PREDICTING PERFORMANCE OF HIGHER EDUCATION INSTITUTES WITH PATTERNS OF EXPENDITURE

Capstone Project: Machine Learning Engineer Nanodegree

Wei-Chuang Chan

I. DEFINITION

PROJECT OVERVIEW

As post-secondary education becomes a common experience of the commonwealth, the budget of an education institution is often limited and funding is competitive. While working closely with many colleges and students, knowing the expectation of their investment is crucial for both. The project aims to figure out whether the allocation of expenditure can help predicting the performance of institution - which is defined by the number of awards, degree, and certificate granted - as it has been widely used to evaluate the performance of higher education institutes.

This project is inspired by Udacity's capstone project guidance and uses dataset downloaded from Integrated Postsecondary Education Data System Delta Cost Project Database, which include the data from the academy year 1987-1988 to 2012-2013. The Delta Cost Project Database is derived from the Integrated Postsecondary Education Data System (IPEDS) surveys on finance, enrollment, staffing, completions, and student aid, and the data have been translated into analytical formats to allow for longitudinal analyses of trends in postsecondary education with a focus on revenues and expenditure.

PROBLEM STATEMENT

The question asked with this project is whether expenditure can be used to predict completion number and if expenditure pattern can make a better predictor for the same target. Expenditure categories will be retrieved from the dataset acquired through delta cost project database, The goal of this project is to identify if the expenditure pattern will be a better reference to predict high completion number than original data.

Dealing with this problem, the target variable and the expenditure variables will be extracted from the dataset. 75% of the randomly selected dataset will be used as the training set, while the rest will be the testing set. The pattern of expenditure has to be found through expenditure variables in the training set. The performance of the model trained by original data will be compared with the performance of the model trained by expenditure pattern.

Multiple models will be compared and the best model will be further tuned to achieve a better result. Metrics for performance evaluation will be discussed below.

METRICS

The models being evaluated will be regression models, and the estimate of error between predicted values and actual value would be the reference of performance. The options for regression models include Mean Absolute Error (MAE), Root of Mean Squared Error (RMSE), and R-squared score (R2).

Mean Absolute Error calculates the average of the difference between predicted value and actual value.

Root Mean Squared Error takes the root of the average squared distance between predicted value and actual value. Compare with MAE, RMSE weighs more on the error that is further away from the mean. Both MAE and RMSE are negatively oriented – meaning the less the error the higher the accuracy is. When MAE equals to RMSE, it means every error are of the same magnitude.

R-squared calculates the variance of the true dataset and calculates the proportion that the predicted data can be accountable for. The residual of the variance indicates the variance caused by the difference between actual data and predicted data. The much variance being explained by predicted data the higher the score is, which also indicates higher accuracy. R2, however, inflates when adding more predictors (variables) to the metric. The variance caused by predicted value hence increase even without model improvement. Adjusted R-squared score (Adj R2) is developed to counter the inflation and adding a penalty for extra variables entering the metric. Adj R2 is always smaller or equals to R2 score. Predicted R2 is another score that helps to determine if a model fits the original data but less capable of providing valid predictions on new data points.

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are also popular tools to evaluate the information loss of a model. The computation of AIC and BIC is complicated, but the concept is based on maximum likelihood estimate of the model parameters, which is the estimate of parameters that gives the highest probability. Log of the value of maximum likelihood function is taken, which ranges from negative infinity to 0. Negative 2 will multiply the value of log. For AIC, this will add 2 times of the total number of parameters. For BIC, the number of the parameters will multiply the log of the number of observations, which is the number of data points in our case. AIC and BIC are all numbers for comparison and definite value cannot be interpreted. For both criteria, the model with a smaller number of the result will be preferred. AIC penalize unnecessary parameters less heavily than BIC, and BIC should be always smaller than AIC when evaluating the same model.

