Final Report: Data Science Platform

Henry Goldkuhle, Zander Hartsuff

DS 398

Problem Description

The problem we were given by our advisor Carl McQueen was to take a non-profit organizations dataset, perform data analysis to answer various questions such as, what donors should ample attention be paid to in marketing campaigns? When should the organization expect to receive their donations? Future behavior of donors? How important are different variables to the amount donated by donors? The next task was to then take the insights and knowledge gained from this analysis to build an application that allows a user to input a dataset and run different models to help answer the questions above.

Problem Background

The problem given was then divided into two different subsets. The first being the data analysis performed as well as the wrangling and variable creation involved to use our non-profit data on the application created. When given the dataset we were told not to disclose what organization was behind the dataset and to not give away any personal identifiable data. If these two requirements were met, we were allowed to share any findings from this dataset. Additionally, the dataset was given in raw format and included millions of rows where each row represented a single donation provided by a donor.

The second was the actual creation of the application to account for various modeling needs. When looking there were tools to perform certain modeling techniques but not one that combined many different modeling sets. It was decided that in our application we would focus on five different modeling techniques which included linear and logistic regression, random forest classification and regression and finally XGBoost. Each model had its own unique purpose and served to provide the user with many choices based on the variable they wanted to model to. These were chosen because the models were able to cover both numeric and categorical variables. There were also multiple choices based on your variable type and therefore you could easily see which model was best for your given variable you wanted to model to. One of the things that was discovered through our data analysis was the importance of time series analysis when looking at nonprofit data. Nonprofit data is gathered at periodic intervals, and we were given many years of data collected. This made this data perfect for time series analysis and through this insight we decided to add another page in the dashboard that allowed the user to perform time series analysis on a given dataset.

Design of Solution

The solution was built in two parts again. The data analysis and wrangling were performed on a remote virtual station that ran an R server. We did all our coding for the project in R and used this dataset for most of our work on the application. We then used our wrangled dataset and took a random sampling of 2000 rows to use for the application. This was done as the dataset stored on the virtual machine was too large and therefore to do testing on our application it was imperative to have a smaller dataset. The application was built in R specifically using the library shiny to create and implement our vision. The first part implemented was the modeling page, the dashboard takes in a dataset which is important because we wanted this dashboard to not only work with our dataset but any wrangled dataset making the application able to be used for more than one context. The user then selects the predictor variables desired as well as the response variable. Each model allows the user to see the model summary, data

table that incorporates the response variable, model prediction, residual and then all the predictor variables, scatter plot of actual versus predicted and finally a histogram showing the residuals. There is then a time series analysis page that again takes in data, then the user selects which column represents time and then what variable you want to analyze against time. The user then runs the time series and is given the overall plot, the forecasted plot which shows what is predicted for the future of the model and finally the cycle boxplot which shows the variance among the time sections selected.

Data Wrangling

There is a lot of work that went into preparing the data that was eventually used for the shiny app. The data wrangling section of work is very useful for both understanding the data demographics, and also helping you ask deeper questions about the data. Before we focused on how to prepare the data for our model, we had to do a deep exploration of the data to understand it better. This is an essential piece of our work in order for our model to run effectively and give us interesting results.

The data we used throughout the semester was a financial database supplied by Carl McQueen that was collected from a nonprofit organization that Carl is very familiar with. The data at the start was formatted in two tables. One contained a long list of donations that was over 14 million rows that we labeled `rev_db`, and the other was a list of donors that was about 150 thousand rows. This table contained demographic information about the donors that was labeled `bio_db`, and it had several NA values in the table. Most of our work was done by manipulating rev_db.

