Data Mining Project

Seed: 934672
Partition: 70/15/15

2.Dataset link:
https://data.iowa.gov/Local-Government-Finance/City-Budget-and-Actual-Expenditures/jy6h-2e5x

Executive Summary:
The aim of this project is to determine the accuracy of government budget estimates within the state of Iowa over a period of 11 years. The data was used from data.iowa.gov shows budget and actual expenditures by each county based on their yearly budget forms.

Business Goal:
Determining the accuracy of government budget estimates within the state of Iowa based on budget data from each county over a 11 year period. Specifically, evaluating if there are certain parts of the total budget that have a greater impact on the total expenditure budget for the year. Estimating accurate expenditure can help Iowa's government better allocate those extra funds to underfunded resources or underdeveloped counties which normally don't get much funding.

Data Mining Goal:
The target variable in our data set is the "total actual" that specifies the total actual expenditures from the counties during the fiscal year. The data mining goal is to build a supervised learning model to understand how the different budget expenditure types affect the total actual expenditure.

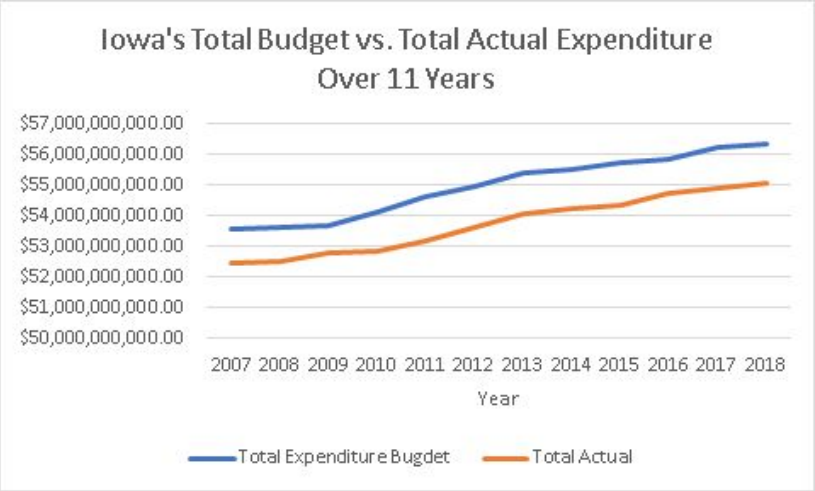
Data Preparation: First, the Iowa counties were categorized into 6 regions using Iowa Homeland Security and Emergency Management's website. The reason for this grouping was to reduce the amount of levels upon uploading the csv into rattle. Upon regrouping, we saw some counties have had more budget predictions in a specific year than others. We could not find a reason for that and thus left the extra rows in the data set. Then we deleted all data from the year 2019 since the actual amount spent by counties had not been uploaded on the csv. After the cleaning, we had a data set of Iowa counties split into 6 regions between the years of 2007 to 2018. The columns in the data set included the year, county name, region, all of the budget columns, and total actual (target variable). Source for how and which of the counties are split into the 6 regions
https://www.homelandsecurity.iowa.gov/county_em/county_em_overview.html

Exploratory Analysis:

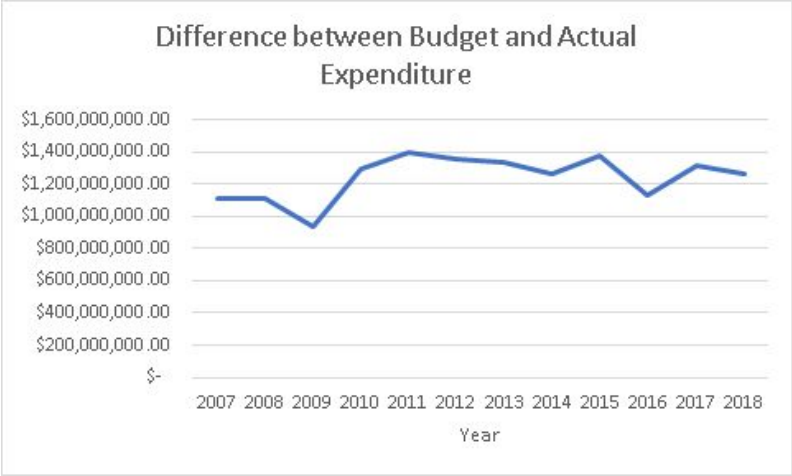
The aim of this project is to determine just how accurate the estimations are for the government budget within the state of Iowa over a period of 12 years. The models were used to conduct further analysis on whether or not the estimations were accurate. The dataset used in this analysis is from data.iowa.gov and consists of over 11.3k rows and 30 columns. The counties were grouped into 6 regions based on the homeland security planning regions (homelandsecurity.iowa.gov). The models that were used for the analysis of this project were: Linear Regression, Decision Tree, Neural Networks and Random Forests. All models were partitioned and executed on rattle.

Overview of Spending

Upon making a line chart of the data we saw a trend which shows a steady increase in total budget and total actual expenditure. We can also see that during 2009 and 2016 the Total Budget was closer to the Total Actual Expenditure. This could mean that there was a possible unexpected cost which arose during those years. On the other hand, if we look at other years there seems to be consistency in the difference between the total estimated budget and total actual expenditure.To analyze the difference between the estimated and actual expenditure let's look at the next graph.