When comparing MAE and RMSE, RMSE puts more weights on the larger error, and MAE behaves less sensitive to outliers. When comparing the model of original data and the model of the transformed dataset, it is likely the dimensionality changes. To avoid the inflation affects the R2 score should not be used as the metric of evaluation but Adj R2 can be used instead. While R2 ranges from 0 to 1, Adj R2 can be negative, which makes it more complicated than R2 to be understood. This feature makes Adj R2 rather easy to understand and straightforward. AIC and BIC are well-known tools while evaluating model performance, and both have the feature that penalizes over-fitting. This makes AIC and BIC the most stands out metrics among above. AIC and BIC provide a relative reference of model performance, but no absolute reference that determines whether the model predicts accurately or not. To add an extra reference, Adj R2 will be used.

II. ANALYSIS

DATA EXPLORATION

The dataset retrieved from IPDES consists of revenue and expenditure of institutes across the country. By selecting data points that contain total completions, the total number of data is 148261. Given the features chosen are numerical, the dataset contains a significant amount of null data - which is likely missing data or unreported.

For the data subset took into this project, a brief explanation of each feature will be provided. The expenditure on salary and wage, total expense reported by FASB, and all subcategories were excluded.

academicyear - the academic year of data reported.

totalcompletions - the total number of degree, award, certificate granted of the current year. This will be the indicator of performance in this project.

has_completion - Indicator of whether totalcompletions has been reported. (0=not reported; 1=reported)

instruction01 - instructional expenses for the institution and excludes administration, operations and maintenance.

research01 - expense used to produce research outcomes excluding operation and maintenance, interest amounts attributed to the research functions.

rschpub01 - expense for research and public service of current year

pubserv01 - expense category in order to provide noninstructional services beneficial to individuals and groups external to the institution such as conferences. Operations and maintenance, interest amounts attributed to the research function are excluded.

acadsupp01 - expenses to support instruction, research, and public service. This category includes retention, preservation, and display of education materials. Operations and maintenance and interest amounts attributed to the academic support function have been excluded.

studserv01 - expenses associate with admissions, registrar activities, and activities that contribute to students' emotional and physical well-being and to their intellectual, cultural, and social development outside the formal instructional program. Operations and maintenance (and interest in the 2009 aligned form) amounts attributed to the student services function have been subtracted.

instsupp01 - expense for day-to-day operational support of the institution such as space management, employee personnel, and records. Operations and maintenance and interest amounts attributed to the institutional support are excluded.

acadinststud01 - academic and institutional support and student service total of current year

opermain01 - expenses for operations providing service and maintenance related to campus grounds and facilities. Institutions may optionally distribute depreciation expense to this function.

depreciation01 - total depreciation of current year

grants01 - the sum of all operating expenses associated with scholarships and fellowships including payment in support of the cost of education or third parties for off-campus housing. Operations and maintenance and interest amounts are excluded.

auxiliary01 - expense of all operating associated with essentially self-supporting operations of the institution such as student health services. The amount of interest is excluded.

hospital01 - operating expenses associated with a hospital operated by the postsecondary institute.

independ01 - expenses associated with operations that are independent of or unrelated to the primary missions of the institution. The amount of interest attributed to the independent operations function is excluded as well as for the expenses of operations owned and managed as investments of the institution's endowment funds.

otheroper01 - All expense other than categories above which is completely discontinued after the academy year 2010.

othernon01 - All other non-operating expense of current year

other01 - All other expense

Data points that contain totalcompletions value will be included for analysis due to the role of this feature being reference of performance.

The dataset contains a significant amount of missing values, the feature contains more than 95% of missing values will be removed, and data points do not contain the value for each remaining features will also be removed to avoid bias. The procedure will be taken in order to keep sufficient amount of data while removing data points with missing value first will result in less than 5% usable data.

