Optimizing our outreach: increasing donor retention using predictive modeling

William Tran

Summary

In regards to our upcoming fundraising campaign, it is of utmost importance that we utilize our outreach resources as efficiently as possible. More specifically, I am interested in **donor retention**; maximizing the amount of returning donors that we gain would ultimately increase the funds collected from each successive year while lowering the volatility of future campaigns. The question then becomes: *How can we increase the amount of long-term donors that we obtain in this campaign?*

In this report I analyze and interpret the donor dataset provided to produce a machine learning model predicting the likelihood of a first-time donor returning to contribute to our organization in their second year. From these results, **I recommend that the Executive Staff focuses our outreach efforts on the three following demographics *in order of priority*:**

1. Individuals in the high income range.
2. Individuals living in metropolitan areas.
3. Individuals that are above the age of 40.

More details on statistical methods and results can be found below, as well as a conclusion containing ideas about going further with this project. All code used in this report was written in Python and can be found in the appendix at the end of the document.

Methods

I began the initial data exploration phase using a *pivot table* to gleam broad insights on the highest-contributing groups of donors based on all metrics available in the dataset, these being homeownership, marriage status, gender, income range, age range, and area of residence. From there, I generated *bar plots* visualizing various relevant analyses such as the amount of donation renewals based on income range and homeownership. Finally I created a *logistic regression model* to predict the likelihood of the "t2_renewal" column, which contains records of whether a given donor contributed to the year after their first donation year. Using all metrics available, I arrived at a very acceptable base level precision of 73%.

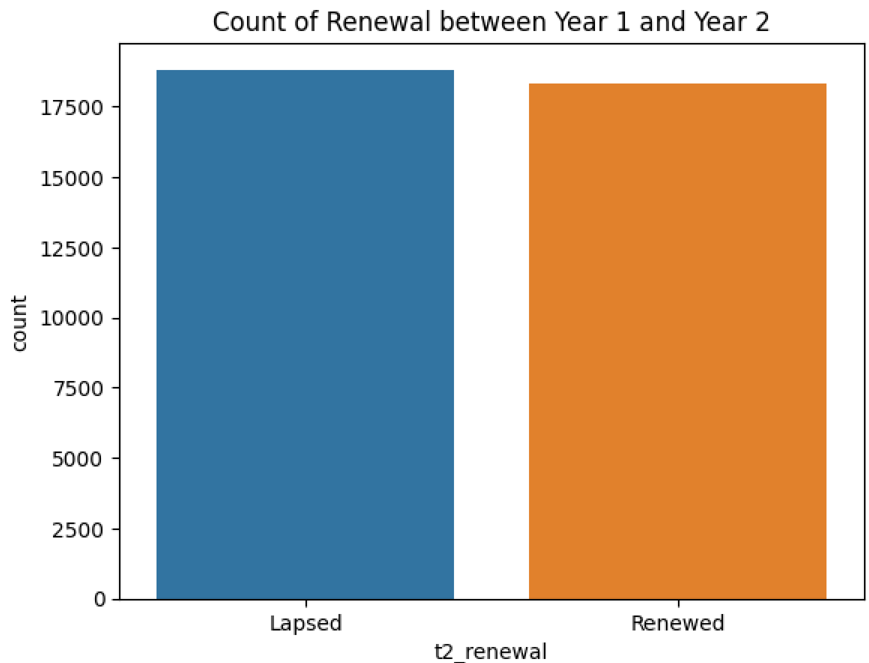
Results

Initial exploration revealed that married female homeowners aged 40-59 in the high income range living in metropolitan areas renew their donations for the second-year the most. This is reflected in the pivot table below:

