**Final Report- Financial Analysis of Irrevocable Gifts to LCMS Foundation**

**Introduction**

Each year, the Lutheran Church- Missouri Synod Foundation receives a certain number of irrevocable gifts. To predict the amount that the Church will be receiving on an annual basis is helpful in assisting the Church plan financial operations. We have tasked ourselves with finding an accurate and representative model of the annual amount in donations based on factors such as trends amongst people from different zip codes, ages, marital statuses, and other personal factors, as well yearly fiscal data, birth rate, and unemployment.

There are many factors at play in the giving of gifts. A person may give one year to the Foundation, but not another. There could be an emotional and spiritual aspect of giving habits. These factors are difficult to assign a measure, and there is no standard metric for how a person *feels* toward giving a specific year. The closest metric we could access in this realm of religious giving would be a “measure” of faith. It is a valid assumption to assume that all LCMS Foundation donors are members of the LCMS. Since faith cannot be explicitly measured, the closest numbers available would be church memberships (church-goers that have completed confirmation courses) or the number of those baptized (includes both confirmed and unconfirmed baptized church attendees). The entity of the LCMS that holds this LCMS-specific data is Rosters & Statistics. However, they are particular in choosing who to divulge their data to, and the data comes at a monetary price. So, we did not get the chance to work with LCMS-specific data for this project, since our request for data was not approved, nor do we have the funds to purchase the data.

Instead, we will focus on external factors that are made available to the public. Other than faith, financial and social factors also play largely into giving. When considering the options of features for a specific year, we concluded that four factors would be useful in predicting donation amounts: unemployment rate, GDP growth rate, S&P 500 data, and birth rate. The first three features constitute the financial side of giving for our predictive models, and the birth rate gives us a possible glimpse of how the general population feels about the social status of society. Our logic follows that if a couple believes their society is in a bad place (for financial or other reasons) then they will choose to not have children, and the inverse holds true. This may or may not be a faulty assumption, but nonetheless gives us another feature to work with. We are effectively predicting the decisions of individuals, and since human behavior is both an art and science, we try to choose features that reflect this.

We decided to use linear regression, k-nearest neighbors, random forest, and a neural network to predict the total irrevocable gift amount per year. Although we anticipated that some of these models will have a better performance than others from the get-go (for example, a model such as kNN is highly likely to have better performance over a simple linear regression model as kNN can explore a nonlinear space), we still wanted to attempt to implement these algorithms ourselves and explore differences in outcome.

Linear Regression plots each feature of the data points along an axis, as well as the label and ultimately finds a line of best fit with least sum of squares error between the points and the line of best fit. In order to make predictions, it follows the formula of the line. When inputting features of a novel dataset, each feature is plotted on the line of best fit, and the label that is returned is the value indicated by the line of best fit where all the features lie.

K-nearest neighbors better accommodates many data points with a large number of features. Additionally, since our data is likely to be noisy given the wide range of different donation types and circumstances, k-nearest neighbors would lead to a more developed model compared to other “simpler” data mining methods such as linear regression.

Random forests can address the overfitting problem of a model. The predicted value is calculated by averaging/aggregating each single decision tree’s output, making for a robust model.

Finally, a neural network is a powerful model that can detect patterns within the features to be exploited to better predict our outcome. For example, if GDP growth rate was found to have a greater predicting power than unemployment rate, the neural network automatically takes that into consideration when assigning weights to features. It also utilizes a nonlinear activation function that allows for more accurate predictions.

**Existing Methods and Related Work**

Many researchers have explored the relationship between various individual demographic, financial, and personal factors and charitable contributions using various data mining algorithms such as regressions and classification trees. In his study on two thousand individuals, Meyers examines the demographics and financial information that are important determinants to charitable donations. Meyer used two regression models in order to split the effects of independent variables on two dependent variables. The first model looks at the dependent variable of whether a household will donate while the second model looks at the amount that a donor donates given that they have decided to donate. The independent variables used for both models were individual household characteristics such as income, age, pension, life insurance, and education. For the first model that measured contribution indication, Meyer found that the positive significant independent variables include income, age, life insurance, number of businesses managed and spouse education while saving habits ha d a significantly negative relationship. In the second model which predicted contribution amount of households who chose to donate, Meyer found that education levels were a significant feature and that as income increases, log(charitable contributions) increases at a decreasing rate. Meyer notes that some of the drawbacks of his study is that there exists some instances of collinearity and heteroscedasticity that may have affected the model’s fit (Meyer 7).

Like Meyer, Abreu also explored factors that affect donation practices, but instead analyzed more personal independent variables, such as “egoism, altruism, voluntarism, compassion, religiosity, denial factors, attribution factors, religious affiliation, gender, and age,” using classification and regression trees (Abreu 3). Abreu splits up his analysis into three trees to evaluate three different dependent variables. For each dependent variable, Abreu followed three steps to create a classification tree and conducted eight trials to find the most accurate tree. The first classification tree focused on predicting frequency of donation. Abreu found that “the two most important attributes for distinguishing between regular donors and non-regular donors, are voluntarism (100%), religious affiliation (94.2%), followed by religiosity (88.1%)” (Abreu 5). The second classification tree analyzed the type of organization donors tended to donate to, which included religious organizations and secular organizations. With 62.2% accuracy for the entire sample, Abreu found that the “two most important attributes are religious affiliation (100%) and religiosity (83.1%)” (Abreu 8). The third classification tree evaluates how different factors affect an individual’s level of donation. From this tree, Abreu found that the age and religiosity are the two variables that best distinguish between high level of donations from low level of donations (Abreu 11). From the study, Abreu concluded that voluntarism, religious affiliation and age are the most important attributes when considering frequency of donations, type of organizations donated to, and level of donations, respectively.

