模式识别课程项目

学生成绩预测与回归模型比较

教授 ： 沈莹教授

学生：2233773 黄愈桢

摘要

此研究对于学生的成绩表现进行分析，并通过相关系数指标了解对成绩表现影响最大的因素。

建造了五个预测模型的状况，并设置了三个比较对照组，比较七个较常用的回归模型的差异，使用的是指标是 MSE 与 RMSE。

研究结果指出学生的成绩好坏与父母是有很强的关联性。而外在因素比学生自身学习的影响要大的多。归纳出各个模型的特性与适用场景。

关键词：相关系数、回归模型

目录

第一章　　绪论

学生成绩一直都是在学时最关心的指话题，无论是父母、老师、还是学生自己。考好了，随之而来的是老师的赞扬，父母的喜悦，包括学生自己也会很开心。考的不好，老师会失望，父母会生气。到底是什麼因素在影响成绩的好坏，若是想要提高成绩的话，除了埋头苦读外，该从何处去改变，这应该是所有人都想知道的问题。

回归分析是建模和分析数据的重要工具。回归模型的目的是预测数值型的目标值，它的目标是接受连续数据，寻找最适合数据的方程，并能够对特定值进行预测。目前市面上有很多的回归模型，有些可能是之前较少接触的，想要在实际数据上使用他们，对于参数更加深入了解，了解这些模型的差异和原理并进行比较，并了解这些回归模型各自适合做分析预测的特征与数据。

根据研究动机，本论文研究目的可分为：

1. 找出影响学习成绩的主要因素
2. 探讨父母的因素与成绩高低是否有相关
3. 研究目前常用的几个回归模型，并比较他们的差异与结果

第二章　　文献探讨

第一节　学生成绩表现资料来源

此研究采用的数据集为两所葡萄牙学校的中等教育学生成绩。数据属性包括学生成绩、人口统计、社会和学校相关特征)，并通过学校报告和问卷收集。提供了两个关于两个不同科目成绩的数据集：数学（mat）和葡萄牙语（por)，此研究只使用了数学的数据集。在 Cortez 和 Silva（2008）中，两个数据集在二元/五级分类和回归任务下建模，此研究是使用回归模型建模。

第二节　回归技术与模型选择

回归模型是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。有各种各样的回归技术用于预测。这些技术主要有三个度量（自变量的个数，因变量的类型以及回归线的形状)。此研究将在下面的部分详细讨论它们。

此研究选择了七个模型，分别有线性回归、随机森林、支持向量机、K 近邻算法、极限树、弹性网络、梯度提升等七个模型。挑选了几个之前比较没有接触过的模型进行深入研究，进而了解他们的算法和进行参数调整。

第一项　线性回归(Linear Regression)

线性回归为大家都比较熟知，是最基础的回归模型，通常是人们在学习预测模型时首选的技术之一。在这种技术中，因变量是连续的，自变量可以是连续的也可以是离散的，回归线的性质是线性的。

线性回归使用最佳的拟合直线（也就是回归线）在因变量（Y）和一个或多个自变量（X）之间建立一种关系。

线性回归的特点有

1. 建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快。
2. 可以根据系数给出每个变量的理解和解释
3. 对异常值很敏感

第二项　随机森林(Random Forest)

随机森林是一个包含多个决策树的分类器， 并且其输出的类别是由个别树

输出的类别的众数而定。

随机森林是一种集成算法，它属于Bagging类型，通过组合多个弱分类器，最终结果通过投票或取均值，使得整体模型的结果具有较高的精确度和泛化性能。其可以取得不错成绩，主要归功于“随机”和“森林”，一个使它具有抗过拟合能力，一个使它更加精准。

第三项　支持向量机（Support Vector Machine, SVM）

支持向量机是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。

SVM使用铰链损失函数计算经验风险并在求解系统中加入了正则化项以优化结构风险，是一个具有稀疏性和稳健性的分类器　。SVM可以通过核方法进行非线性分类，是常见的核学习方法之一。

第四项　极限树(Extra Tree)

极限树模型的算法与随机森林算法十分相似，都是由许多决策树构成。极限树与随机森林的主要区别：随机森林应用的是Bagging模型，极限树使用的是所有的样本，只是特征是随机选取的，因为分裂是随机的，所以在某种程度上比随机森林得到的结果更加好，随机森林是在一个随机子集内得到最佳分叉属性，而极限树是完全随机的得到分叉值，从而实现对决策树进行分叉的。