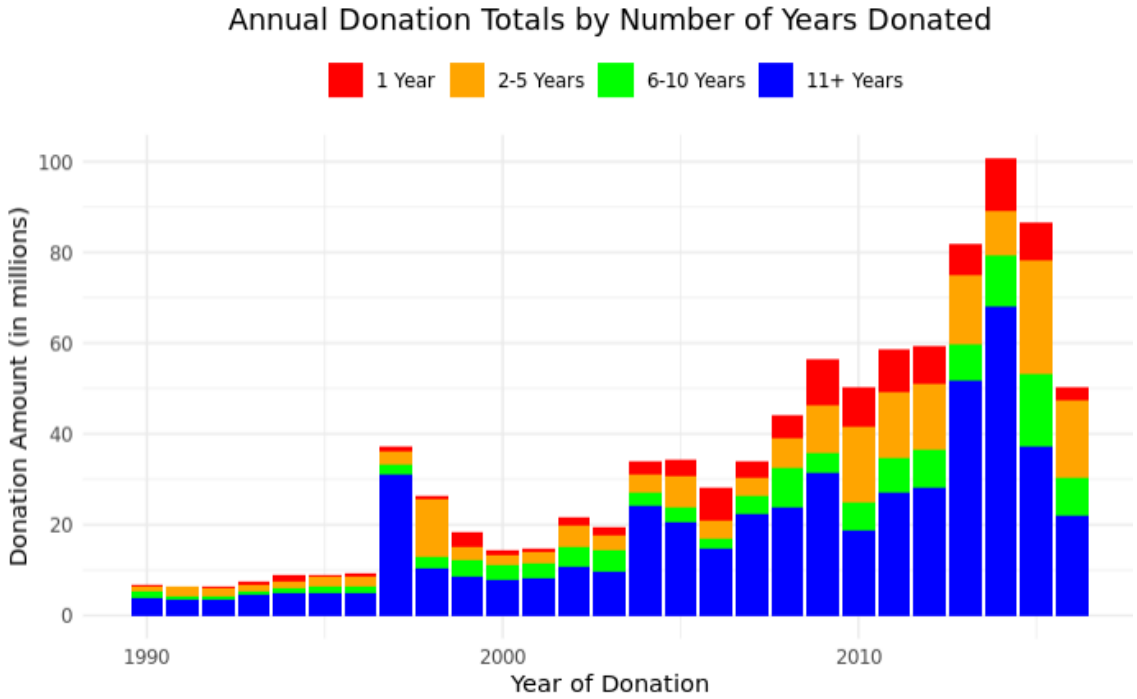
After making a new column called Year by manipulating the date of the donation, we began wrangling the data by grouping by the Year column and the donor ID which was labeled NewID. This essentially means that each row represents one Year of donations for one donor. Each donor has a row for each year they donated. This table was called `donors_by_year`. Looking at the data in this format is useful for looking at trends over time of how individual donors' donating habits change. In this format we summarized a few variables that ended up in our model such as how many times they gave that year (num_gifts), total donations within that year (DonationInYear), most recent year given (most_recent_year), whether or not they gave in December (DecemberGiver), etc. Next, we grouped by NewID again and called the new table `grouped_by_ID`. This allows us to look at the overall statistics of each donor. Some things that we can learn by looking at it this way are: What is the first year they gave, what is the last year they gave, how many total years they gave, how much total donations they gave, a frequency score which tells us how often they were giving, etc.

The next portion of the data wrangling was very tricky. We were trying to observe what the donation trends were like over time, but the formatting was very difficult. Eventually Carl helped with figuring out the correct approach. The approach was to join the donors_by_year table and grouped_by_ID for a new table called `Gift_All`. This allows us to look at each year someone donated and for each row it would tell us what year they started donating. From this we can extract what donation year it was for them—first year, second, tenth, twentieth, etc. Finally, we grouped by year so we could look at trends of the data over time. This table was called `year_breakdown`. The columns for this table included: TotalGivers, YearRevenue, ReturningGivers, ReturningGifts, AttritionScore, and columns for each donor class that