In contrast to Meyer and Abreu, who both focused on the individual and his or her lifestyle, John A. List conducted a model looking at external variables’ effects on charitable donations by analyzing the relationship between the Standard & Poor’s (S&P) 500 stock index of the previous year, GDP, unemployment and consumption expenditures on annual charitable gifts. Using regression to analyze the data, List discovers that a 1% increase in last year’s S&P 500 index is correlated with a 0.19% increase in charitable giving this year and that individuals are significantly more responsive to economic improvements than to economic declines. This same pattern is prevalent for independent variables: GDP, unemployment and consumption expenditures. This suggests that religious giving is nearly unaffected by economic times.

Most existing research has been more focused on individual motivations and characteristics as opposed to macroeconomic factors. And, although List and Meyer do incorporate economic variables into their studies, we believe that both demographics and economic factors are important variables that affect how much is donated every year. As a result, our model will be exploring a combination of demographics and a wider range of economic factors when predicting annual donation.

**Methods Adopted in Analysis**

We used four main methods in the analysis of our dataset in order to predict donations for future years:

Linear Regression: a single “line of best fit” used to predict a label that is assumed to be scalar, using multiple features to fit the line. The line is fit to the data based on minimizing the residual sum of squares, and predictions are made based on fitting parameters into the function of the line and realizing its output.

k Nearest Neighbors: a method in which all the datapoints are plotted as points on a graph where the axis are features, and the label is an attribute of the node. The novel datapoint is introduced to the graph and plotted along the graph according to its features. The most common label of the k nearest points to the novel point (k is a parameter) determines the label of the novel datapoint.

Random Forest: a model that conducts independent regression decision trees to predict the dependent variable. Each tree is developed from a different subsample of the original dataset and uses information gain to evaluate which feature to split the tree on.  A new data point’s dependent variable is predicted by calculating the mean value all decision trees in the model.

Neural Network: an artificial neural network is a machine learning model designed the mimic the transmission of information in actual neurons. It consists of an input layer, hidden layer(s), and an output layer, each layer connected by weighted edges. The nodes in the hidden layer transform the inputs into nonlinear space, allowing us to find trends that simpler classifiers such as linear regression cannot. The model “learns” the weights, updating them to minimize the loss function each time training data is back propagated through the network.

**Setup and Exploration of Data**

We evaluated each method’s performance based on comparing the mode’s prediction results to the ground truth values from the original dataset using statistics such as r2 and standard error. In addition, each method was also compared qualitatively by looking at how efficient the model is and how well the model represents the dataset.

The main comparison criteria that was used is R2 and standard error. The predicted annual donation for 2018 was also used to compare results between models and will be used as an evaluation criterion as the 2018 year ends, as at the time this was written the 2018 year has not yet ended and there is no ground truth data for the total amount of donations the church has received.

We can begin our data mining by reviewing the format of our data. Data was collected when donations were made, and whatever personal information the individual felt comfortable divulging was recorded. The data included gift date, amount, gift code (which refers to the type of gift that was given), zip code of the location in which the gift was given, state, gender, marital status of the donor, birth day, age of donor when gift was received, current age, and finally, whether the individual is deceased. Figure 1 shows five rows of sample data.

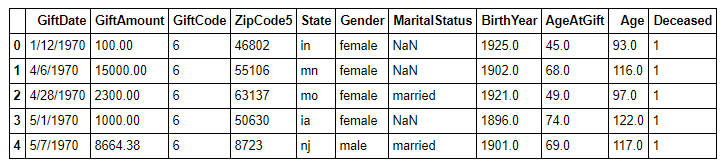
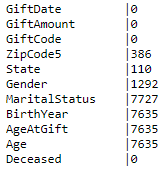


Figure 1. Five rows of sample data.

We have a total number of 26,897 data points, some of which have missing attributes. While certain attributes such as gift date and gift amount do not have any missing rows, other attributes such as marital status, birth year, age at gift, and current age have over a quarter of the rows missing, as can be seen in Figure 2. When creating our model, the missing data is considered and processed accordingly, depending on the model.

We ultimately decided to run our data mining algorithms in Python and R for two main reasons. Primarily, Python and R is commonly used for data science and already has a multitude of frameworks that simplify data science through better data organization and variable management. Secondly, our team members are already familiar with Python and R.

Figure 2. Written next to each attribute is the number of missing data points of that attribute.

We began our data analysis by analyzing several independent variables. The most obvious and trivial would be to determine the frequency of each amount of donation over the entire dataset. Since amount donated is not a discrete value, we designated a series of evenly spaced buckets where the range of each bucket is 500. These buckets spanned from the lowest recorded donated amount, which was $0, to the highest amount donated, which was $5,402,615. This resulted in approximately 10,000 buckets. We assigned each donation to a bucket if it existed within the bucket’s range, and then counted the number of elements in each bucket. We plotted our results on a log-log scale.

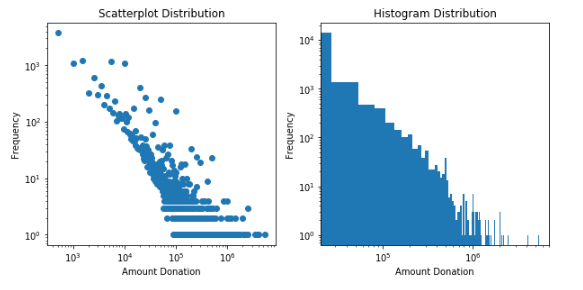
From the scatterplot in Figure 3, it is evident that, while the majority of people donate within $10,000, there are potentially around two hundred people who donate well above $100,000 and probably less than 20 who contribute amounts into the millions. This pattern is even more evident in the histogram, where the outliers are included in the bars. We can see that the frequency of donations versus the amount donated follows a power-law distribution, which is to be expected.