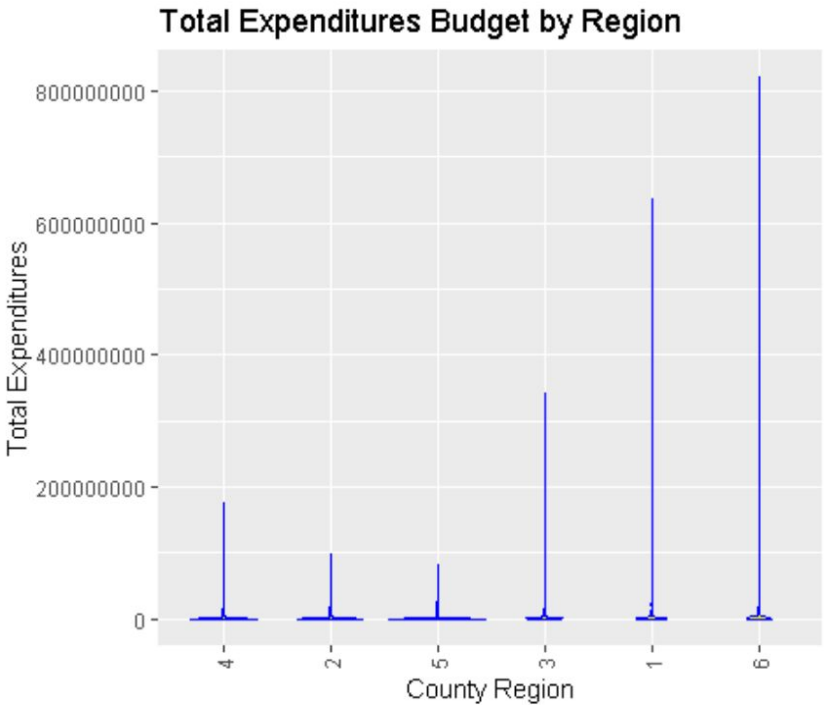


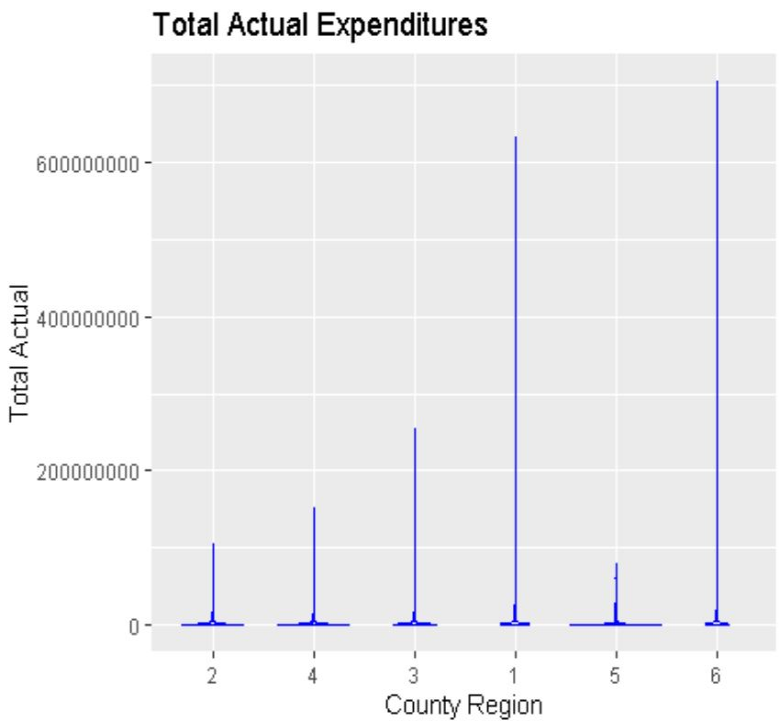
In this graph, we can clearly observe the pattern between the estimated total budget for each year and the actual total expenditure. As the line indicates, each year there has been an overestimation of the actual expenditure between the range of $800,000,000 to $1,400,000,000. Fortunately, we believe these differences in estimates can be lowered by using predictive modeling.

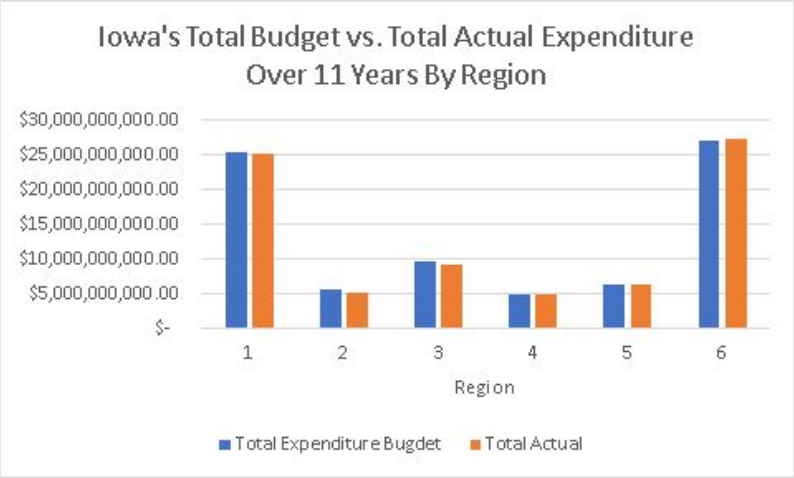


County Regions With Highest Total Expenditures and Budgets

As you can see on the graph below, the county region with the highest Expenditure Budget over the 12 year period is County Region number 6. This was also consistent with the Total Actual Spending as county region 6 also had the highest total actual expenditure. Based on the graphs below, out of the 6 regions, region 2 is the only county region that has exceeded their total expenditure budget over the past twelve year. While the other county spending on average has been equal to or below their expenditure budget.







Modeling & Evaluation

ANN

For the Neural Network, I transformed the categorical variables into numeric and then I imputed the missing values. After that, I rescaled the numeric variables into a [1-0] scale. I tested five different combinations with different Hidden Layer Nodes and used the MAE. As you can see below, four Hidden Layer Nodes produced the best MAE which was .002226. I then ran the best combination which was 4 on the testing set and it produced a MAE of .00235.