极限树与经典决策树的构建方式不同，为一个非常随机的树回归器。在寻找将节点样本分成两组的最佳分割时，会为每个 max_features 随机选择的特征绘制随机分割，并选择其中的最佳分割。当 max_features 设置为 1 时，这相当于构建一个完全随机的决策树。

第五项　弹性网络（Elastic-Net Regression）

弹性网络是一种使用 L1， L2 范数作为先验正则项训练的线性回归模型，是岭回归和套索回归的混合技术。 这种组合允许拟合到一个只有少量参数是非零稀疏的模型，就像 Lasso 一样，但是它仍然保持了一些类似于 Ridge 的正则性质。

弹性网络在很多特征互相联系的情况下是非常有用的。Lasso 很可能只随机考虑这些特征中的一个，而弹性网络更倾向于选择两个。

第六项　梯度提升（Gradient Boosting）

梯度提升也是 Boosting 中的一大类算法，它的思想借鉴于梯度下降法，而梯度提升是属于前向分布算法。梯度提升每次通过拟合上一次的残差来减小损失，基学习器限定为 CART。其基本原理是根据当前模型损失函数的负梯度信息来训练新加入的弱分类器，然后将训练好的弱分类器以累加的形式进行拟合并计算出该弱分类器的权重，最终实现对模型的更新。

第七项　k 近邻算法（K-Nearest Neighbors Regressor）

KNN 回归器通过找出测试样本的 k 个最近邻样本，将这些近邻样本做平均赋给该样本，就可以得到测试样本的预测值是监督学习算法，意味着训练数据集需要有 label 或者类别，KNN 的目标是把没有标签的数据点（样本）自动打上标签或者预测所属类别。通常用于分类，但也可用于回归。

第三章　　研究方法

第一节　相关系数指标

相关系数是在直线相关的条件下，说明两个变量之间的相关关系密切程度的统计分析指标。简单相关系数的计算公式为：

$$Corr(x,y)=\frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

其中，Cov(X,Y)为 X 与 Y 的协方差，Var(X)为 X 的方差，Var(Y)为 Y 的方差

此研究以相关系数指标的方式去查看各个特征与目标值的关系了解哪个特征与目标值的相关程度最高

第二节　模型比对实验设置

设置三个比对组
比对一：挑选 10 特征模型调参与不调参，特征选择是基于 G3 的相关系数指标，选择了相关系数指标前十个特征
比对二：使用全部特征与挑选 10 个特征，使用的是调过参数的回归模型
比对三：只删除特征”G2”和只删除特征”G2”、”G1”与不删除任何特征做对比，使用的也是调过参数的回归模型

模型比对评估标准
平均绝对误差（MAE）

$$MAE=\frac{\sum_{i=1}^{n}|predicted_i-actual_i|}{n}$$

均方根误差（RMSE）

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n}(predicted_i-actual_i)^2}{n}}$$

第四章　　实验结果

第一节　相关系数指标分析

以下的图表为经过独热编码后，与目标值”G3”的相关系数指标，并从大到小排序。”G3”、”G2”、”G1”这三个特征提出另一个图表来看。
通过以下的图表，除去”G2”、”G1”特征正相关性最高的是母亲的教育程度，再来是学生想要接受更高的教育，第三是父亲的教育程度，第四是学生没有对象的，第五是妈妈的职业是健康护理相关。