Figure 3. Frequency of donation amounts.

We were then curious about the discrepancy between the average amount that females donate and the average amount that males donate. (Note that I say females and males instead of women and men, because some donors are less than 10 years old.) Our results showed that females donated on average $29,625.99, while males donated on average $31,557.35. While these results suggest that males donate on average almost $2,000 more than females do, the standard deviation for each average was 128,406.85 and 123,429.65, respectively. This indicates that we have humongous variance in our data, and that averages are not necessarily a good representation of population. Nevertheless, the median donation for females was $4,025.04, and for males was $5000- indicating that there is a trend where males tend to donate more than females do, although a specific value is difficult to determine.

We wanted to analyze a few more independent variables before developing our model, so that we have an idea of how independent variables effect on donation amount. We found the five states with the highest average donation, as well as the five states with the lowest average donation, shown on the following page. It becomes apparent that the states with the highest average donation have, in general, very few donors, indicating there exists a small group of people who donate very large amounts. It is also possible that the far-right outlier from our scatterplot in Figure 3 resides in one of these states, skewing our results far to the right.

It was difficult to find a correlation within the lowest average states. The top three states had very few donors, but the fourth lowest state had well over 1500 donors.

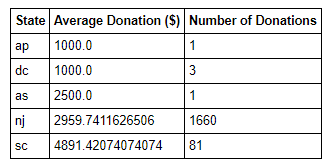
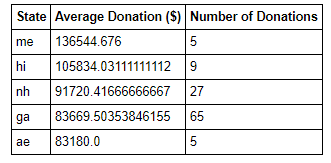


Figure 4. Left: Top 5 highest averaging states. Right: Top 5 lowest averaging states.

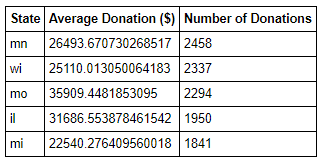
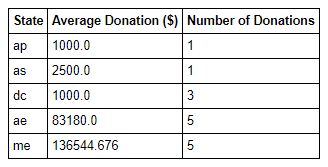
 We also pulled the top 5 states with the most donations, as well as the top 5 states with the least donors. In Figure 5, the states with the greatest number of donations all have a relatively high annual gift value. However, the top 5 states with the lowest number of donations all have extremely low annual gift values. Combining the data derived from the top 5 lowest average donation in Figure 4 and top 5 lowest donations in Figure 5, we can begin to glean a correlation between a low number of donations and a low average donation.

Figure 5. Left: Top 5 highest states with greatest number of donations. Right: Top 5 states with least number of donations.

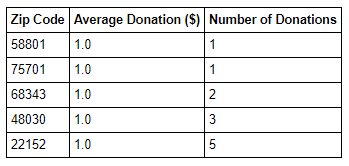
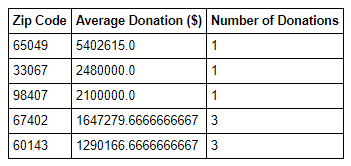
A similar pattern can be observed within zip codes. Note that because of a certain way in which the method that was used to calculate the average donation was implemented, an average donation of $1 actually indicates an average donation of $0, since for some reason, donations of amount $0 were included in the original dataset. Additionally, zip code “nan” indicates that no zip code, or an invalid zip code, was provided. The method used grouped all of those values together.

Figure 6. Left: Top 5 highest averaging zip codes. Right: Top 5 lowest averaging zip codes.

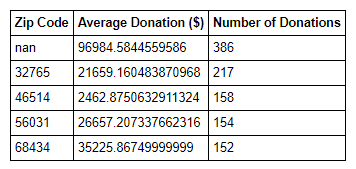
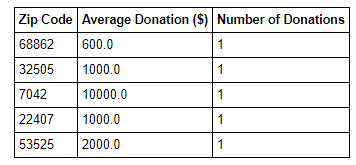
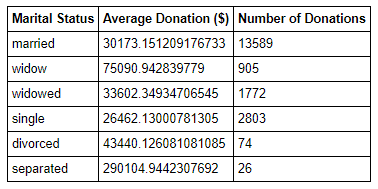
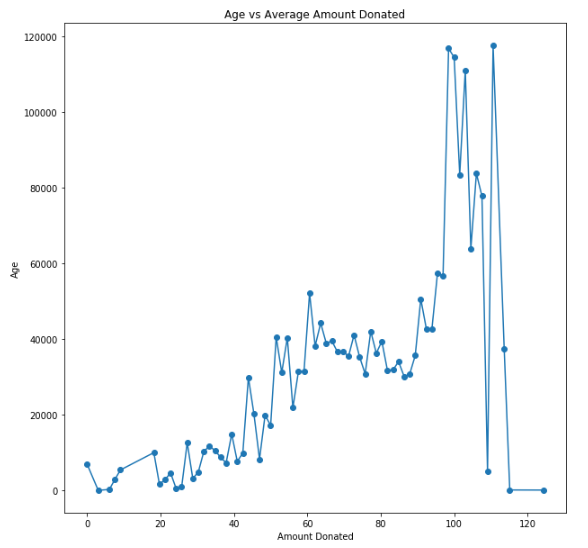


Figure 7. Top 4 highest zip codes with greatest number of donations (excluding nan). Right: Top 5 zip codes with least number of donations.