I also tested five different combinations with different Hidden Layer Nodes using the Pseudo R Squared. As you can see below, out of the six combinations, the highest Pseudo R Squared was 98% which was at 4 Hidden Layer Nodes. I then ran the best combination which was 4 Hidden Layer Nodes on the testing set and received a Pseudo R Squared of 98.28%.

Hidden Layer Nodes	MAE
3	0.002333
4	0.002226
5	0.002284
7	0.002956
10	0.002699
15	0.002456

ANN

Hidden Layer Nodes	Pseudo
3	0.9677
4	0.98
5	0.9766
7	0.974
10	0.9737
15	0.9731

Random Forest

For our Random Forest, I started by cleaning our data to fit the Random Forest model's preferences, which included imputing our missing values with a median of the data as well as rescaling numeric values to a 0-1 scale. With Random Forest, we have control over both the amount of trees as well as the amount of variables, so I used the outputted Pseudo R^2 value to assess how well the model was performing.

	Random Forest	
Trees	Variables	R^2
500	3	0.9721
1000	3	0.9739
1000	10	0.9658
1500	2	0.973
2000	3	0.9742

	Random Forest	
Trees	Variables	MAE
2000	3	0.002285

In my testing using the validation set, the best combination I could find was using 2000 Trees as well as 3 Variables, where it produced a Pseudo R^2 of .9742. When running this setup with the testing set, it produced an output value of .9826. With the same setup, testing for MAE yielded a result of .002285, proving it again to be a valuable and accurate model.

Decision Tree

Deal with missing values:

- Public Safety Budget had 1 missing value
- The target variable Total Actual had 7 missing values
- I then imputed both variables with the mean because they are numeric variables.

Finding the right sized tree

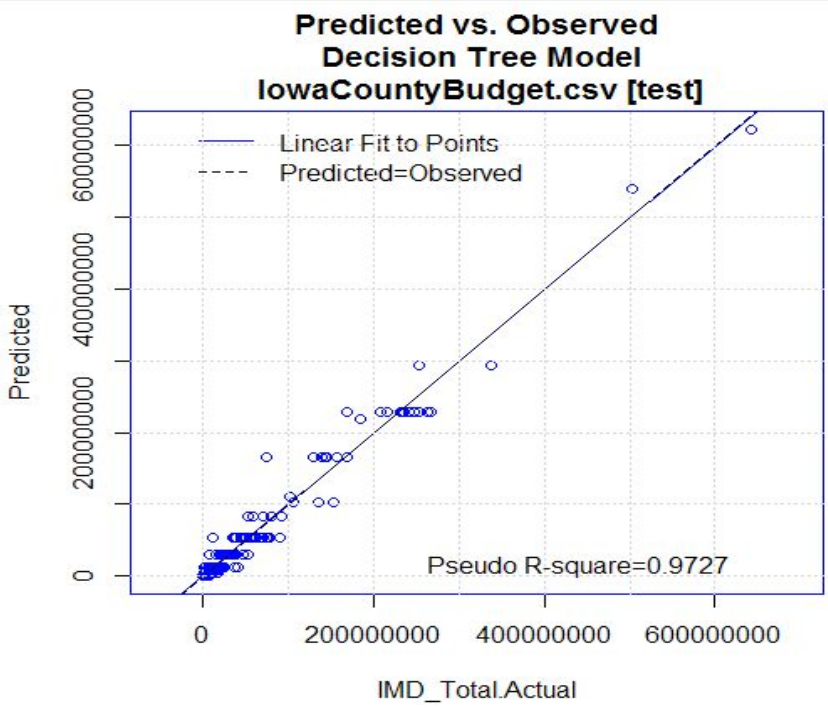
- No pruning:
    - Min split = 0
    - Min bucket = 1
    - Complexity = 0
    - Max depth = 30

  Choosing the best complexity level: look for the smallest xerror (validation error), use the CP associated with this smallest x-error and set the complexity to this number