负相关性最高的是学生之前是否有挂科，再来是学生不想接受更高的教育，第三是学生的年龄，第四是学生出门的频率，第五是学生有对象。

由正负相关的前五名可以看出，父母对学生的成绩相关性是很高的，尤其是父母的教育程度。

Index	G3	Index	G3	Index	G3
Medu	0.217147	nursery_yes	0.0515679	Pstatus_T	-0.058009
higher_yes	0.182465	famrel	0.0513634	health	-0.0613346
Fedu	0.152457	school_GP	0.0450169	famsize_GT3	-0.0814071
romantic_no	0.12997	famsup_no	0.0391571	schoolsup_yes	-0.0827882
Mjob_health	0.116158	absences	0.0342473	guardian_other	-0.0877744
address_U	0.105756	guardian_father	0.0324932	Mjob_other	-0.0964774
sex_M	0.103456	guardian_mother	0.0223378	internet_no	-0.0984834
paid_yes	0.101996	activities_yes	0.0160997	reason_course	-0.0989496
internet_yes	0.0984834	freetime	0.0113072	paid_no	-0.101996
studytime	0.0978197	Fjob_at_home	-0.0133846	sex_F	-0.103456
reason_reputation	0.0956922	activities_no	-0.0160997	address_R	-0.105756
Fjob_teacher	0.095374	Fjob_services	-0.0161078	Mjob_at_home	-0.115634
schoolsup_no	0.0827882	reason_home	-0.0213592	traveltime	-0.117142
famsize_LE3	0.0814071	famsup_yes	-0.0391571	romantic_yes	-0.12997
Mjob_services	0.0784289	school_MS	-0.0450169	goout	-0.132791
Pstatus_A	0.058009	nursery_no	-0.0515679	age	-0.161579
Mjob_teacher	0.0577124	Walc	-0.0519393	higher_no	-0.182465
Fjob_health	0.0571105	Fjob_other	-0.0534834	failures	-0.360415
reason_other	0.0520077	Dalc	-0.05466		

由下图可以看出”G2”、”G1”这两个特征对”G3”有非常强的相关程度。所以代表说学生要是前面表现好，后面也大概率成绩会较佳，反之亦然。

Index	G3
G3	1
G2	0.904868
G1	0.801468

再来对所有的相关系数做绝对值处理，排出前 12 的特征，后面的模型训练有删掉相关系数相同的”higher_no”与”romantic_no”，为之后训练模型作准备。

```
G2              0.904868
G1              0.801468
failures        0.360415
Medu            0.217147
higher_yes      0.182465
higher_no       0.182465
age             0.161579
Fedu            0.152457
goout           0.132791
romantic_no     0.129970
romantic_yes    0.129970
traveltime      0.117142
Name: G3, dtype: float64
```

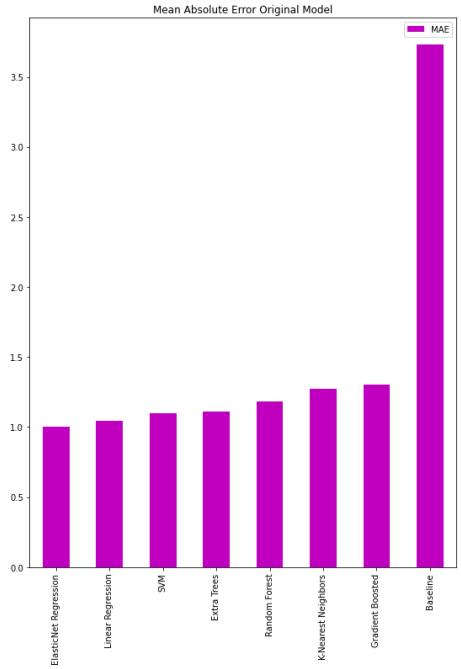
第二节　回归模型预测与比对结果

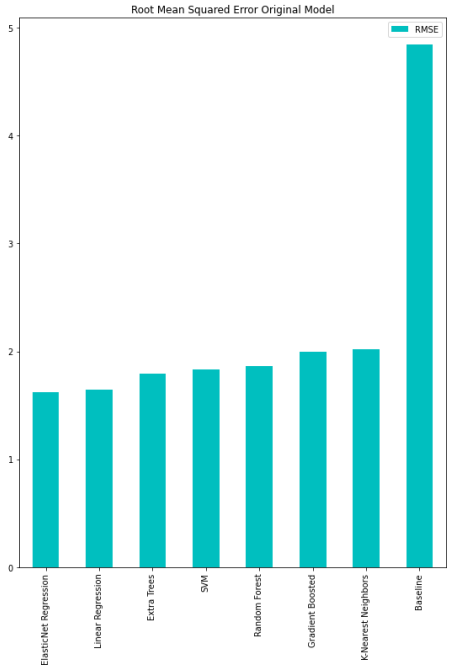
此研究建造了五个实验模型，包括未调整过参数并有挑选特征的原始模型、调整过参数并有挑选特征的模型、调整过参数且未挑选特征的模型、调整过参数且只删掉”G2”特征的模型和调整过参数且删掉”G2”、”G1”特征的模型，下面可以看到五个实验模型的预测结果，而后进行比对。