With regards to the marital status of the donor, we wanted to see the average of each category. Interestingly, widowed individuals donated the most on average, whereas single individuals donated the least on average.

Figure 8. Average donation of different marital statuses.



Age was also an interesting independent variable. It’s intuitive that the older an individual is, the more money they have earned, and the more likely they are to donate. However, we did not expect to see a trend in which individuals who were close to or over a hundred years old tended to donate the most. Ages climbed as high as 120 years old. As it is unlikely that there exists so many centenarians, we figured the majority of people who are donating at such an advanced age are donating the rest of their worth to the Church after death. A quick cross-reference with the “deceased” column of our dataset confirmed these suspicions.

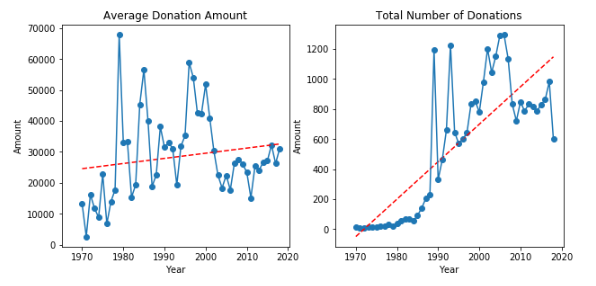
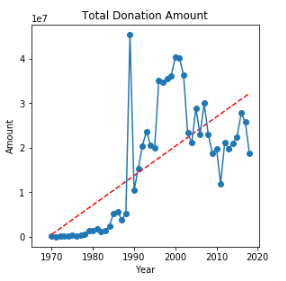
Finally, we analyzed the average and total donation amount from year to year, as well as the total number of donations each year. We observed that the trend for average donations stayed comparatively stable- there would be years where the average donation would peak, and years where the average donation would drop, but the trend line stays relatively flat. However, the trend for total donation amount as well as the total number of donations is a considerably sharp positive slope, although the data points seems to flatten in the recent few years. Finally, it is interesting to point out that just before 2010, there is a global peak in the total number of donations. While a peak also existed in the total donation amount, its value was far from close to the maximum peak, and the local peak in the average donation plot just prior to 2010 is actually underneath the average trend line. Unmistakably, many factors play into the donation amount, and there are complex interactions at play.

Figure 9. Figures derived from donations as a function of time

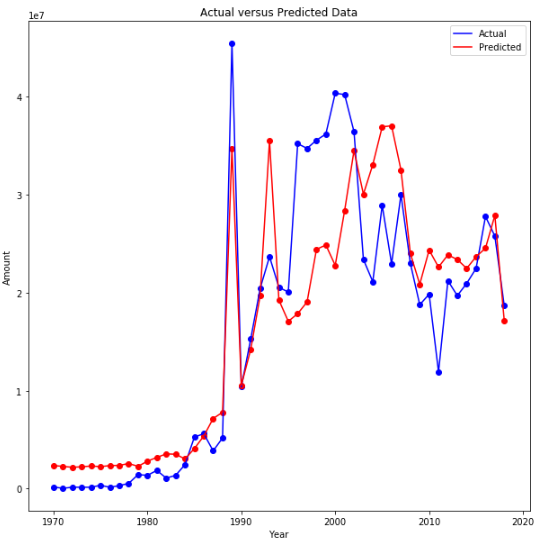
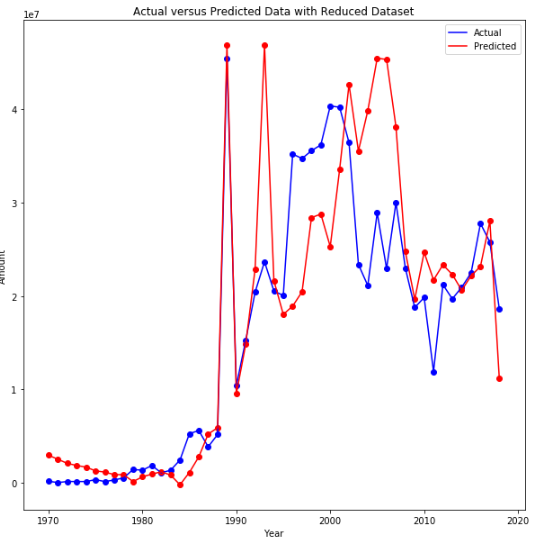
**Results- Linear Regression**

Figure 10. Left: Actual vs Predicted Values. Right: Actual vs Predicted Values, except several datapoints have been removed before training model.

Now, we can begin to develop our models for our data, which we thoroughly understand. The first model we develop is a linear regression model. Our first linear regression model was simple- we used the year and the number of donations for that year as features, and we also used the total value of donations for that year as our label. We scaled and whitened our features, while placing slightly more weight on the number of people who donated to generate our model. We expected an extremely poor score for this model, but after generating 1000 random models, we found a mean coefficient of determination (R2 score) of 0.664- much better than what we had originally anticipated. The standard deviation of scores however, was larger than ideal, at 0.20.

On the left of Figure 10, we can see the results of the linear regression model which we created. Although the trend of our model follows in general the trend of the actual data, one serious issue to consider is the possibility of overfitting, which means that the model we have trained is good at predicting only our model, and may scale extremely poorly to real world data. Since gift data has only been collected for approximately 40 years starting 1970, if we are to estimate total amount of gifts in a particular year, this indicates that we only have approximately 40 rows of data. In order to fix this, we randomly selected approximately 10% of datapoints to be left out while creating our model. The resuting predicted trend is plotted on the right. Surprisingly, this model performed much better in estimating our data- after 1000 random trials, it has a mean R2 value of 0.745 with a standard deviation of R2 score of 0.009. The average standard deviation for these trials was 0.08 with very little variance.