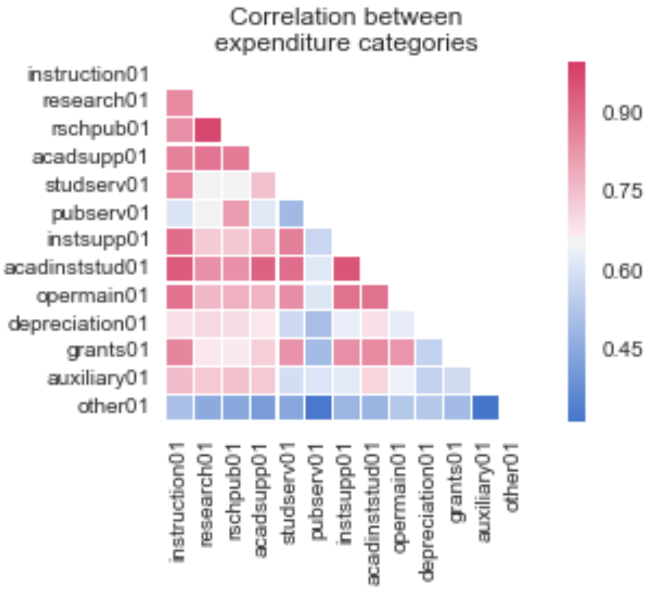


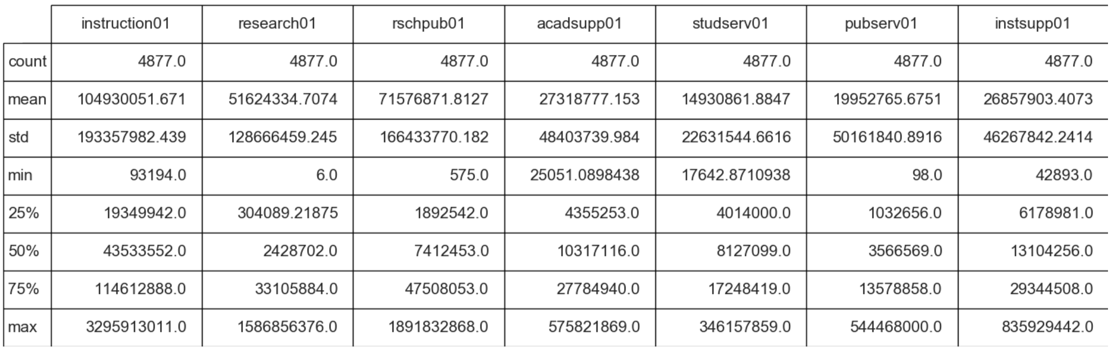
fig 1. Correlation plot between variables

There are high correlations between some variables; instruction01 seems to correlate with most variables but other01, which has low correlation with all other variables. In this cast, PCA can help setting new dimensions that explained shared variance.

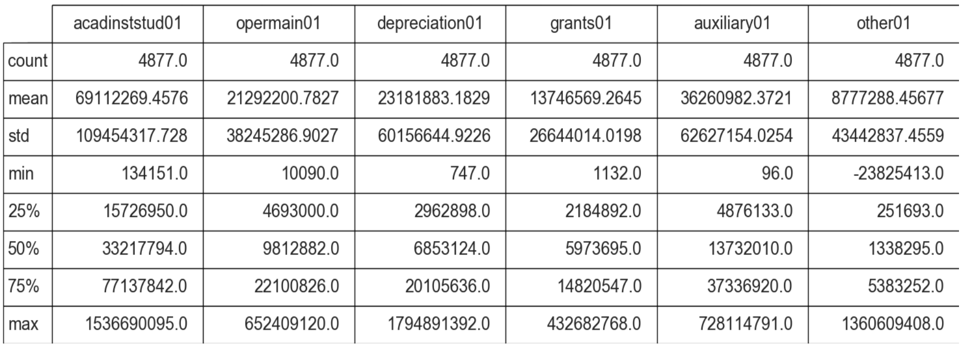
According to the heat map, instruction01 appears to have a high correlation with other features, and the other expenditure has low correlation with any other features. Features including depreciation01 and pubserv01 have no significant correlation with most other feature in general, but pubserv01 has more than 0.5 of correlation with research01 and rschpub01. Features studserv01, grants01, and auxiliary01 have a high correlation with around half of the other features but also show low correlation with the other half of the features.

As the plot shows, instruction is highly correlated with nearly all kinds of expenditure, and so does research expense. While other01, pubserv01, and depreciateion01 could be the most distinguishable features among all.

