合优度达到74.8%。

当子树数量为80, 随机数为42,最大深度为3时，最小样本叶片大小为15，随机森林模型对于测试数据的拟合优度达到81.1%。可通过调试找到最大深度，使得决策树回归模型的准确率达到最优。当相关的属性特征越多时，我们得到的预测模型的准确率越高。

# 五、参考文献

[1] 江三宝 毛振鹏. 信息分析与预测〔M〕.北京; 清华大学出版社 北京交通大学出版社 ,2018

[2] 林彬. 多元线性回归分析及其应用[J]. 中国科技信息,2010,36(5):10-12.

[3] 方匡南，吴见彬，朱建平等. 随机森林方法研究综述[J]. 统计与信息论坛, 2011-03,26(3): 32-68.

[4] Peter H. 机器学习实战 [M]. 北京:人民邮电出版社,2013: 161-170.

[5] Breiman L. Random Forest.[J]. Machine Learning, 2001,45(1):532.

[6] Dede. GRADUATE ADMISSIONS. Kaggle