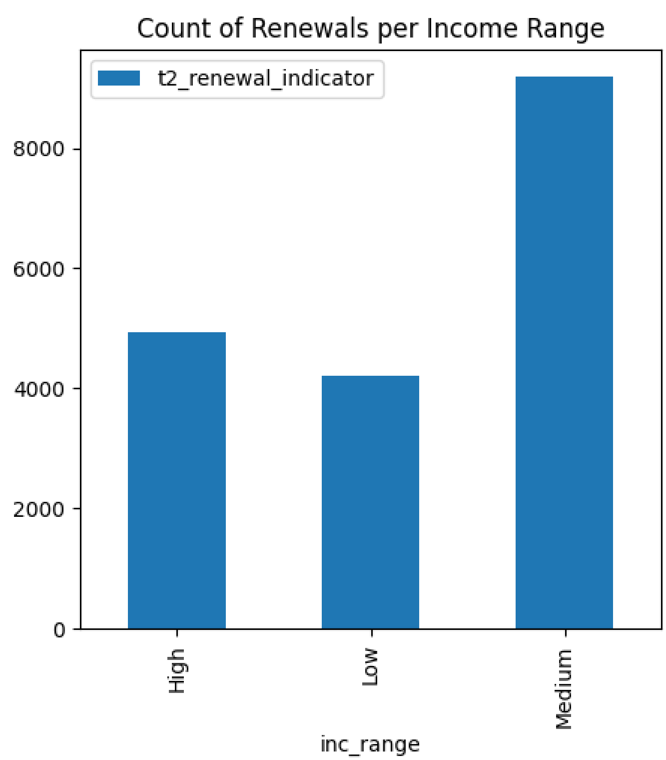
						t2_renewal_indicator
homeowner	married	inc_range	age_range	metropolitan_area	gender	
Owner	Married	High	40_to_59	Metropolitan area	Female	977
					Male	996
					Non-binary/Additional responses	43
				Rural	Female	13
					Male	4
...	...	...	...	...	...	...
Renter	Unmarried	Medium	Under_40	Metropolitan area	Male	29
					Non-binary/Additional responses	0
				Rural	Female	4
					Male	1
					Non-binary/Additional responses	1

While focusing our attention on this hyper-specific subset may appear to increase retention, it will ultimately result in a reduction of total funds collected due to the amount of people excluded from the campaign. To account for this, I opted to create a machine learning model that predicts donor retention based on all available metrics in order to determine which metrics affect retention the most.

In preparation for the construction of the model, some further information was needed. I found that there were slightly more lapses than renewals between year 1 and year 2.

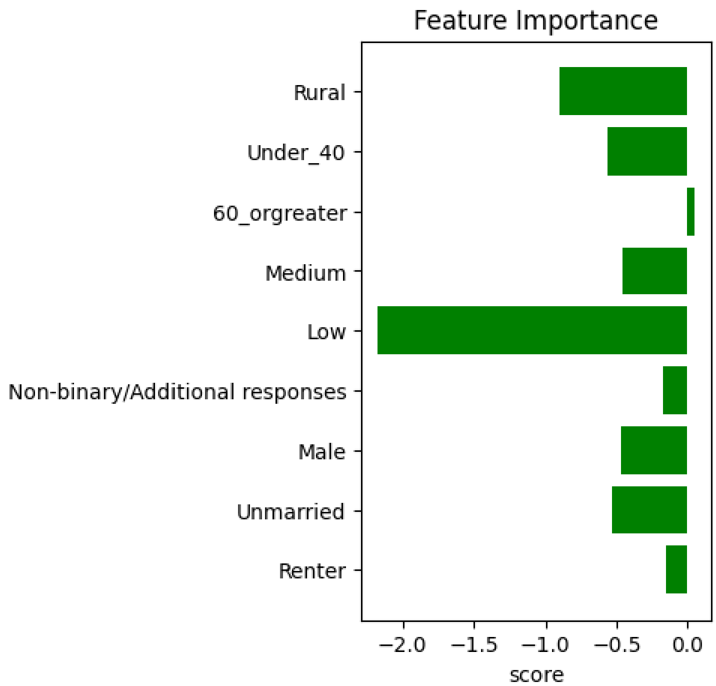


I also visualized the effect of certain metrics I initially thought were the most significant. Something interesting I noted here was that while the pivot table showed that individuals in the high income range renewed their donations the most, the bar plot shows that individuals in the middle income range renewed their donations more than any other range. This is most likely due to the middle income range possessing a higher amount of donors.



Finally, after constructing my predictive model I arrived at a relatively high precision level of 73% with an overall AUC score of 0.756, indicating high usability. The ROC curve for this model can be found in the appendix section.

