Who Will Donate?

For this project, the dataset was gathered from Washington and Lee University Office of Development. This department deals with the annual fund and contacting alumni for possible donations. We collected this data with the intent to find which “type” of alumni is most likely to donate to W&L. The “type” of alumni refers to the features attached to them. The features that we collected were “School Name”, “Class” (Graduation Year), “Majors”(If multiple it was recorded in a separate feature column), “Minors”(Similar to majors), “Academic Activities”, “Academic Honors”, “Activities”, “Varsity Athletics”, “All American”, “Honor Societies”, “UG Academic Honors”, “Age”, “Gender”, “Given?”, “PG?(Pledged to Give)”, “Industry”(Industry the person works in, if multiple were recorded in multiple columns like majors), “Alumni Admissions Program”, “Alumni Board”, “Chapter Volunteers”, and “Reunion Class Committee”. The data excludes the names of the donors and any other identifying information for security reasons. When we received the dataset, we were explicitly told this would only to be used for our project and not public means. This means even if our project was impressive enough to be published, we wouldn’t be allowed to publish the dataset along with the project. The target function of this project is to create the model with the most accurate predictions for what type of alumnus will donate. The training examples come from the dataset previously mentioned, the graduating classes from 1990 to 2010 donating history.

We decided to test 3 different models and learning algorithms and compare them to see the most effective type: SVM, Decision Tree, and Random Forest. The hypothesis will consist of all possible support vectors with their respective margin to divide the data (SVM) or will be a decision tree (Decision Trees) or a large group of decision trees from which the best option is selected (Random Forest). The final hypothesis we gathered was a Random Forest model built around multiple decision trees to classify alumni into “will give” or “will not give”. We chose these three models for these reasons: SVMs are a standard good classifier model to try and easy to implement overall; Decision Trees and Random Forest models are capable of handling a large amount of features relatively well, and we knew we would have a large number of features from the start, so deciding on these was a no brainer.

First, let’s address the EDA file and the initial changes that were made to the dataset. We began by altering all “yes” values in the entire dataset to 1. For all numerical features we imputed the NaN values to 0s. The features unaffected by this were “School”, “Class”, “Age”, “Majors”, “Minors”, and “Industry”. We then decided to look at which features we could drop. We determined as a group that the features “School” and “Age” were useless or redundant. “Age” was misleading due to students possibly taking gap years or going on mission trips, and was also redundant since we also have the “Class” feature. The “School” feature was straightforward as all donors on the list are W&L alumni, so we dropped this feature.

We also looked at the heatmap and the correlation scatterplot matrix for our dataset. This did not reveal much to us, unfortunately; each individual feature had a low correlation with the “Given” column, with the highest correlation being 0.19 for the “Years Since Graduation” feature. The scatterplot matrix did not reveal anything to us, as the data was binary for the most part and a chart with 30 features on it is impossible to accurately read. Because there were no clear outliers within these plots, we did not drop any features because of these.

For the categorical features, namely “Major” and “Minor”, we binned them into 6 different subjects. The bins for these two features are STEM, Business, Arts, Humanities, Social Sciences, and Other. We would add a person’s total values of major or minors to their respective groups. For example, if an alumnus was a double major in Physics and Chemistry but obtained a minor in Art History, they would have a 2 in the STEM group, a 1 in the Arts group, and a 0 in the other 4 groups. There were some outliers present in “Major” and “Minor” features, but we threw those into the “Other” category.

For the “Industries” feature we did something like what we did in “Major” and “Minor”. We binned the feature and created 12 groups or categories. These categories were Legal Services, Investments, Real Estate, Education, Healthcare, Government, Retail, Public Relations, Utilities, Tech, Entertainment, and Other. The values were assigned to these categories like what was done in “Major” and “Minor”. We had a difficult time deciding on how to approach the “Industry” feature. We were unsure if the best method would be if people working multiple jobs in the same industry deserved a higher count than 1 or if it should be 1 for if they’ve ever worked in the industry. The reason we felt torn was the reason for ranking in “Major” and “Minor” was to value people with multiple major or minors in an area to have a higher ranking than those with less. This made us believe that the ranking in “Major” and “Minor” wasn’t the best route but the features themselves did not hold a lot of weight where it would’ve made a difference. Ultimately, we decided to do the same thing we did for Majors and Minors and add one to the score for each industry if they switch around jobs. This only applied to a few alumni, so we felt it did not make a large difference overall.

After adjusting the dataset to our liking, we decided on which scalars to use. We tested the Standard, MinMax, MaxAbs, and Robust. The Standard scalar provided us with the best metrics, so we decided to use that.

Finally, we applied the following models to our dataset. We used Random Forest, SVM (Support Vector Machine), and Decision Trees. To determine the correct parameters for each model, we implemented a grid search algorithm on each and determined the best parameters for each type of model. We also looked into sequential backward selection to determine the best amount of features for each model, but didn’t find that dropping any features resulted in a large metric improvement. We kept each feature in because of this.