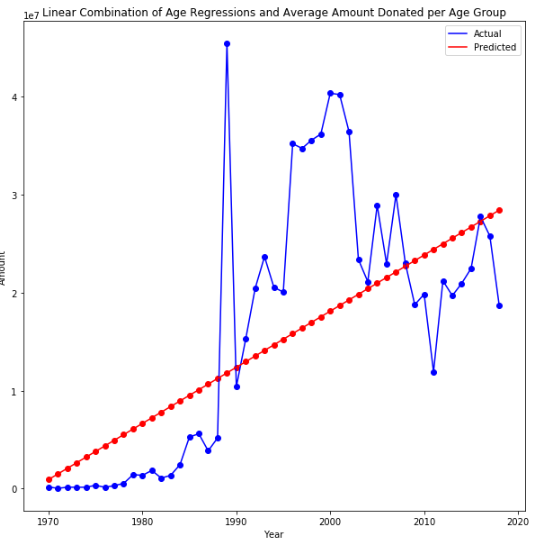
We began considering what we could do to improve our linear regression model. We tried to break down our independent variables find a regression for each independent variable. For example, we found that on average, each age group donates different amounts, and that as years progress, the number of donors in each age group changes. Perhaps there was a linear pattern in how the number of individuals in each age group changes versus time. Following this idea, we performed 10 separate regressions on each of the 10 age groups, using a method that similarly follows the one we used when finding frequency of donation amounts. We divided the age distribution into buckets of even ranges, and measured how the count of each bucket changed over time. We performed a linear regression on each bucket, and used this data to predict the number of individuals donating from that particular age group for each year. Finally, we performed a linear combination of each of these age groups for any particular year, in which we multiplied the expected number of donations within that age group with the average amount per donation in that age group, and summed up the values over all bins. We noticed a sharp decline in the R2 results of our data to approximately 0.475, but the standard devaition of our scores also decreased significantly to 0.177. (NOTE: These results are different than the ones indicated in our midterm progress report- a mistake in methodology was made). Thus, while our scores became less accurate, they also became more precise- which can be obviously derived from the graph generated in Figure 11, depicting the predicted amounts as well as actual ground truth data. Note that the trend line appears to a simple line of best fit for the plotted data points. However, recall that the prediction line was generated without using any of the labeled data (the total amounts donated per year). Keeping that in mind, we can see that our linear combination model did a fairly good job of estimating a linear trend for the total donations per year.

Figure 11. Linear combination regression model.

However, developing a model from aggregating linear regression and linear combination only points us down into a deeper rabbit hole. For example, the average donation of each age group also changes from year to year- which is a pattern we did not analyze and must also try to find a linear regression for if we are to account for how each variable changes. Simply put, there are too many variables and dependencies to perform linear regression thoroughly. Furthermore, when using a linear model for predictions, we are constraining ourselves- especially if the pattern of change in a particular variable is nonlinear. In order to generate a more powerful model, we can use a non-linear method, such as k-nearest neighbors.

**Results- k Nearest Neighbors**

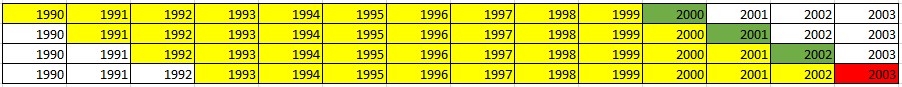
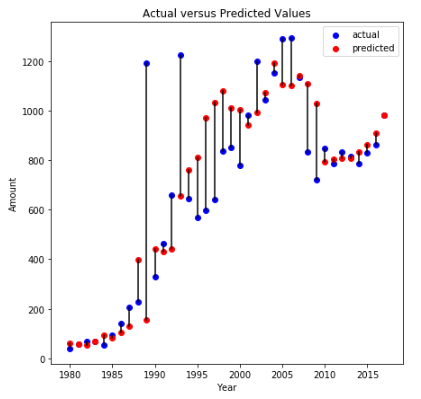
k-nearest neighbors works by finding the k most similar data points closest to the new datapoint, and predicting the feature based on the labels of those k nearest data points. However, due to information the algorithm needs to work, we run into a slight complication. The output for our data should be the amount that is expected to be donated in any given year. This output relies on features such as how many individuals of each age are donating, how many of those who are donating are female or male, and how many individuals from each state are donating. The issue is, each of these features are time correlated, so in any given year in which to predict the label, none of the features will be available to develop our prediction. Thus, we have the additional complication of having to predict what each feature will look like before being able to predict what our generated output is like.

Figure 12. Yellow values indicate the frame of features (10 years) used to develop a model for the label label, which is in green. The red value is a year in which no data has been collected. We can use the model developed from the previous years in order to predict what the value of this particular feature in 2003 will be.

The most obvious answer to predict how a single feature changes as time progresses is to use linear regression based on the pattern presented by the feature in previous years. However, as we have stated previously, our features may not necessarily develop in a linear fashion. We can utilize this in our kNN analysis. In order to develop a time correlated model for kNN, we can treat the value of each feature as a label for a certain frame of predecessing x number of years. For example, assume we have a complete set of data from 1990 all the way until 2002. To develop our kNN model with a 10 year time frame, for the value of a feature in the year 2000, we look at the previous 10 years of values for this particular feature- 1999, 1998, and so on until 1990, in order to have features to predict the value of the feature in the year 2000. We proceed down our dataset with a 10 year frame- for the label in 2001, we look at the past 10 years for features, until 1991, and for 2002, we look for the past 10 years to 1992. When it comes to 2003, an unknown year with no data, we can use the model we developed from the past three years, use the datapoints from 1993 to 2002 as features, to predict the value of the feature in 2003. Figure 12 explains this methodology graphically.