第一个训练的模型为未调整过参数的原始模型，使用默认值的结果，预测结果由下面的图表表示，图表显示的出使用原始模型的预测结果，左边为 MAE 指标，右边为 RMSE 指标。

可以由图表明显看出弹性网络在原始条件下的预测结果是最好的，而梯度提升的模型预测结果最差。

```
Using origin models:
                            MAE      RMSE
Linear Regression       1.04491  1.648211
Random Forest          1.183608  1.861875
SVM                    1.101555  1.830157
K-Nearest Neighbors    1.273418  2.020277
ElasticNet Regression  0.999633  1.624568
Extra Trees            1.111013  1.796036
Gradient Boosted       1.305414  1.995885
Baseline               3.734177  4.848333
```

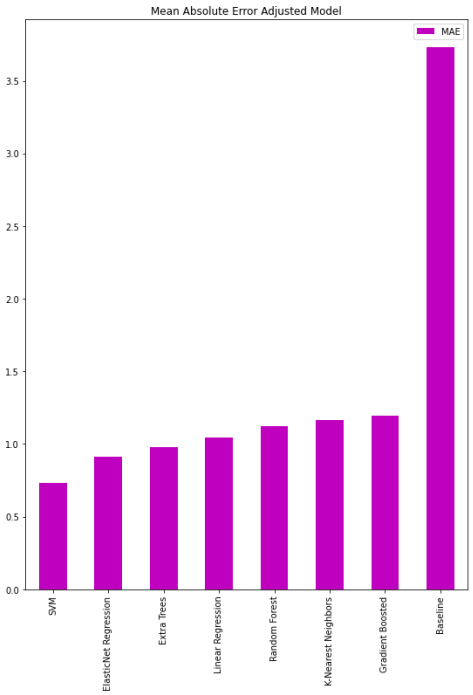


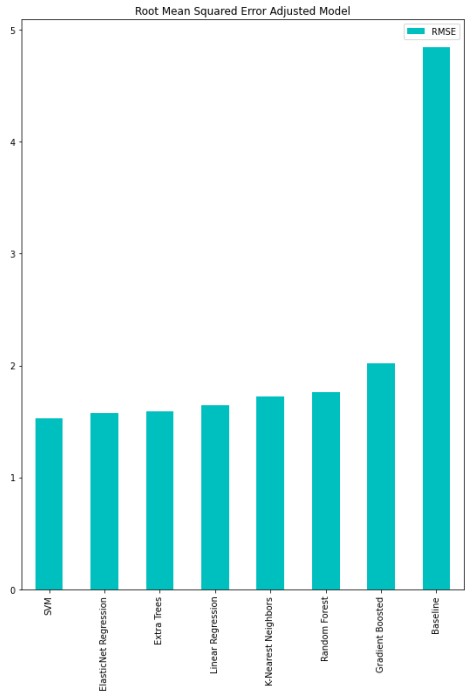


再来是调整参数过后的模型，可以由图表看出 SVM 回归模型在调整参数后的预测结果是最好的。

After Adjust :

	MAE	RMSE
Linear Regression	1.04491	1.648211
Random Forest	1.123901	1.760135
SVM	0.734531	1.528037
K-Nearest Neighbors	1.16261	1.722141
ElasticNet Regression	0.914716	1.576389
Extra Trees	0.976772	1.594153
Gradient Boosted	1.193252	2.016166
Baseline	3.734177	4.848333



