**Final Report**

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.

We begin 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.

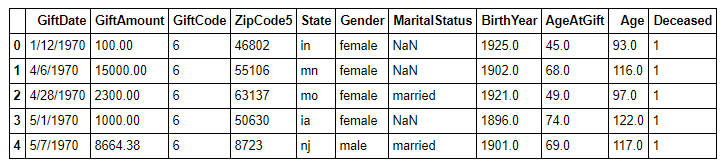
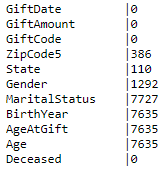


Figure 1. Five rows of sample data.

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 for two main reasons. Primarily, Python 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.

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.

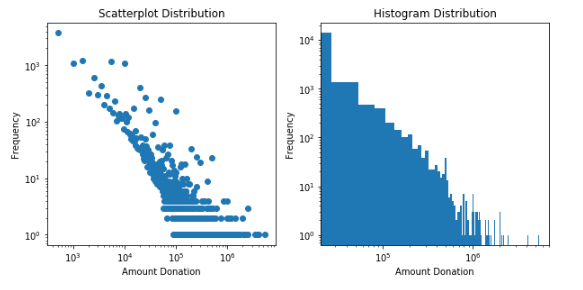
From the scatterplot in Figure 3, it is evident that, while the majority of people donate within $10,000, there are potentially around two hundred people who donate well above $100,000 and probably less than 20 who contribute amounts into the millions. This pattern is even more evident in the histogram, 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.

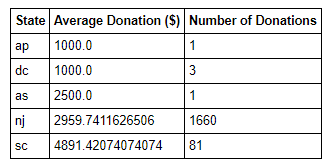
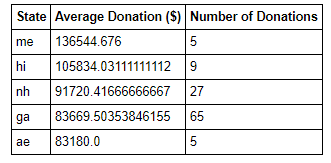


Figure 4. Left: Top 5 highest averaging states. Right: Top 5 lowest averaging states.

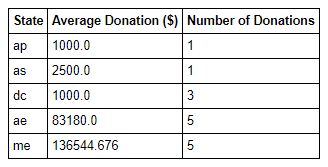
 We also pulled the top 5 states with the most donations, as well as the top 5 states with the least donors. In Figure 5, the states with the greatest number of donations all have a relatively high annual gift value. However, the top 5 states with the lowest number of donations all have extremely low 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.

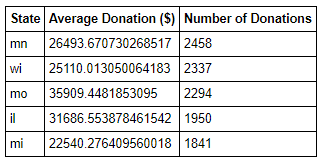
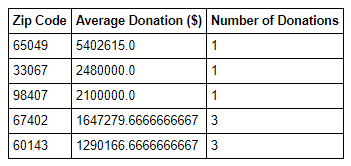
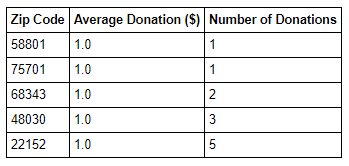
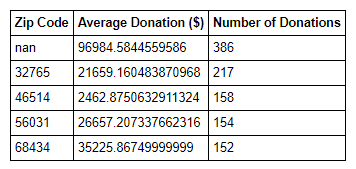
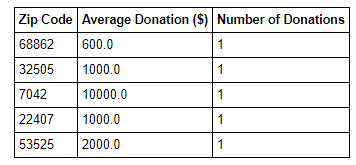
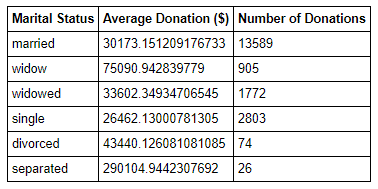
A similar pattern can be observed within zip codes. Note that because of a certain way in which the method that was used to calculate the average donation was implemented, an average donation of $1 actually indicates an average donation of $0, since for some reason, donations of amount $0 were included 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.

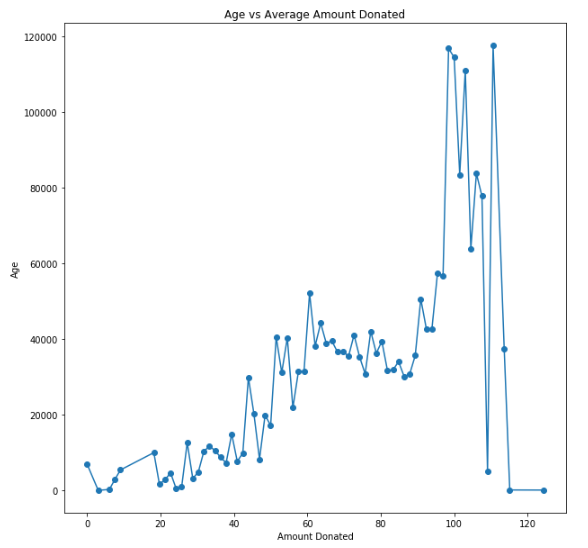
Figure 6. Left: Top 5 highest averaging zip codes. Right: Top 5 lowest averaging zip codes.