We first applied this methodology to the number of people who donate per year. Since we only have a complete set of data from 1970 to 2017 (2018 isn’t over yet!), we only had 48 data points to work from. Thus, when cross validating, we took out four data points and trained the rest. Surprisingly, our predicted results were more consistent with the actual data than expected. On average, we had an R2 value of 0.640, and a standard error of 0.100. Results to our cross validation can be seen in Figure 13, where the blue points represent actual values and the red points are predicted. The farther away a red point is from a blue point, the more error there was in calculating that point.

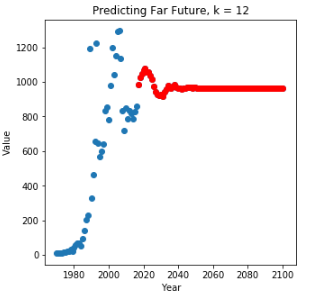
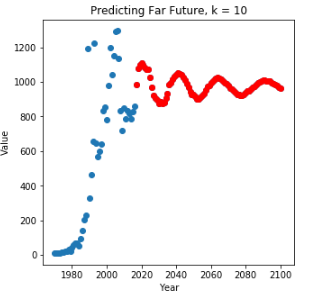
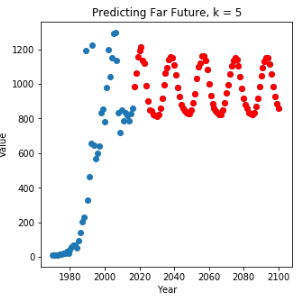
It becomes clear that our kNN model would probably perform relatively well for data that is not far from the endpoint of our dataset, since we have ground truth data to base our predictions off of. If we wanted to predict the number of people who will donate in 2018, for example, we have ground truth data from 2007 to 2017 to generate our 2018 prediction. However, if we wanted to generate a prediction for number of people who will donate in 2019, we will have to use our predicted value for 2018 in conjunction with ground truth values from 2008 up until 2017. The problem then is apparent. If we wanted to predict data for the year of 2028 and we’re using a frame of 10 indeces, then all of the data which we use to predict the 2028 year- the data from years 2018 up until 2017, are all previously predicted values. We will be making predictions based on previous predictions only, and no ground truth data. We decided to see what would happen if we attempted to predict data 100 years into the future, which is depicted by Figure 13.

Figure 13. Blue indicates actual data, red indicates predicted data. Predicted data is generated via a kNN model with number of neighbors being 5, 10, and 12, respectively, from the left.

Figure 14. The farther away the blue points are from the red points, the greater the error was in that prediction.

When using a smaller number of neighbors, such as k = 5 in the first graph in Figure 14, we see that our data has taken on a sinusoidal pattern. This is likely due to the fact that the last approximately 10 years of our data form a convex parabolic shape with the leading edge curving up just ever so slightly. However, using higher numbers of neighbors such as 10 or 12, as can be seen in the center or rightmost graph, tends to “average” out the data more and eventually converges to predict one continuous value- the average of the peak and valley of our previously generated sine function. This verifies our hypothesis that using kNN to predict future variables based on previous data works best for predictions for years immediately following our ground truth data, and is less able to replicate data as time passes. Furthermore, we can accentuate local patterns if we choose a lower number of neighbors for our kNN model, or we can choose a safer “average” for our predictions if we select a higher number of neighbors. In all further mentions of kNN, we stayed with a more conservative value of k, where k=10.

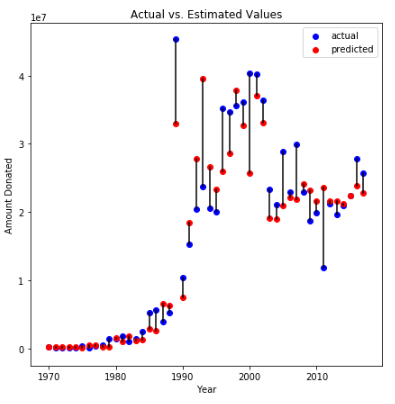
In line with our next steps when we were developing a regression model, we divided the range of ages into 50 buckets, and then placed each individual who donated in a given year into their respective bucket. We then counted the number of items in each bucket and used those numbers as features for our kNN model. In order to test the model, we randomly removed four data points and trained the model, then used the data points previously removed to test the accuracy of our model. We did this repeatedly until each datapoint had been removed at least once and tested against the rest of the data. Results can be viewed in Figure 15. Incredibly, using this model, we had an R2 value of 0.86, with a standard error 0.059. This is a significant improvement from our previous kNN model where we measured only number of people who donated per year as the feature and had a resulting R2 score of 0.640. This kNN model also performs better than previous linear regresison models, where the highest scoring model had an R2 score of 0.745.

Figure 15. Predicted versus Actual Values

Due to the accuracy of this model, we decided to make an estimation on the total amount donated in the year of 2018. To do so, we used our kNN model to predict the total amount donated in the 2018 year. We did so by measuring how each age group changed per year, and running kNN regression on the previous data to find the next datapoint, as described in the methodology above. We performed this for each age group to have the age group data for the year of 2018. Finally, we fit our model with the age distributions of each year as features and the total amount donated that year as a label, and we ran a prediction based on the 2018 age group data. The output for year of 2018 is $21,965,646.

**Results- Random Forest**

A Random Forest model uses many independent decision trees to predict the dependent variable: annual donation. Since annual donation is measured continuously, not categorically, regression trees were used instead of classification trees. The features examined were annual unemployment rate, birth rate, GDP change percentage, SP500 change percentage, gender, marital status, state, age when donated, and whether the donor is deceased.