- Lowest x error =

	CP	nsplit	rel error	xerror	xstd
1	0.685482902	0	1.00000000	1.000129	0.1486974
2	0.130086682	1	0.31451710	0.318364	0.0322812
3	0.107329091	2	0.18443042	0.217639	0.0223659
4	0.022834811	3	0.07710132	0.085705	0.0104320
5	0.013949325	4	0.05426651	0.064397	0.0091689
6	0.006657993	5	0.04031719	0.050347	0.0090226
7	0.003002561	6	0.03365920	0.042640	0.0061926
8	0.002954900	7	0.03065663	0.041789	0.0064727
9	0.002861525	8	0.02770173	0.041488	0.0064638
10	0.002257285	9	0.02484021	0.036976	0.0060231
11	0.001764817	10	0.02258292	0.034184	0.0058655
12	0.001443172	11	0.02081811	0.033300	0.0059237
13	0.001231228	12	0.01937493	0.032515	0.0051572

0.032515, Cp = 0.001231228



MAE	MSE
0.002359	8.37475E-05

From the results above, using a Decision Tree model on our dataset shows that the predicted and the actual are very similar. The MAE follows this conclusion with a very small value of 0.002359. This states that the average error between the predicted and the actual is extremely small in value.

Linear Regression

For Linear regression we had to rescale all numeric variables to zero and 1 then we ignored the county variable and used our county region variable we made to avoid multi level errors.

16	IMN_Public.Safety.Budget	Numeric [0.00 to 109377359.00; unique=7764; mean=745820.52; median=38561.50; ignored].
17	IMN_Total.Actual	Numeric [0.00 to 705214812.00; unique=9410; mean=6876903.61; median=567758.00; ignored].
18	R01_Fiscal.Year	Numeric [0.00 to 1.00; unique=12; mean=0.50; median=0.45].
19	R01_County.Region	Numeric [0.00 to 1.00; unique=6; mean=0.49; median=0.40].
20	R01_Public.Works.Budget	Numeric [0.00 to 1.00; unique=8165; mean=0.01; median=0.00].
21	R01_Health.and.Social.Services.Budget	Numeric [0.00 to 1.00; unique=1182; mean=0.00; median=0.00].
22	R01_Culture.and.Recreation.Budget	Numeric [0.00 to 1.00; unique=7278; mean=0.02; median=0.00].
23	R01_Community.and.Economic.Development.Budget	Numeric [0.00 to 1.00; unique=3633; mean=0.01; median=0.00].
24	R01_General.Government.Budget	Numeric [0.00 to 1.00; unique=8265; mean=0.01; median=0.00].
25	R01_Debt.Service.Budget	Numeric [0.00 to 1.00; unique=4920; mean=0.01; median=0.00].
26	R01_Capital.Projects.Budget	Numeric [0.00 to 1.00; unique=2251; mean=0.00; median=0.00].
27	R01_Business.Type...Enterprise.Budget	Numeric [0.00 to 1.00; unique=7487; mean=0.01; median=0.00].
28	R01_Transfers.Out.Total.Budget	Numeric [0.00 to 1.00; unique=4435; mean=0.01; median=0.00].
29	R01_Total.Expenditures.Budget	Numeric [0.00 to 1.00; unique=10257; mean=0.01; median=0.00].
30	R01_IMN_Public.Safety.Budget	Numeric [0.00 to 1.00; unique=7764; mean=0.01; median=0.00].
31	R01_IMN_Total.Actual	Numeric [0.00 to 1.00; unique=9410; mean=0.01; median=0.00].