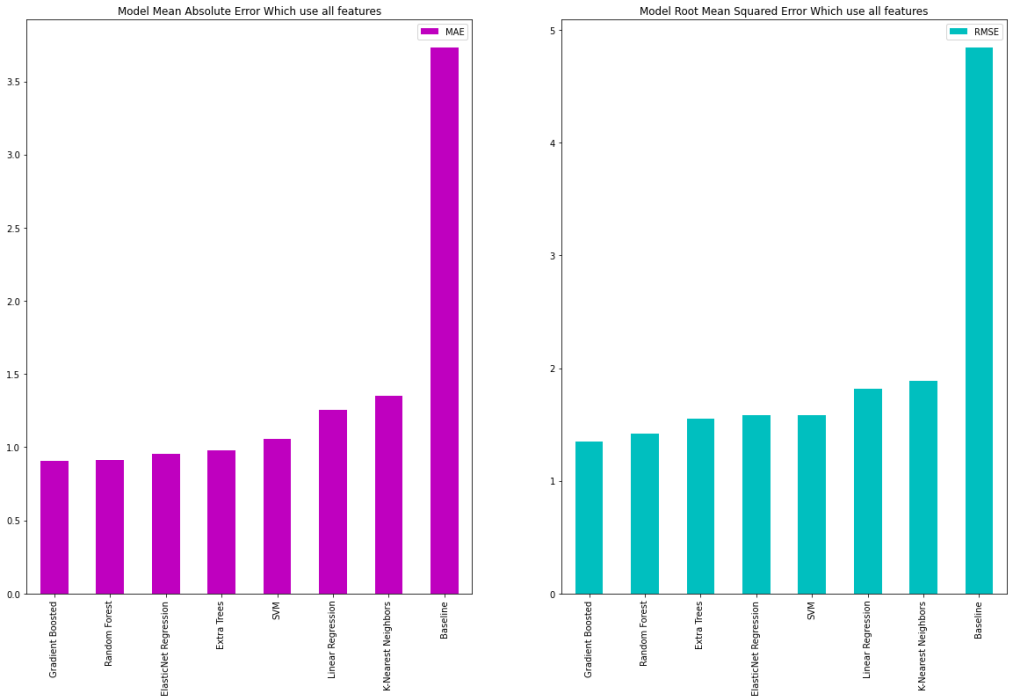


第三个模型为使用全部特征训练模型，下两张图表示了此模型的预测结果，可以看出梯度提升在使用全部特征训练的条件下的预测结果是最好的，而KNN模型预测结果最差。

在这里就能明显看出决策树算法的优势，越多特征值反而是增加了他们的准确度。

Using all features :

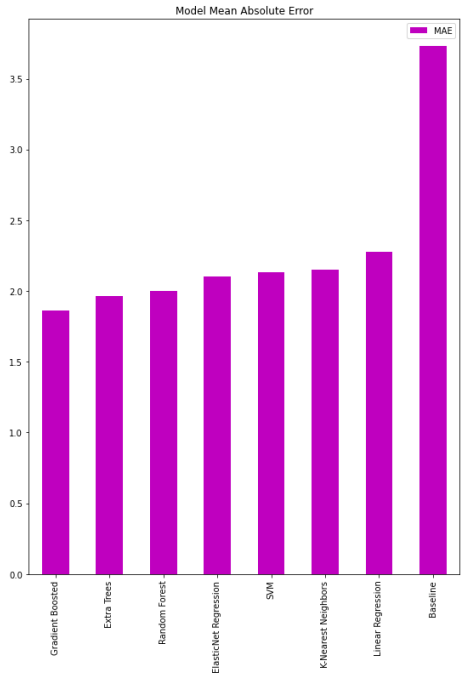
	MAE	RMSE
Linear Regression	1.252847	1.813371
Random Forest	0.914237	1.415721
SVM	1.054314	1.583814
K-Nearest Neighbors	1.352483	1.887658
ElasticNet Regression	0.953365	1.578778
Extra Trees	0.978027	1.549028
Gradient Boosted	0.909187	1.346797
Baseline	3.734177	4.848333

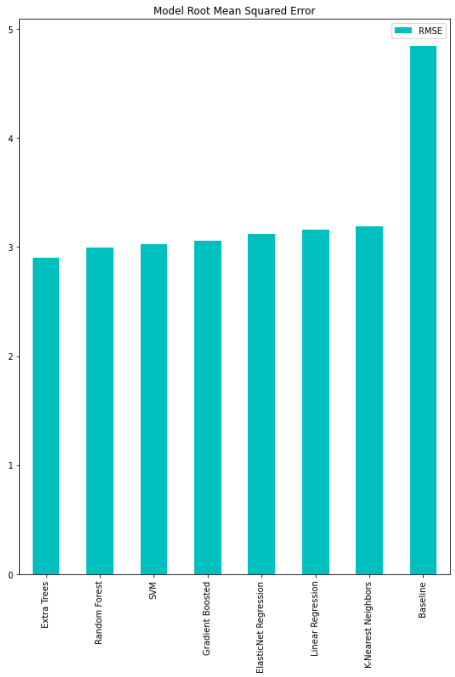


第四个模型为删除掉”G2”特征后再进行建模，由下图的数据明显看出预测结果的准确度突然就差了两倍，由此可见”G2”特征对于前面的预测模型的重要性。

Drop G2 :