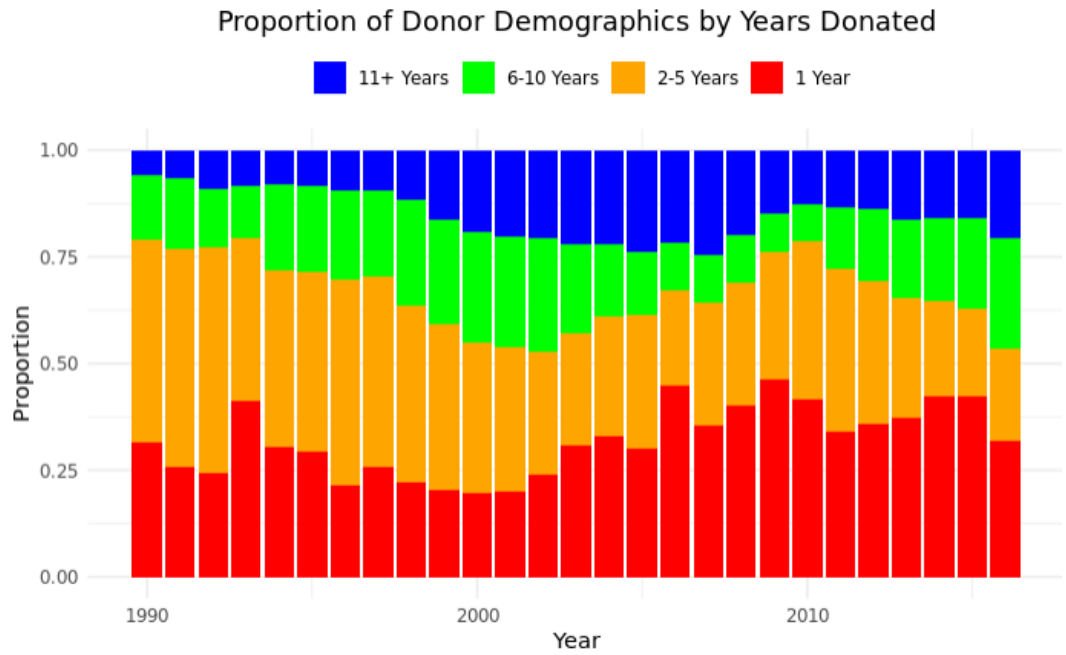
represented the number of donors from each class, the donation amount from each class, and the average donation size from each class. Using this data, we could begin answering some of the questions we were most interested in for this project: are donation amounts increasing or decreasing as time goes on, what type of donors are donating the most, how well are donors being retained over time.

Data Visualization

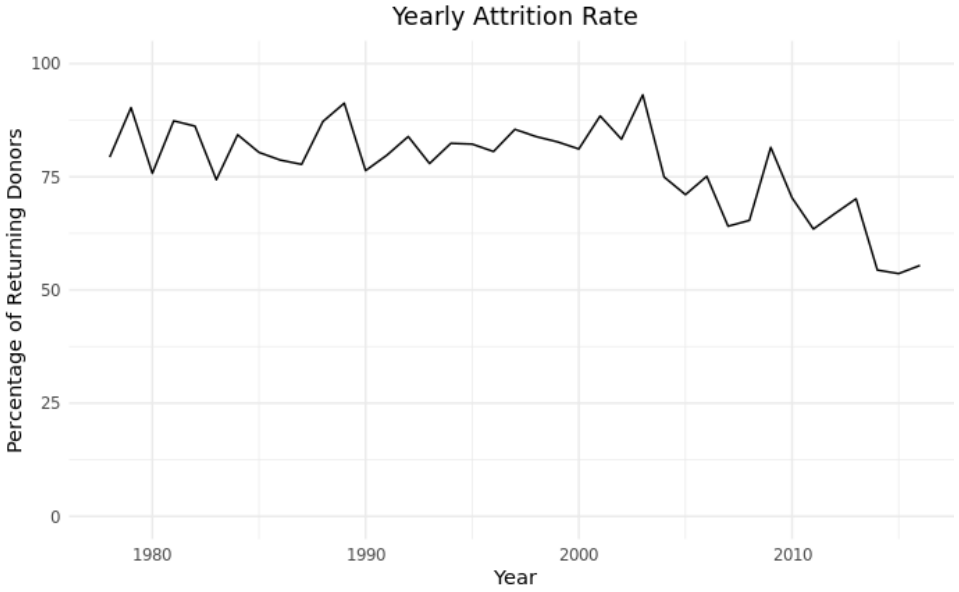
Now that we have done a sufficient amount of data wrangling, we can observe some interesting demographics in the data.