We then had to re-assign these new rescaled variables as the inputs except for Total actual since it is our target variable

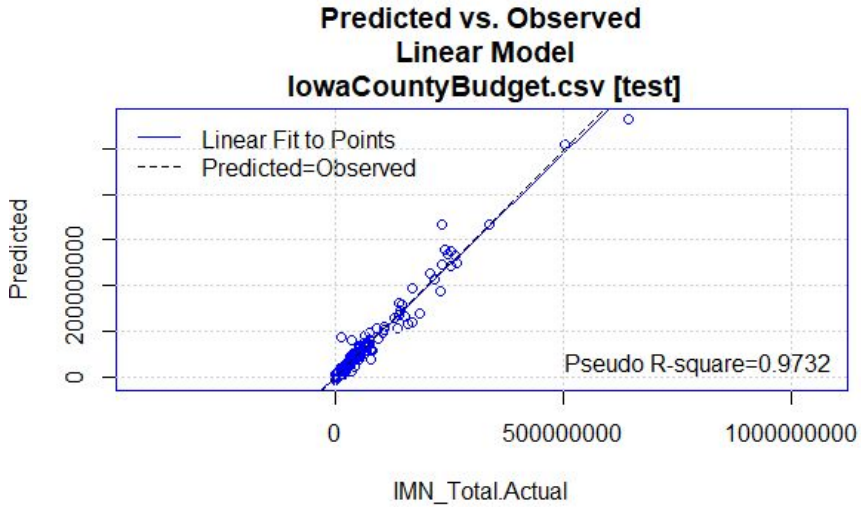
18	R01_Fiscal.Year	Numeric	◉	○	○	○	○	○	Unique: 12
19	R01_County.Region	Numeric	◉	○	○	○	○	○	Unique: 6
20	R01_Public.Works.Budget	Numeric	◉	○	○	○	○	○	Unique: 8,165
21	R01_Health.and.Social.Services.Budget	Numeric	◉	○	○	○	○	○	Unique: 1,182
22	R01_Culture.and.Recreation.Budget	Numeric	◉	○	○	○	○	○	Unique: 7,278
23	R01_Community.and.Economic.Development.Budget	Numeric	◉	○	○	○	○	○	Unique: 3,633
24	R01_General.Government.Budget	Numeric	◉	○	○	○	○	○	Unique: 8,265
25	R01_Debt.Service.Budget	Numeric	◉	○	○	○	○	○	Unique: 4,920
26	R01_Capital.Projects.Budget	Numeric	◉	○	○	○	○	○	Unique: 2,251
27	R01_Business.Type...Enterprise.Budget	Numeric	◉	○	○	○	○	○	Unique: 7,487
28	R01_Transfers.Out.Total.Budget	Numeric	◉	○	○	○	○	○	Unique: 4,435
29	R01_Total.Expenditures.Budget	Numeric	◉	○	○	○	○	○	Unique: 10,257
30	R01_IMN_Public.Safety.Budget	Numeric	◉	○	○	○	○	○	Unique: 7,764
31	R01_IMN_Total.Actual	Numeric	○	◉	○	○	○	○	Unique: 9,410

From the results of running the linear regression we can see that the variables Fiscal Year, County Region, community and economic involvement budget, and debt service budget are statistically insignificant since they are greater than alpha of .05 and therefore have small if any impact on the total actual budget of each county.

	Estimate	Std. Error	t value
(Intercept)	-0.0003447	0.0002180	-1.581
R01_Fiscal.Year	-0.0003853	0.0002971	-1.297
R01_County.Region	0.0008035	0.0002620	3.067
R01_Public.Works.Budget	-0.0539709	0.0074089	-7.285
R01_Health.and.Social.Services.Budget	-0.0226790	0.0081093	-2.797
R01_Culture.and.Recreation.Budget	0.0713456	0.0058175	12.264
R01_Community.and.Economic.Development.Budget	-0.0063663	0.0059246	-1.075
R01_General.Government.Budget	0.1049626	0.0096866	10.836
R01_Debt.Service.Budget	0.0091038	0.0097816	0.931
R01_Capital.Projects.Budget	-0.3795300	0.0355858	-10.665
R01_Business.Type...Enterprise.Budget	-0.1405459	0.0299509	-4.693
R01_Transfers.Out.Total.Budget	-0.2097902	0.0144939	-14.474
R01_Total.Expenditures.Budget	1.4135077	0.0719191	19.654
R01_IMN_Public.Safety.Budget	NA	NA	NA