Random Forest outperformed the others slightly, but SVM came in a close second with Decisions Trees in third. The metrics for each model are as follows: For Random Forest, a 10-fold cross validation yielded an accuracy of 0.929, a precision of 0.935, a recall of 0.929, and an f1 of 0.929. For SVM, the 10-fold cross validation yielded an accuracy of 0.919, a precision of 0.926, a recall of 0.919, and a CV of 0.919. For Decision Tree, the 10-fold cross validation yielded an accuracy of 0.910, a precision of 0.916, a recall of 0.910, and an f1 of 0.909.

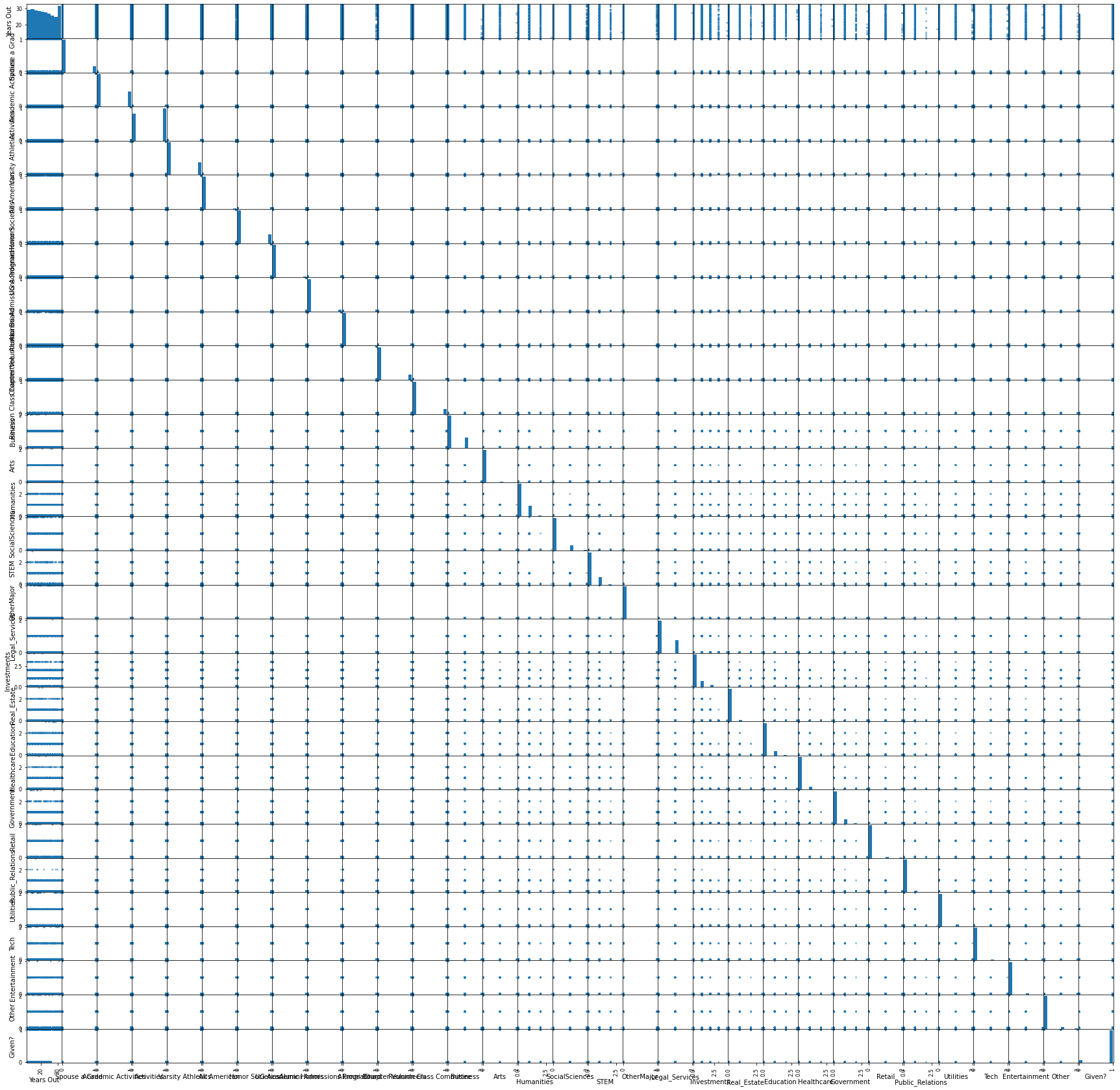


Figure : Correlation scatterplot matrix for dataset

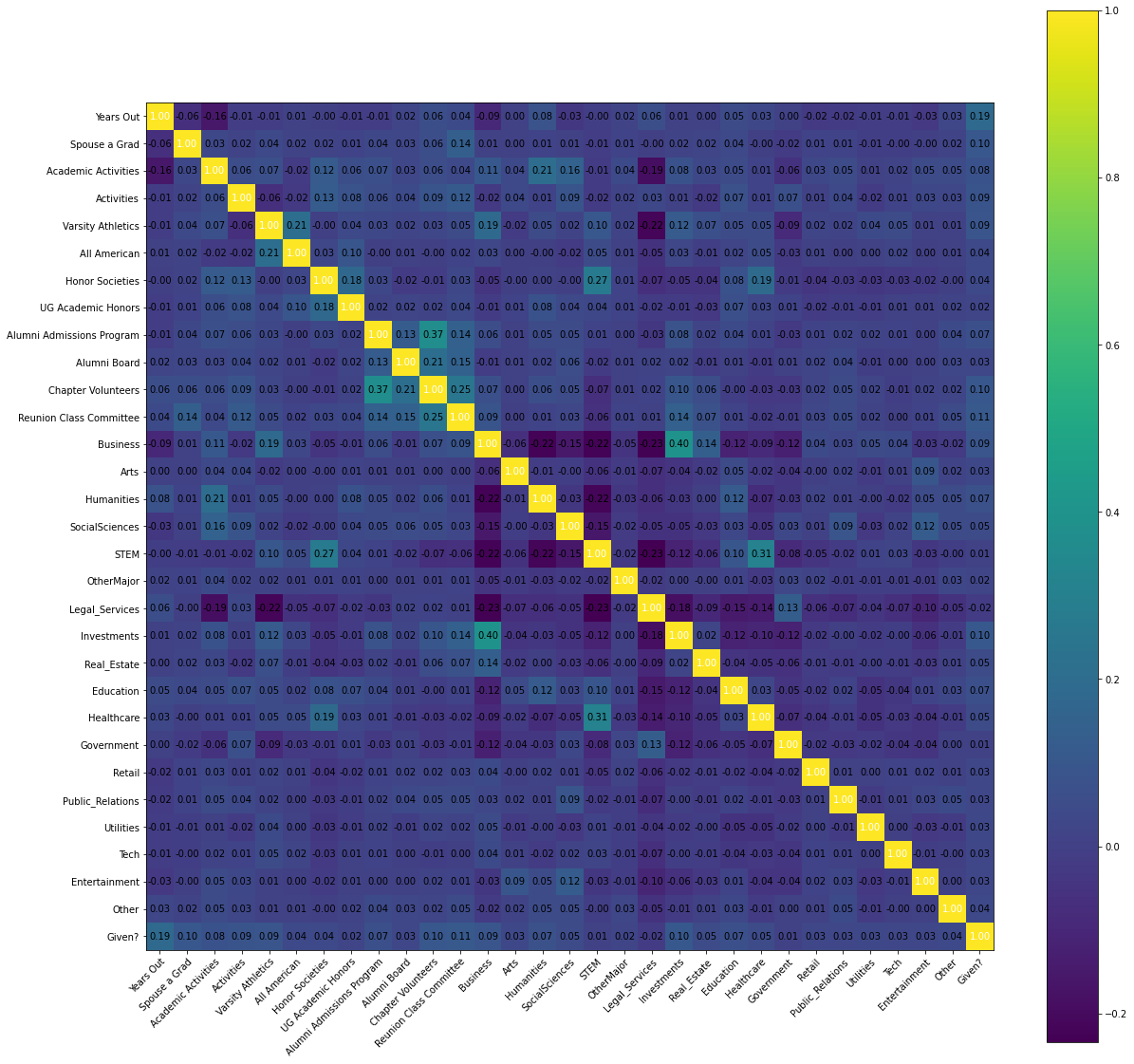


Figure : correlation heatmap for dataset

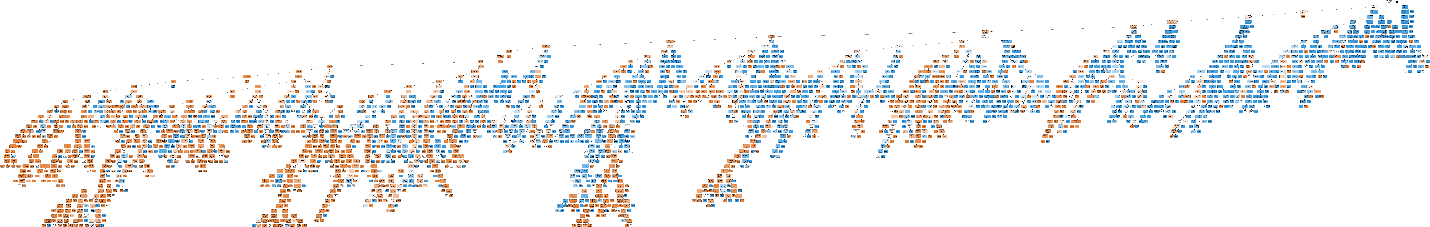


Figure : graphical representation of decision tree created