The statistics of each feature are listed below:



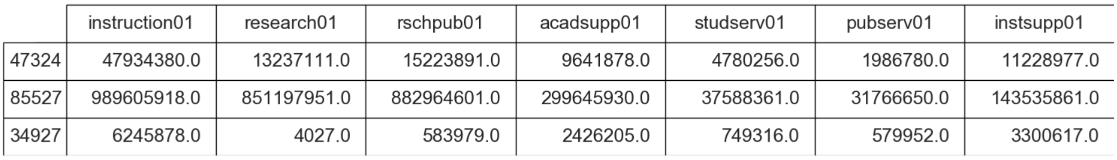
	instruction01	research01	rschpub01	acadsupp01	studserv01	pubserv01	instsupp01
count	4877.0	4877.0	4877.0	4877.0	4877.0	4877.0	4877.0
mean	104930051.671	51624334.7074	71576871.8127	27318777.153	14930861.8847	19952765.6751	26857903.4073
std	193357982.439	128666459.245	166433770.182	48403739.984	22631544.6616	50161840.8916	46267842.2414
min	93194.0	6.0	575.0	25051.0898438	17642.8710938	98.0	42893.0
25%	19349942.0	304089.21875	1892542.0	4355253.0	4014000.0	1032656.0	6178981.0
50%	43533552.0	2428702.0	7412453.0	10317116.0	8127099.0	3566569.0	13104256.0
75%	114612888.0	33105884.0	47508053.0	27784940.0	17248419.0	13578858.0	29344508.0
max	3295913011.0	1586856376.0	1891832868.0	575821869.0	346157859.0	544468000.0	835929442.0



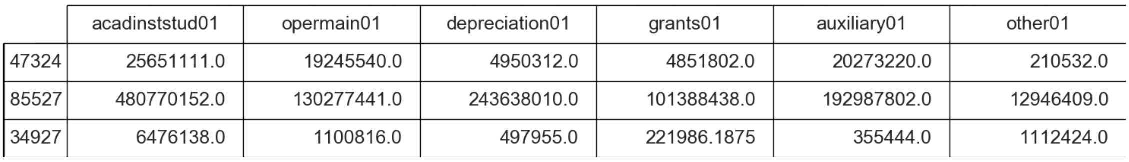
	acadinststud01	opermain01	depreciation01	grants01	auxiliary01	other01
count	4877.0	4877.0	4877.0	4877.0	4877.0	4877.0
mean	69112269.4576	21292200.7827	23181883.1829	13746569.2645	36260982.3721	8777288.45677
std	109454317.728	38245286.9027	60156644.9226	26644014.0198	62627154.0254	43442837.4559
min	134151.0	10090.0	747.0	1132.0	96.0	-23825413.0
25%	15726950.0	4693000.0	2962898.0	2184892.0	4876133.0	251693.0
50%	33217794.0	9812882.0	6853124.0	5973695.0	13732010.0	1338295.0
75%	77137842.0	22100826.0	20105636.0	14820547.0	37336920.0	5383252.0
max	1536690095.0	652409120.0	1794891392.0	432682768.0	728114791.0	1360609408.0

Table 1 Statistics of features

The first thing to be noticed is that huge range within each feature, and large standard deviation indicates the distribution is spread out. Further calculating the skewness and kurtosis of each feature, the distributions are found to be positively skewed with high kurtosis.



	instruction01	research01	rschpub01	acadsupp01	studserv01	pubserv01	instsupp01
47324	47934380.0	13237111.0	15223891.0	9641878.0	4780256.0	1986780.0	11228977.0
85527	989605918.0	851197951.0	882964601.0	299645930.0	37588361.0	31766650.0	143535861.0
34927	6245878.0	4027.0	583979.0	2426205.0	749316.0	579952.0	3300617.0



	acadinststud01	opermain01	depreciation01	grants01	auxiliary01	other01
47324	25651111.0	19245540.0	4950312.0	4851802.0	20273220.0	210532.0
85527	480770152.0	130277441.0	243638010.0	101388438.0	192987802.0	12946409.0
34927	6476138.0	1100816.0	497955.0	221986.1875	355444.0	1112424.0

Table 2 Sample Data

From the first 3 data point in the training set we found that:

47324 is between the 50th percentile and the 75th percentile on instruction, between the 25th and the 50th percentile, and relatively small number of all features, 85527 has relatively large value on most features. 34927 is at the end of the lower side, but it spends more on grants and 47324.