This graph shows the donation amounts over time, but it is categorized by how many years they have been a donor. You can see that there is a clear upward trend in the overall quantity of donations, but within each year we can see where the donations come from. It is clear that most of the revenue is coming from people who have been donating for over ten years often making up over half of the donations, and much fewer donations made from any of the other groups.

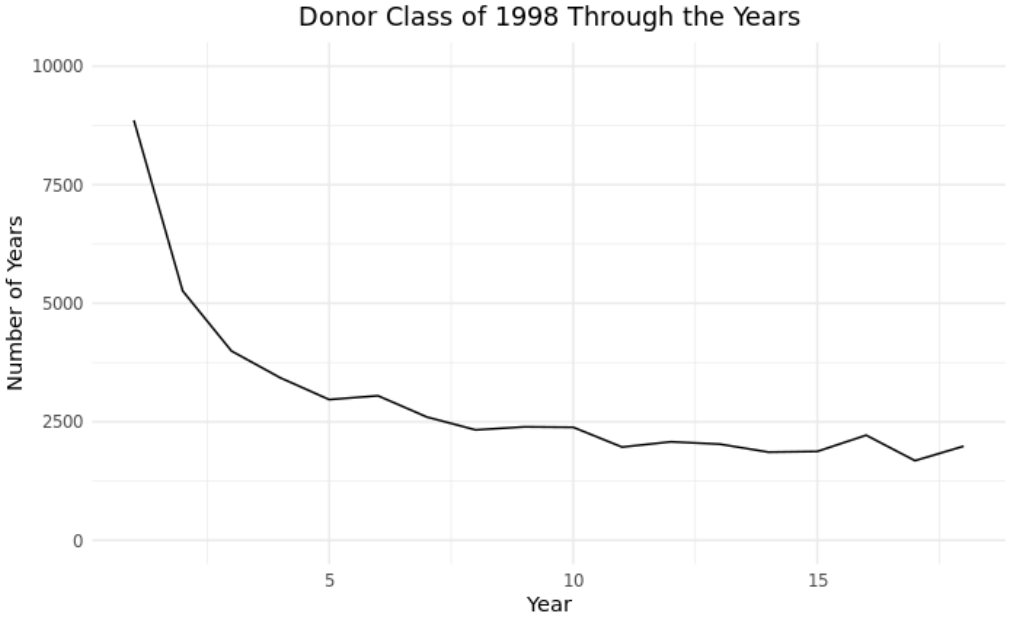


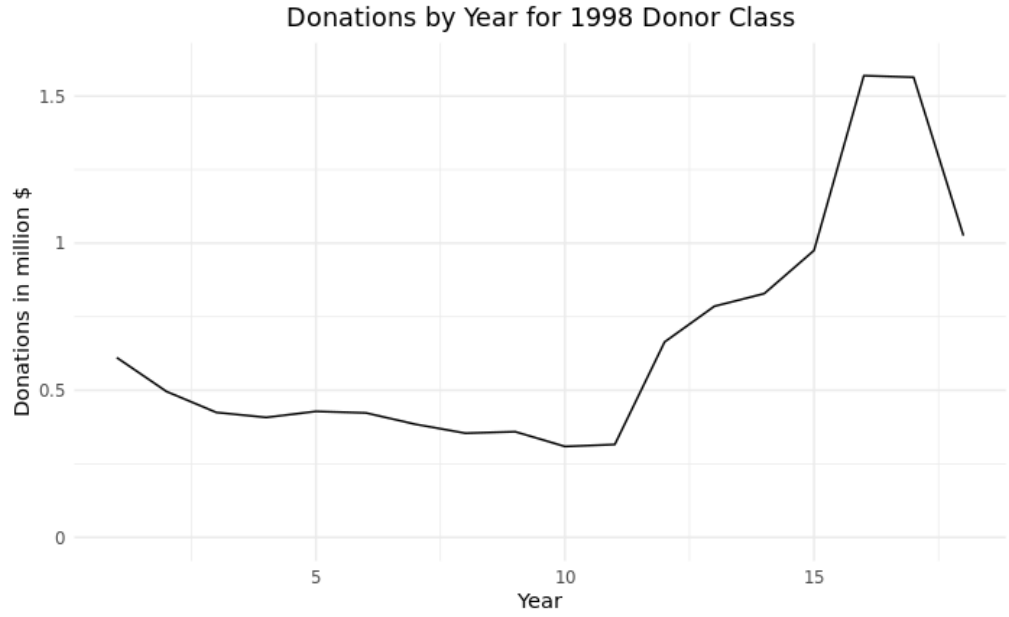
This graph shows the proportion of donors that fits into each category. Here we can see that a large portion of the donors each year are first year donors, and many of the others are in their first 5 years. This tells us that in a given year, many of the first-year donors will not donate again considering how large the proportion of first year donors is each year. When we look at the greater than 10-year donor group each year, it is always a very small group usually around 10-15 percent of total donors. But if we look at the graph before this we know that more than ten year donors are contributing over half of the donations every year. This tells us that the people who have been donating for over ten years often give very large gifts and they do it every year.



This graph shows the yearly attrition percentage. This represents the proportion of donors who were retained from the year before. For the majority of the past 50 years, the

attrition has remained fairly high, usually around 80 percent donor returning rate. However, there seems to be a shift that begins in about 2004 when fewer and fewer people are returning each year. Interestingly, we can see in the proportions graph that there is a huge decrease in the proportion of over ten-year donors in the late 2000's. Considering the economic circumstances of that time, we could reasonably conclude that there is a correlation. During the housing crisis in 2008 and the following years, some of their more reliable donors were donating less frequently, so the attrition fell.