Before developing the model, we had to preprocess the original dataset because the input data was of individual donors. Many of the features were either continuous (age) or had many different categories (state), and there were many records with missing information. Since we were looking to predict annual donations, we had to split features into sub-features in order to aggregate the features’ values under each year. For example, marital status was split into multiple categories. And, for each year, the percentages of the people who were married, single, widowed, separated and divorced were listed as separate categories, as can be seen in Figure 16. Secondly, if we trained a model using continuous features or features with many different categories, it would have taken a lot of processing power, since a single node could have been split into more than fifty branches depending on the number of categories existed. This would have greatly increased the size of each tree and the cost of the entire algorithm. To address this challenge, we split features into buckets, as was done previously. For example, age was split every 10 years and state was split into regions (East, Midwest, South, West, Other). Once features were aggregated for each year and percentages were calculated to normalize the data, the values under each feature was numerical and continuous. Therefore, we used the mean and standard deviation for each feature to categorize into five buckets, represented in Figure 17. Lastly, Random Forest models are not the best with features that having missing information, therefore we made the decision to remove all records with missing feature information. By manipulating the input dataset into categorical features, as shown in Figure 18, and removing records with missing data, this increased the efficiency of training the A picture containing furniture

Description automatically generatedA screenshot of a cell phone

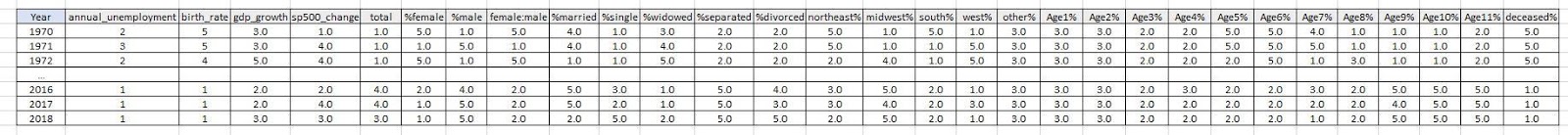
Description automatically generatedmodel at the cost of accuracy.

Figure 16. Manipulating the input dataset into categorical features.

Figure 17. Categorizing features into buckets based on mean and standard deviation.

Figure 18. Percentage of people in each category.

A screenshot of a social media post

Description automatically generatedIn order to cross-validate our results and to examine the accuracy of the model, we used 80% of the dataset randomly to train the model and the remaining 20% to test the accuracy. In addition, we developed a random forest model using 300 decision trees and a minimum leaf node observation size of 5 to address overfitting. For each iteration, we used 12 different years that were chosen at random from the overall dataset to test the model and the remaining 36 years were used to train the model until the model predicted values for all 48 years. Figure 19 displays the prediction results for all years against the actual donation amounts. From years 1970 to 1980, the percentage error between the prediction and the actual value was extremely high (not reflected in graph due to small donation value). These years were when they first started collecting data, and therefore the high error could be attributed to the lack of information and that technology was not as widespread at that time. For years, 1992, 1993, 2011, 2014, 2015 and 2017, the predicted value was extremely close to the actual value (error < 2%). This could suggest overfitting in the model. In addition, the R2 value for the model is 0.86 and the standard error is 0.049. The R2 value and standard error indicate that the model is moderately accurate, but because the dataset only takes in 48 years of data, with 36 years used for training and 12 years used for testing, there could still be signs of overfitting, and therefore more data and information is needed.

Figure . Error in predicted data versus actual in Random Forest Model

A close up of a map

Description automatically generated After comparing the prediction results to the ground truth, we looked at predicting annual donation for future years. Since decision trees take into account features for each year, our model faced the same challenge as K-Nearest Neighbors because none of the feature values will be available for the future. For example, the model needs future GDP values, unemployment rates as well as individual information such as percentage of married people donating and percentage of 60-year-old people donating. All these values need to be predicted prior to making a prediction for future annual donation. Using the model developed in the K-Nearest Neighbors algorithm, we were able to predict the future values of the features given the past values (Figure 20). In the future years, the annual donation seems to be leveling off, which could be a result of using the K-Nearest Neighbors model to predict annual features. As mentioned above, since K-Nearest Neighbors takes in k number of past values when predicting future values, as the year get further away from our true data points, the feature value will also converge because it is based off predicted value as opposed to ground truth values. The annual donation for 2018 is predicted to be 17.8 million.

Figure . Future predictions of annual donations

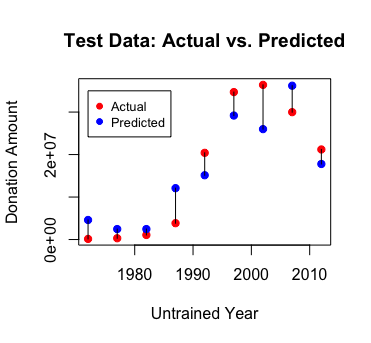
  
  
**Results- Neural Network**

Figure . Predicted vs Actual amount

The neural network accepts four features for a particular year: GDP growth rate, births per 1,000 people, unemployment rate, and S&P 500 open index.  The input data is scaled to resemble a standard normal distribution, as that was required by the neural network builder. The number of hidden layers hyper-parameter was tuned to offer the most realistic and accurate results, which was with just one hidden layer. In the hidden layer, the robust rectifier function was used to prevent the vanishing gradient problem. The output layer was a simple linear mapping since we were looking to make a regression model and not a classification model.

First, we used 80% of the data to train and 20% of the data for testing. The results of this neural network can be viewed graphically in Figure 21.