	MAE	RMSE
Linear Regression	2.277743	3.160424
Random Forest	2.000725	2.991932
SVM	2.131389	3.024241
K-Nearest Neighbors	2.148978	3.188167
ElasticNet Regression	2.104269	3.117224
Extra Trees	1.962826	2.905042
Gradient Boosted	1.895278	3.066728
Baseline	3.734177	4.848333

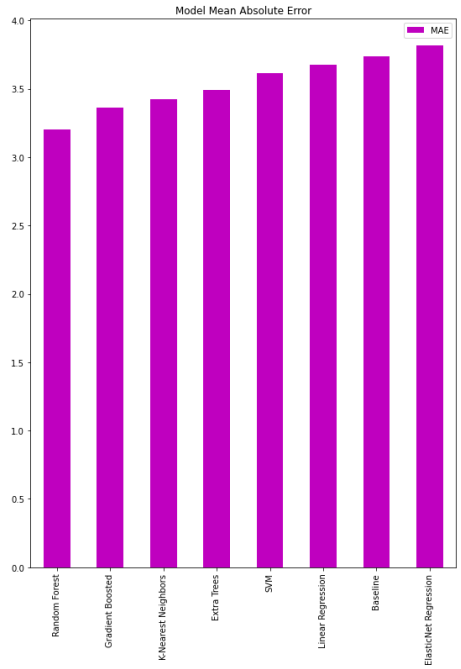


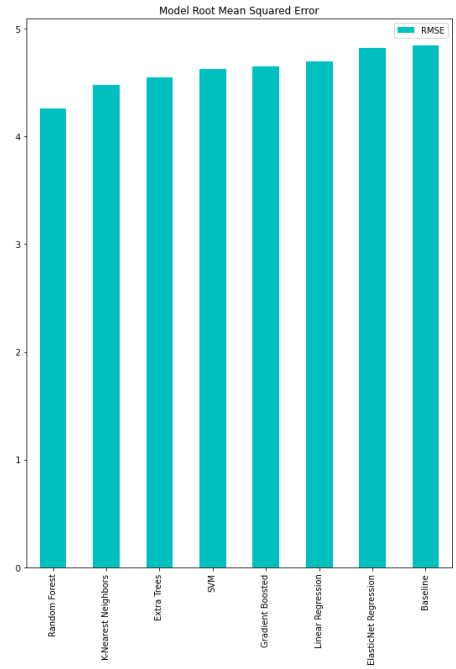


第五个模型为删除掉”G2”、”G1”两个特征后建模，可以看出模型的准确度又再一次降低，甚至有低于 baseline 的模型，如 ElasticNet Regression，此模型在第一次参数未设定时的预测结果是最好的，而在此却是最差的，这个结果令人诧异。

Drop G1 :