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.

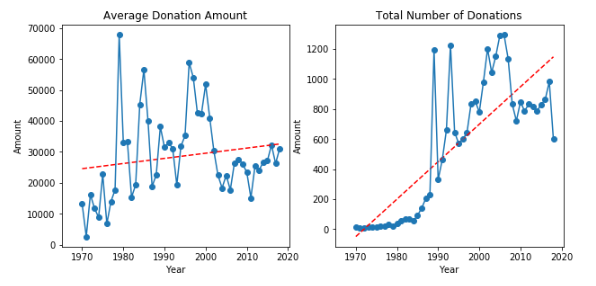
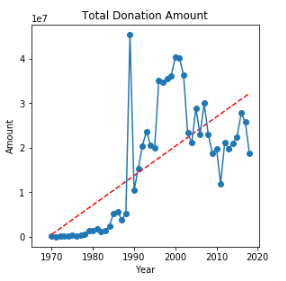
Finally, we analyzed the average and total donation amount from year to year, as well as the total number of donations each year. We observed that the trend for average donations stayed comparatively stable- there would be years where the average donation would peak, and years where the average donation 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

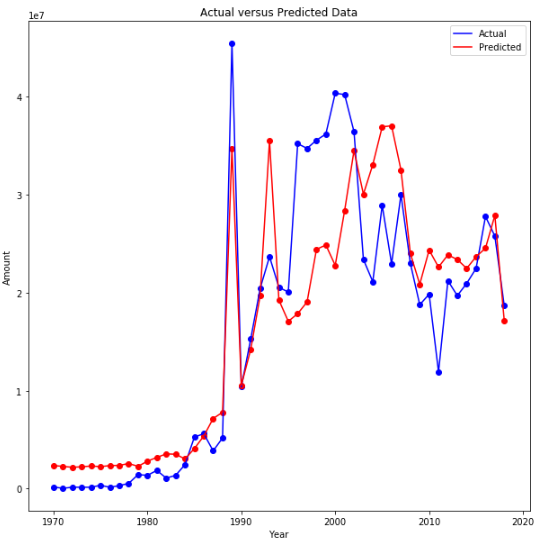
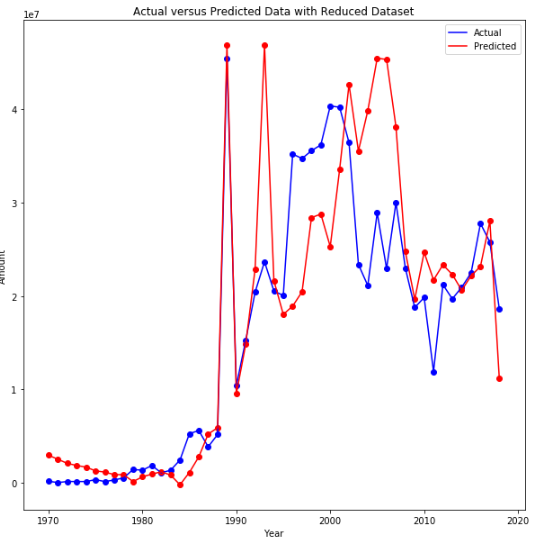
 Now, we can begin to develop our models for our data, which we thoroughly understand. The first model we develop is a linear regression model. Our first linear regression model was simple- we used the year and the number of donations for that year as features, and we also used the total value of donations 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. While we expected an extremely poor score for this model, we cross validating our model 1000 times and 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.

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

On the left of Figure 10, we can see the results of the linear regression model which we created. While 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 follows exactly 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 left. Surprisingly, this model performed much better in estimating our data- after 1000 trials, it has a mean R2 value of 0.745 with a standard deviation of 0.009. The average standard deviation for these trials was 0.08 with very little variance.