It can be seen in the figure that when the features of an untrained year are inserted into the network, some of the predictions are more accurate than others. Overall, the average residual was about $5,000,000. When we predicted the donation amount for 2018, we had to partly predict the features. While data on unemployment rate and certain S&P 500 information can be found for 2018, the number of births per 1,000 people and GDP growth rate is not yet definite since 2018 had not yet ended at the time of building the model. So, current the unemployment rate, the S&P 500 open index, and the estimated values for GDP growth rate and births per 1,000 people are used as 2018 features. When these features are fed into the neural network, the predicted total irrevocable gift amount is about $33 million. Late November totals for 2018 showed $21 million in donations were already received. Many people give in December, especially for tax reasons. December of 2017 alone brought in $11 million. So, the prediction of $33 million is reasonable.

The R2 value on the untrained data is 0.83, so it performs better than simply setting the predicted amount to the average past donations. Note that this is a different value from than given in the class presentation because the neural network has been improved since the presentation. The model now contains more epochs (more passes forward and backward through the network).

Additionally, we performed 10-fold cross validation on the neural network to see how well it could predict other untrained years. Figure 22 shows each year’s prediction when it was in the untrained set and its actual value. The R2 value of this model is 0.5, which is lower than ideal. As we can see from the figure, the neural network learns the general trend in the data, with the average error of about $7,000,000 for this model. The R2 value could be low because of distant outlier in 1989 and the high predictions in the early years when the Foundation was not as developed and had much less employees. Overall, this model is a satisfactory one. Even when the untrained years during the 2000s recession are inserted into the model, the network learns that those features mean a more destitute year and therefore predicts a lower donation amount. Also, the model appears to not overfit nor underfit, and it does a good job at following the overall trend, which is important to take into consideration.

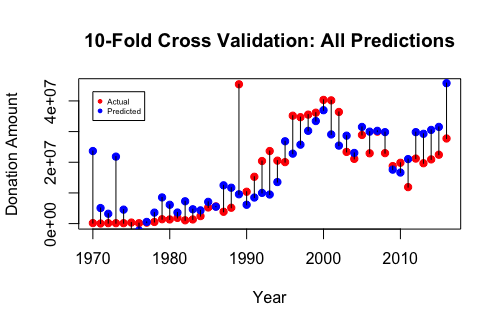
 Tuning the parameters of the neural network to minimize the loss function took careful consideration. When two hidden layers were used, the average residual was around $4 million, but many of the predictions were negative, which doesn’t make sense realistically. When three or more hidden layers were implemented the average residual was also around $4 million, but many of the years yielded the same exact prediction, which also isn’t realistic. Most CEOs do not wish to hear that a model predicted the exact same value as the previous year, as that isn’t the case in the real world. So, we chose to stick with one hidden layer as the results were more pragmatic. Additionally, the number of epochs was tuned to offer the most accurate and realistic results as well. The model worked well with 500 epochs.

Figure . 10 fold cross-validation of dataset using a neural network

**Conclusion**

In summary, after conducting four models and comparing R2 and standard error values, we found that the Random Forest model had the highest R2 value and lowest standard error followed closely by K-Nearest Neighbors. Based off these statistics, one could conclude that these models were the best for predicting annual donations, but these two models both faced the same challenges for time series data as they both need future year features in order to predict the annual donation amount. Since the future prediction is based off feature prediction, these two models are both more susceptible to error and inaccurate predictions. In addition, these models cannot predict too far into the future because the feature predictions are based off the closest past predictions. Neural Networks on the other hand is a more dynamic and complex model when predicting values in the future. Its power can be seen in the 2018 prediction of $32 million which could be the most accurate prediction for this upcoming year. The Random Forest model performs the worst by predicting $18 million while the K-Nearest Neighbors prediction is $22 million. Even though the Random Forest model and the K-Nearest Neighbors model have the best statistics when comparing the testing data predictions to the ground truth, their future prediction power is not as good as the Neural Network model.

We learned a lot form this project. The main takeaways from this project include learning how to work with a real-world dataset and which models were best for various types of data. When we first learned about the different data mining techniques in class, we worked with “perfect” sets of data which included relevant features, no missing information, and a small number of observations. The dataset that we received from the LCMS had a lot of missing information, not enough features and too little years to develop the most accurate model. By working with this real dataset, we learned how to preprocess this dataset to transform it from information about individual donors to information about the entire year. In the process of developing the model, we also learned that K-Nearest Neighbors and Random Forest were not the best models to use for time-correlated data in the long term. The Random Forest model also has a high computation cost for datasets with features that had many different categories or continuous features. Overall, we learned a lot about the different data mining techniques as well as some of the challenges that data scientists face when working with datasets.

We have many ideas for how we may be able to continue this project into the future. We presented our models to the marketing director who was originally interested in predicting donation amounts. Overall, he was impressed with what the performance of the models. However, more work could be done to improve the methods. We could use other algorithms in predicting the features of a year, like the k-nearest neighbors model did. Additionally, the neural network could be made into a recurrent neural network since the output of one year could affect the next. Given more time, we could acquire more relevant features. If the LCMS Foundation were to implement and support this project, then Rosters & Statistics would approve the data request, and the Foundation would cover all data costs. With this new data, we could integrate LCMS-specific features into models. This could boost the performance of the model since it takes into consideration the subculture that we are interested in.

**Distribution of Credit**

Work amongst the three team members were evenly distributed-

Harrison Lu- 33.3% Data Analysis, Linear Regression, kNN

Michelle Xu- 33.3% Research into Previous Methods, Random Forest

Meghan Woodruff- 33.3% Data Pre-Processing, Neural Networks, Liason