These two graphs took some additional data wrangling to work properly, but they show us a very clear picture of the different types of donor patterns. In order to create these graphs we had to mess with the year_breakdown dataset a bit more. We used the lag function to mutate a column to shift the values up or down any number of rows. Using this function the

proper number of times for each column and then filtering by the year 1998 (randomly selected), we were able to get the values for the same group of people to follow their trend of giving throughout the years. As we saw before, the first year had a significantly higher amount of donors, and then over time the number slowly decreased. We can see a similar trend in the donations graph. But then around ten years in there is a large increase in donation amounts. General trends we see from these graphs are that some people gave only one year of gifts likely varying largely in size. Some people decided to stop giving altogether after a few years, some were inconsistent givers who gave average sized gifts. And some continued giving long term. Of those who gave long term, several of them likely started to have a lot of extra income after enough years of working, and others likely continued giving the same amounts.

Model Input Design

After completing an in-depth analysis of the data, the focus turned toward developing a dataset that we would be able to use for the model analysis. We decided our desired output would be a donor database (each row representing a donor) and include any and all possible predictors that we had used in our data assessment. This brought us back to the GiftAll table which had almost everything we needed. Then we perform a join on GiftAll with bio_db to include some donor information—age, state, gender, etc. Then we group by ID again and add a couple more useful predictors. Then we simply select from this table the columns we want to use in our modeling. The columns we ended up with were: NewID, TotalYearsGiven, TotalDonations, FirstYear (Year they began giving), LastYear, MultiYearGiver, DecemberGivings (proportion of their gifts that were given in December), YearsSinceLastGift (2016 – LastYear), frequency_score (how often they give), AverageDonationPerYear (TotalDonations / TotalYearsGiven), Gender, Age, State, and AgeGroup. Finally we took our random sample of 2000 rows to use for our model testing.