Algorithms and Techniques

The tasks of this project include: Finding expenditure patterns and training 2 separate regression predicting models with the original dataset and the transformed dataset.

The dataset will be processed to eliminate data points that are not helping the training process: data that misses target value, outliers that heavily influence the first randomly split into a training set and a testing set at 0.75:0.25 ratio. A copy of training set will be made, and the predicting features of a copy will go through Principle Component Analysis.

Principle Component Analysis will be implemented to find the dimensions that explained the most of the variance. The captured principle components will be used as predicting features for comparison. To find the best model to be trained with transformed data, multiple models will be compared with the metric chosen.

Acquired dimensions will be used as the new feature to predict the target. K-fold cross validation will be used to evaluate the fit of the model. The best model will be tuned through grid search cross validation. The final performance will be evaluated via AIC and BIC.

Benchmark

The aim is to compare the performance of the model on original data and PCA transformed data, and the performance score of the model trained by original data will be the benchmark for the model trained with the transformed dataset.

The benchmark for this project is obtained with a k-nearest neighbor model fitted to original training data. AIC: 16445.95; BIC: 16512.33, Adj R2: 0.955671

III. Methodology

Data Preprocessing

The features are selected for analysis as stated above, among all the features taken. As discovered in data exploration, the dataset contains a significant amount of missing data. The features have more than 95% of missing value will also be removed. To obtain the complete data from the remaining data, any data contains missing value are removed.