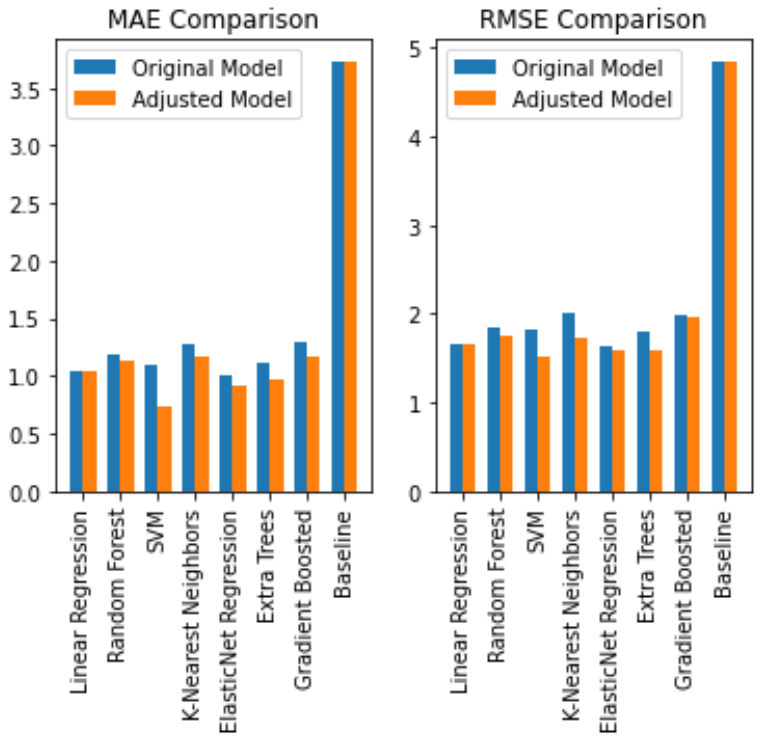
	MAE	RMSE
Linear Regression	3.674465	4.694262
Random Forest	3.204545	4.261656
SVM	3.613903	4.625048
K-Nearest Neighbors	3.420643	4.47827
ElasticNet Regression	3.819969	4.825123
Extra Trees	3.490741	4.550244
Gradient Boosted	3.36165	4.637701
Baseline	3.734177	4.848333





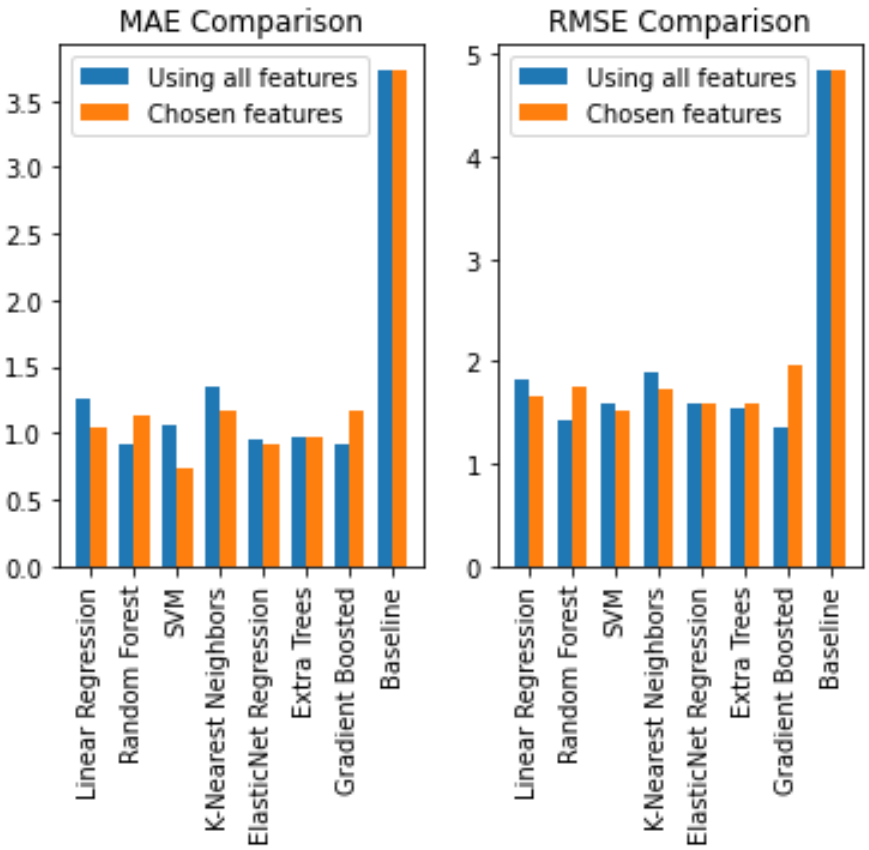
比对一

此图是比对原始参数模型的预测结果和调整参数后模型的预测结果，可以下图的比对图表看出，调参过后的结果优于原始模型预设参数的，尤其是 SVM 的预测结果差别很大，MSE 指标就下降了 0.367。