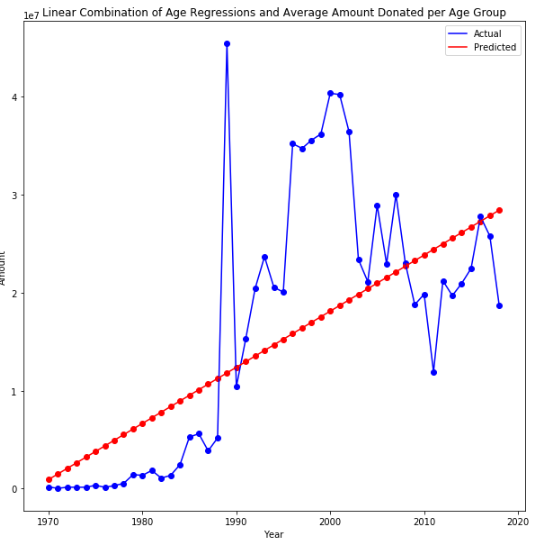
We began considering what we could do to improve our linear regression model. We tried to break down our independent variables find a regression for each independent variable. For example, we found that on average, each age group donates different amounts, and that as years progress, the number of donors 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. Method explained on left.

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 include 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.

k-nearnest 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, here, 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 for the label to be predicted, 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.

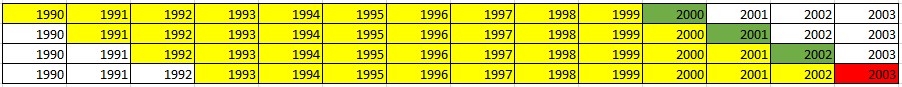
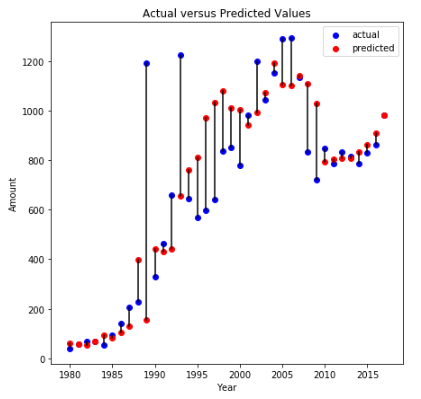
The most obvious answer to predict how a single feature changes as time progresses is to use linear regression based on the pattern presented by the feature in previous years. However, as we have stated previously, our features may not necessarily develop in a linear fashion. We can utilize this in our 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. 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 for features, until 1991, and for 2002, we look 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. We Figure 12 explains this methodology graphically.

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.

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 47 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 12, where the red line represents ground truth. Each blue point is a predicted value. Since the axes are predicted vs actual values, the farther away a blue point is from the red line, the more error there was in calculating that point.

Figure 13. The farther away the blue point are from the red line, the greater the error was in that prediction.

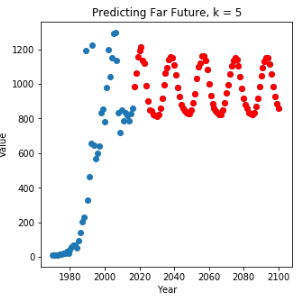
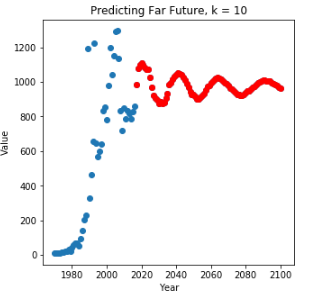
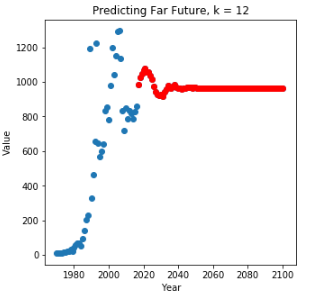
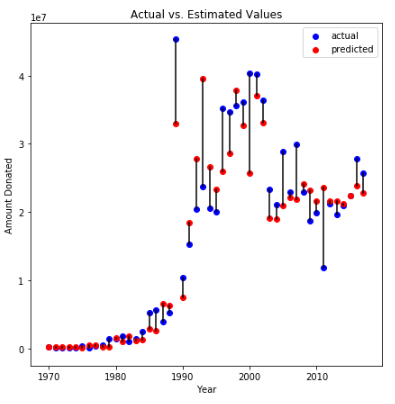
It becomes clear that our kNN model would probably perform relatively well for data that is not far from the endpoint of our dataset, since we have ground truth data to base our predictions off of. If we wanted to predict the number of people who will donate in 2018, for example, we have ground truth 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 14. 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.

When using a smaller number of neighbors, such as k = 5 in the first graph, 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.

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.

Figure . Predicted versus Actual Values