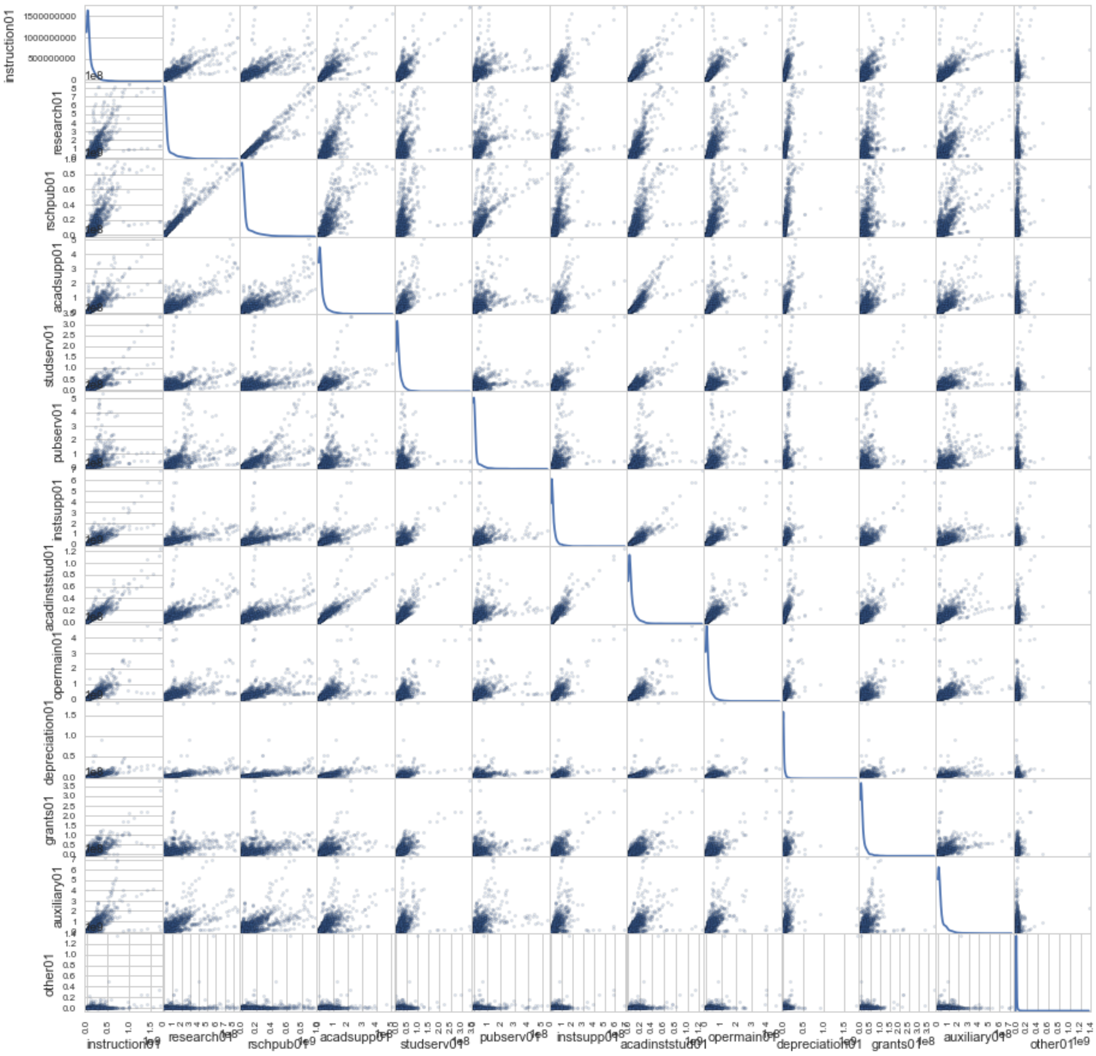


fig 2 Scatter Matrix of features

As described above, the data set is highly skewed and the range of feature is large. The data set should be normalized prior to PCA to avoid too much weight being put on the feature with higher variance. The outliers of the data set will not be dropped in order to keep most of the information.

IMPLEMENTATION

The data set will be loaded with python pandas library as a data frame. After removing useless feature and data points, the training set and the testing set were randomly selected at 3:1 ration.

The metrics chosen were written as functions calculated through basic formula based on the residual sum of squared error and built-in R2 score function of scikit learn's base model.

Training data set will be standardized before PCA is implemented, and transformed data set will be used as new features to predict the target.

Candidate models are compared by the mean of metrics calculated by k-fold cross-validation, which is set to 10 folds here. Models to be compared are retrieved from scikit-learn's library, which is a widely used library that contains various commonly used base models. Among all the selection, the following models are used for comparison and the reason of selection follows:

K-Nearest Neighbor Regressor: The model estimates value by the closest data points, and the model costs less with minimal tuning.

Decision Tree Regressor: The model is easy to implement and also performs very fast. Once trained the model predicts very quick. The model gives a clear structure of how it makes prediction, but the downside is the prediction may lack interpretation.

Support Vector Regressor: The model is also a commonly used model. SVR has strong founding theory, less prone to over-fitting, and need less tuning

Multiple Layer Perceptron Regressor: The model uses the structure of neural network and is regaining popularity due to vast applications on voice and image recognition. The model consists 3 or multiple layers and each layer contains perceptrons which follow their own activation functions to decide whether to send an output or not. The neural network model is easier to conceptualize and has huge amount of related research, but to outperform other modern models it requires more tuning and usually harder to train. However, it can be used to tackle a lot of problem without too much understanding.

The following two algorithms use multiple trees and combine the outputs.

Random Forest Regressor: The model creates multiple decision trees that are trained with data through bootstrap sampling. Every node on the branch will randomly choose a small amount of the features. The trees will be tested on the data not sampled. This method can avoid over-fitting to training data by its randomness.

Gradient Boosting Regressor: The model builds decision trees sequentially to make a strong model. It uses later model to correct the previous model. The process can take longer than Random Forest, but the result shows better performance than Random Forest. On the contrast, Gradient Boosting is more prone to over-fitting.

The model performance will be evaluated with the metrics chosen combining with k-fold cross validation within training set. The model with the best performance will be tuned with grid search.