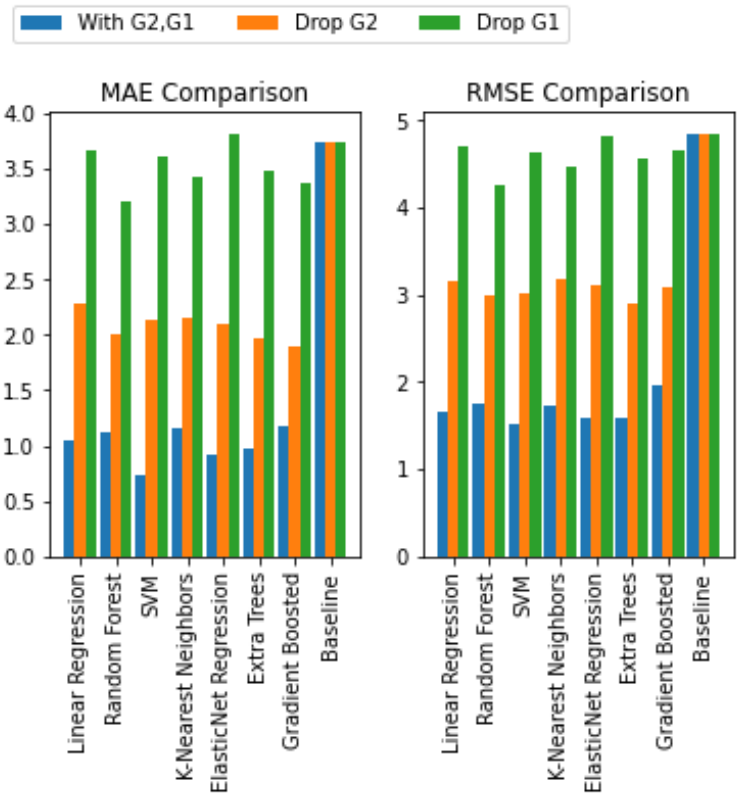
比对二

由图表可以看出有些模型使用了全部特征的预测结果比挑选特征的预测结果还要好，预测结果明显更好的有 Random Forest、Gradient Boosted，预测结果差不多的有 Extra Trees、ElasticNet Regression，以 MAE 指标下去看的话预测结果并没有超过 SVM 模型，但以 RMSE 指标去看的话，预测结果是有优于 SVM 模型的。



比对三

比对删除特征”G2”、删除特征”G1”、”G2”与未删除任何特征的预测结果来看，这个对比非常的明显，基本上删除了特征后，准确度以倍数在下降，这更可以看出”G1”、”G2”对”G3”的重要性，也就是代表说，学期一开始的成绩就非常重要，这会影响到自己最后的成绩。



第五章　　结论

此研究发现父母在很大程度上会影响学生的成绩，尤其是父母的教育程度。另外令人意外的是其实学生的读书时间或是在课业上的努力影响度反而没有想象中的高，影响高的反而是外在的因素，除了父母以外如学生的升学意愿、对象的有无、年龄、通勤时间，这就表示了其实学生除了读书外，应该也要多在意身边的。

由此研究也显示了，从一年的考核中，前面表现出成绩高的学生，到最后成绩的时后基本成绩也会高，所以表达学生该从头到尾都努力学习，不该在最后的一次考试才用功读书，临时的抱佛脚是无法拟补之前的缺失，并且会落后别人许多。

经由训练回归模型，查看预测结果并进行比对后，了解了几个模型的特点，由结果可以发现，当使用全特征值之后使用随机森林算法的模型准确度就明显提升很多，而像线性回归模型就会被过多的特征干扰，由此可以归纳出几个回归模型的特点，如线性回归和 SVM 模型和 KNN 模型是容易欠拟合的，适合在特征相关度较高的数据集中使用，这几个模型对离群值较敏感，而随机森林和梯度提升模型是容易过拟合，适合在特征较多、相关度较低的数据集中使用，弹性网络和极限树就比较没有限制的数据偏向，是大部分的数据集都可使用的模型。

虽说此次研究有更深入学习回归模型的原理和算法，但可能没有完全发挥所有模型的功能，有些模型可能适合分类的，所以之后有机会希望可以尝试做分类模型的分析与比较与这些模型各自适用的场景。

参考文献

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3－42.

数据集来源：
https://archive.ics.uci.edu/ml/datasets/student+performance