	Pr(>|t|)	
(Intercept)	0.11392	
R01_Fiscal.Year	0.19474	
R01_County.Region	0.00217	**
R01_Public.Works.Budget	3.54e-13	***
R01_Health.and.Social.Services.Budget	0.00518	**
R01_Culture.and.Recreation.Budget	< 2e-16	***
R01_Community.and.Economic.Development.Budget	0.28261	
R01_General.Government.Budget	< 2e-16	***
R01_Debt.Service.Budget	0.35203	
R01_Capital.Projects.Budget	< 2e-16	***
R01_Business.Type...Enterprise.Budget	2.74e-06	***
R01_Transfers.Out.Total.Budget	< 2e-16	***
R01_Total.Expenditures.Budget	< 2e-16	***
R01_IMN_Public.Safety.Budget	NA	

Using Pseudo R-squared with the testing set to evaluate the linear regression models you can see that all points are on or very close the regression line therefore yielding a R-squared of .9732. meaning it is an effective model for explaining the variation in finding the total actual budget for any given county in Iowa.



MAE
0.001785

Evaluation:
To evaluate the 4 different models, we ran the models on the testing set using the Pseudo R Squared. Pseudo R Squared shows how much of the variance in the data is explained by this model. Out of all the 4 models, the ANN had the highest Pseudo R Squared of .9828. We chose the ANN to be the best model because it gives us the highest percentage of variation for our target variable which is Total Actual explained by the other variables in our data set which include public safety budget, public works budget, health and service budget, business enterprise budget and a few more.

Models	**Pseudo R Squared**
Random Forest	.9826
Decision Tree	.9727
Linear Regression	.9732
ANN	.9828

Weaknesses and Possible Improvements:
A weakness in our models that we were able to notice is that some of the different budget types like Capital Project Budget, contain a lot of zeros. This means that the Capital Projects Budget does not impact the findings significantly because it does not have as strong a correlation to our target variable as the other budget types.

An improvement we could make to improve the accuracy of our models in the future is to work with data for each county's spending budget. This would give us better insight of whether there a certain counties that tend to exceed their budget. For example, instead of saying the county region 2 exceeded their annual spending budget, we could be able to determine what specific county from that region was the main cause that led to exceeding the annual expenditure budget.

Business Insights:
Insight models provide: With the models' output and their respective Pseudo R^2 and MAE values, anyone viewing the data can gain a better idea of how well the data is correlated when placed in different models. For models with higher R^2 values and lower MAE values, their data can be trusted to fit more accurately.

How model can be used in real application:
In real application, we can use the models and their resulting values to predict future values and how different inputs can predict different outputs. For example, in our models, we can input different values (not even necessarily all of the ones including in creating the models) and

predict our target value total.actual. With this info, Iowa counties could build better budgets with more certainty that they will turn out as they expect.

What related problems can it solve:Forecasting more accurate budgets can help Iowa counties reallocate the amount of money they reserved to handle fluctuations in their budgets. This reallocation of funds can help support projects which could benefit from additional government support. This additional allocation can also be used to support homeless shelters, public schools, public parks, and any other initiatives that were struggling financially.

What improvement can be made: As stated in the evaluation, by including actual expenditures for each estimated budget. We could have pinpointed which area has more faults in their budget estimation. Unfortunately, due to the time constraint we could only work with one target variable. Nevertheless, including specific budget and actual expenditure for all the categories would have provided more insight and is something that can be improved upon next time.

Summary:
To sum, through evaluating Decision Trees, Random Forest, Linear Regression, ANN models and using Pseudo R Squared we discovered that the ANN model could most accurately predict our target variable. We believe using the ANN model, we can help predict more accurate total actual expenditure. This accuracy in prediction can help the Iowa government allocate the funding to other projects.