REFINEMENT

The initial performance of models are reported as following:

	Decision Tree	Support Vector Machine	Multi-Layer Perceptron	K Nearest Neighbor	Random Forest	Gradient Boosting
AIC	5251.27	6050.15	5533.21	4871.94	5075.76	5024.20
BIC	5298.09	6096.97	5580.03	4918.76	5122.58	5071.02
Adj R2	0.8731	-0.03	0.7402	0.9573	0.9208	0.9357
Time	0.5293	5.8469	18.6735	0.3771	3.0982	3.2151

K-nearest neighbor regressor stands out on all metric and will be tuned with grid search.

After grid search, the best parameters are set. The final model will be using 1 closest neighbor for estimates. Performance is as following:

AIC: 15405.51

BIC: 15466.80

Adj R2: 0.9811

To further enhance the performance, I tried applying AdaBoost to the model. AdaBoost further reduces the dimensionality of data set and more frequently used with weak learner. The result improves, but the time cost is relatively high – fitting time is nearly 1000 times the time cost of final model and predicting time is 55 times of the final model.

AIC: 15351.16
BIC: 15412.44
Adj R2:0.9819

IV. Results

Model Evaluation and Validation

With all variables retrieved from original data set, 13 were kept for unsupervised learning. PCA reduced the dimensionality to 12 dimensions. Using data transformed by PCA as predicting features, multiple models are compared together. K-Nearest Neighbor has the best performance, which has AIC and BIC more than 4% better than other candidates and hence chosen to be fine-tuned. The best parameter is set as:

```
   KNeighborsRegressor(algorithm='auto', leaf_size=30,
metric='minkowski', metric_params=None,
n_jobs=1, n_neighbors=1, p=2,
weights='uniform')
```

The tuned model uses uniform weights for all neighbors, and choose the 1 closest neighbor to make an estimate, and metric= minkowski and p=2 meaning the model uses the straight distance between points as metric.

AIC and BIC evaluate the information loss and apply a penalty to unnecessary parameter added. The final model applied the transformation of original data, but the reduction of dimensionality makes it reserves relatively more information with fewer parameters. As reported in Refinement, K-Nearest Neighbor has already been the best performance among all candidate models without tuning. The final model is able to make the prediction without over-fitting while achieving performance over the benchmark.

Justification

The performance compared with the benchmark, the model's AIC is 8.02% lower, BIC is 8.02% lower, and Adjusted R2 score is 3.86% higher.

According to Bayes Factors (Kass, Raftery, 1995), the strength of the evidence against the models with higher BIC is considered strong when **ΔBIC** is between 6 to 10. The final KNN model appears to have significant difference against the benchmark, and the adjusted R2 score is also better than the benchmark.