Design Norms

There are three design norms that are most pertinent to our project they are justice, trust, and caring. The first being justice, our application is built to bring people who don't have the coding background or education complex modeling techniques so they can better answer questions about their dataset and draw conclusions by using our application. The second is trust, our application was designed to be reliable as well as being usable in many different settings the trust norm was implemented by account for different datasets and modeling needs. Additionally, it was built with error handling to allow for a reliable experience when using the application. Finally, the caring norm was implemented, our application was not only built to accommodate many different users' needs but additionally it was designed to help people better understand easily why we use certain models. This was implemented in our application using a help button, which when clicked gives details on each modeling technique as well as what different numbers that are produced mean.

Developmental Approach

When building the model and wrangling the dataset we instituted multiple development approaches. We maintained an asana board to help us keep track of what we were working on and future work that needed to be done. This allowed us to easily see our progress and what

was still needing work. This allowed us to use agile programming where we had iterative development allowing us to create a working application quickly and then continue to build off it through continuous integration. Additionally, to keep on track and build communication throughout the project we had weekly meetings with our advisor, where we had an organized meeting. We would start by showing our progress and answer any questions our advisor had and then we were given the opportunity to ask our advisor questions to help us keep progressing.

Problems Encountered

While working on our project there were many problems encountered. The first being the dataset was mostly raw so there were many problems with wrangling the data and figuring out our vision for how we wanted the dataset to look. This problem persisted throughout the first semester and a lot of time was sunk into this problem. In addition to this neither of us had any experience working with financial data and this turned out to be a big hindrance as to answer the questions about the dataset there were many financial terms and practices that are common when addressing this issue that we were unaware of. To overcome this problem, we luckily had Carl McQueen our advisor to help answer these questions. Carl made himself readily available throughout our project and this resource was utilized a lot to accomplish our goals as a team. The final major problem we had when instituting our project was our unfamiliarity with shiny. Both of us had never made an interactive dashboard for an end user, therefore there was a major learning curve when creating a suitable dashboard. This problem was overcome through extensive documentation searching as well as looking at previous examples on the internet of how people implemented certain features of their respective dashboards.

Testing

We did not utilize user testing when creating our project however we did internal testing and made sure to use various datasets to ensure our dashboard worked for all its modeling needs. We utilized three different datasets to ensure the dashboard worked when using a dataset other than the one Carl provided for us. Additionally, we chose different response variables that were either numeric, binary, or categorical to ensure that each of the five models worked when given the correct datatype.

Demo

The demo below shows the dashboard being used with the dataset being created on to predict the total number of donations someone gives in their lifetime. This is essential when looking at nonprofits because if you can focus marketing campaigns on these potential targets then you would be able to maximize donations from that user. In the demo we start off since it is a numeric variable by using a linear regression, then xgboost and finally showing the model summary for the random forest regression. Then to show that the model runs on other datasets we pull in our housing dataset which was used for testing and show a linear regression of this model. Then we move onto the time series analysis which uses our nonprofit dataset by monthly data and can show a prediction for the future, how much the nonprofit can expect to make and the cycle boxplot which highlights that December consistently gets more donations,

which is helpful for nonprofits to understand when they should expect to receive the most money.

Possible Future Work

In the future there are multiple things that would be ideal to add, Carl has expressed his interest in using the dashboard for some other projects in the future so these additions may be implemented in the future. The two main things that would have been nice to implement had we more time would be first the implementation of a data wrangling tab. It would have been ideal to have another page where you could input a dataset and then wrangle this dataset by giving the user options to remove NA's, change the variable from a string type to categorical or string to numeric. This work currently must be done separately in a different r script. The second thing that would be implemented is to be able to modify and adapt each model parameters. Currently the dashboard uses each model's default parameters which gives a good basis, but to get the best performance it would be essential to have model tuning implemented. This would look different for each model and would be a significant use of time but would make the dashboard even more adaptable to different needs.

Acknowledgements

In working on this project for the year our advisor Carl McQueen was immensely helpful in making sure this project was a success. He sunk many hours in our weekly meetings or answering various questions we had, as well as providing us with a space to go to work on our project. Without him this project would not have happened and we are grateful for this.