V. Conclusion

Free-Form Visualization

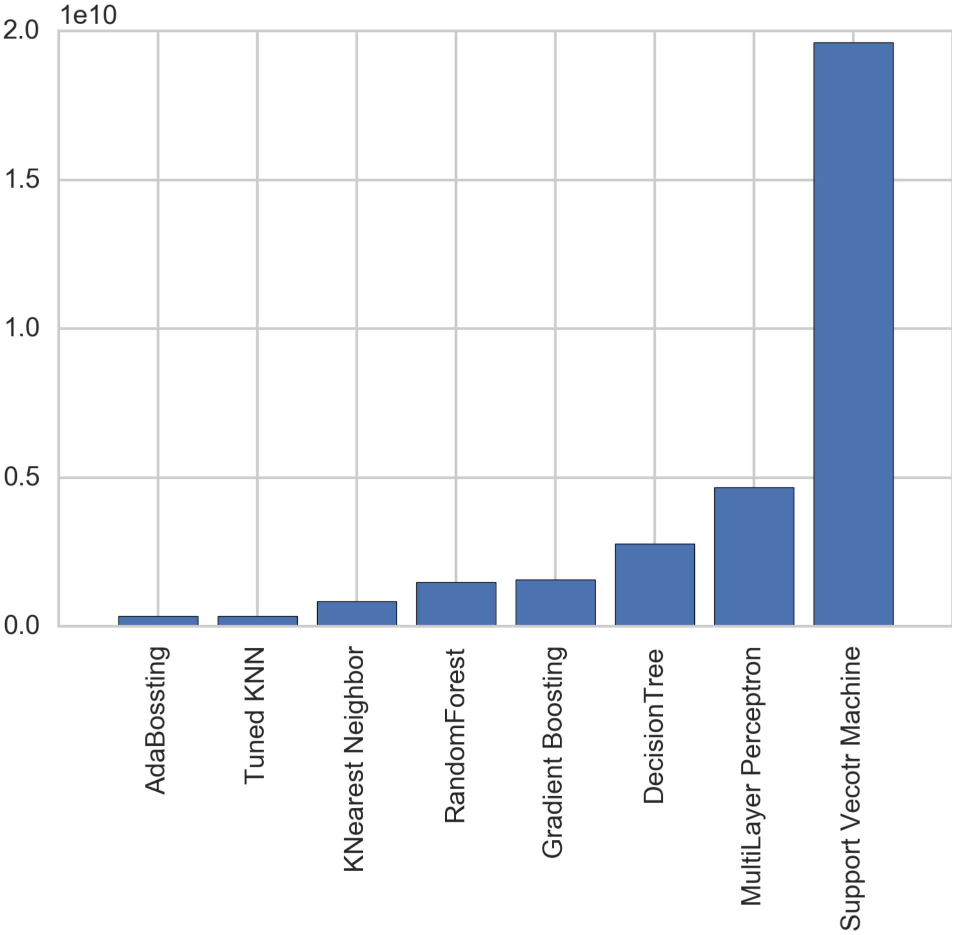


fig 3. Residual Sum of Square Comparison between models

The bar chart shows each model and the sum of squared difference between prediction and actual completion number.

From the bar chart, it can be seen that AdaBoost tuned KNN, but the difference is not as significant as between the 2 models and other models. KNN leads the ranking and followed by ensemble methods, and SVM has far more error than other candidates.

Reflection

This project targets to find the principle components of the expenditure of higher education institutes and used them as a better predictor of performance, which is defined as a total number of degree, certificate, and award completion. The project successfully to find the expenditure combination that serves as better predictors with principle component analysis and creates a model that performs better than the benchmark.

Selecting the features for analysis is the first obstacle in this project. While some institution runs a hospital and others don't, there are some fields are excluded for not relevant to all dataset. Although the project eventually excluded data points with missing data, imputation is an alternative that has been considered. However, exclusion was chosen to avoid changing the statistics of dataset. Choosing candidate models is also a struggling, while most popular models are selected, there are also several models that were widely used listed solely for comparison. Gradient Boosting Decision Trees and Random Forest are both well-known ensemble method, but the characters of these 2 models might not be the best option while dealing with this dataset.

Delta Cost Project provides a dataset with quite a good shape, and simple analysis can bring a lot of information. During the contemplation of this project, there are several other ideas that use the same dataset and take not much time to be carried out. The project itself, even if performance evaluation is passed, also has room for improvement, which will be discussed in the later section.

IMPROVEMENT

While predicting the performance of an institute, rather than using expenditure of current year, previous years should be also taken into consideration for any degree, certificate, awards can take longer than a year to be completed. On the execution part, this project may need a comparison with the version includes all sub-categories of expenditure. Missing value treatment can also be reviewed again in this case.

PCA successfully reduced dimensionality and provides the principle components. With further clustering, we might be able to categorize numerous education institutes into groups and find the high-perform expenditure pattern for each group. Institute can use the result to find out which group they are in or want to identify them, and using the expenditure pattern for future investment reference.

Models comparison includes some models that require tuning to perform better. For a model collection including those models, it might be better to compare after proper tuning. Initial plan also includes xgboost model, but it is removed after research on the dataset and the advantage of models.

In the process of this project, coding becomes a constraint and costs much time. With more practice, the code can be cleaner.

Reference:

http://www.ncsl.org/research/education/performance-funding.aspx

http://www.ed.gov/college-completion

https://www.researchgate.net/post/What_is_the_AIC_formula

http://www.stat.wisc.edu/courses/st333-larget/aic.pdf

https://en.wikipedia.org/wiki/Bayesian_information_criterion

https://en.wikipedia.org/wiki/Akaike_information_criterion

http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/