A more interesting visualization to note is that of feature importance. The chart below can be interpreted as a list of the most significant metrics in predicting donor retention. Metrics leaning farther to the left (in the negative direction) indicate their respective weight in labeling a donor as lapsing. For example, the most significant factor here seems to be income range, with low income individuals being the least likely to renew their donations. This makes sense, as low income individuals would most likely not have room in their personal budgets for donations. My recommendation for the Executive Staff comes from choosing the opposites of the top three negative



Conclusion

The plot above (Feature Importance) is a perfect explanation for my recommendations for the Executive Staff. As stated in the summary, I would advise the executives to focus outreach efforts on the following demographics *in order of priority:*

1. Individuals in the high income range.
2. Individuals living in metropolitan areas.
3. Individuals that are above the age of 40.

Refocusing the scope of the upcoming fundraiser using these three filters will allow the department to reduce effort spent on individuals that will likely not continue donating to us after their first year, significantly increasing the longevity of new donor activity while reducing our expenses.

Going Further

Given more time, my predictive model could be further refined through various technical methods such as cross-validation to increase its accuracy. This would allow the organization to gain unprecedented insight on future donors without any additional cost.

An interesting direction to which this method can lend itself is **location analysis**. Given a donor's latitude and longitude (already included in the dataset), a similar model could be produced to predict *giving interest* (the subject area that a donor is inclined to contribute towards). As a brief window into this analysis I quickly visualized the distribution of total funds over the recorded 10 years given to each different interest, revealing religion as the highest earning field (found in the appendix). Completing this analysis would allow the organization to target different geographic areas with specialized campaigns, maximizing the productivity of outreach efforts while minimizing our work and resources expended.

Appendix

In [1]:
```python
# importing packages
import numpy as np
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from matplotlib import pyplot
```

In [2]:
```python
donors = pd.read_csv('C:/users/Bill/Downloads/dataset.csv') # importing csv dataset
```

In [3]:
```python
donors.head(5) # show first 5 rows
```

Out[3]:

	homeowner	married	gender	inc_range	age_range	metropolitan_area	t1	t2	t3	t4	...	t8	t9	t10	t2_renewal
0	Owner	Unmarried	Female	Low	60_orgreater	Rural	6.0	0.0	0.0	0.0	...	50.0	0.0	0.0	Lapsed
1	Owner	Unmarried	Male	High	Under_40	Rural	NaN	0.0	0.0	0.0	...	0.0	0.0	0.0	Lapsed
2	Renter	Unmarried	Male	High	60_orgreater	Rural	70.0	0.0	250.0	0.0	...	0.0	0.0	0.0	Lapsed
3	Renter	Unmarried	Male	High	60_orgreater	Rural	1000.0	0.0	0.0	0.0	...	0.0	0.0	400.0	Lapsed
4	Owner	Married	Female	Medium	40_to_59	Rural	50.0	100.0	0.0	0.0	...	0.0	0.0	0.0	Renewed

5 rows × 23 columns

In [4]:
```python
donors.isna().sum() / donors.count() # verifying ratio of missing values in each column;
                                     # every column has 3% missing data

df = donors.dropna().reset_index(drop=True) # remove rows with missing values,
                                            # only analyzing complete records

df['t2_renewal_indicator'] = df['t2_renewal'].apply(lambda cell: 1 if cell == 'Renewed' else 0)
df.head(5)
```

Out[4]:

	homeowner	married	gender	inc_range	age_range	metropolitan_area	t1	t2	t3	t4	...	t9	t10	t2_renewal	zipco
0	Owner	Unmarried	Female	Low	60_orgreater	Rural	6.0	0.0	0.0	0.0	...	0.0	0.0	Lapsed	85387
1	Renter	Unmarried	Male	High	60_orgreater	Rural	70.0	0.0	250.0	0.0	...	0.0	0.0	Lapsed	45345
2	Owner	Married	Female	Medium	40_to_59	Rural	50.0	100.0	0.0	0.0	...	0.0	0.0	Renewed	67654
3	Renter	Unmarried	Male	High	60_orgreater	Rural	1000.0	0.0	0.0	0.0	...	0.0	300.0	Lapsed	1451
4	Owner	Unmarried	Male	Low	60_orgreater	Rural	50.0	0.0	0.0	0.0	...	0.0	0.0	Lapsed	4062

5 rows × 24 columns

In [5]:
```python
# initial exploration using pivot table
pd.pivot_table(df, index=['homeowner', 'married', 'inc_range',
                          'age_range', 'metropolitan_area', 'gender'],
               values='t2_renewal_indicator', aggfunc=np.sum)
```

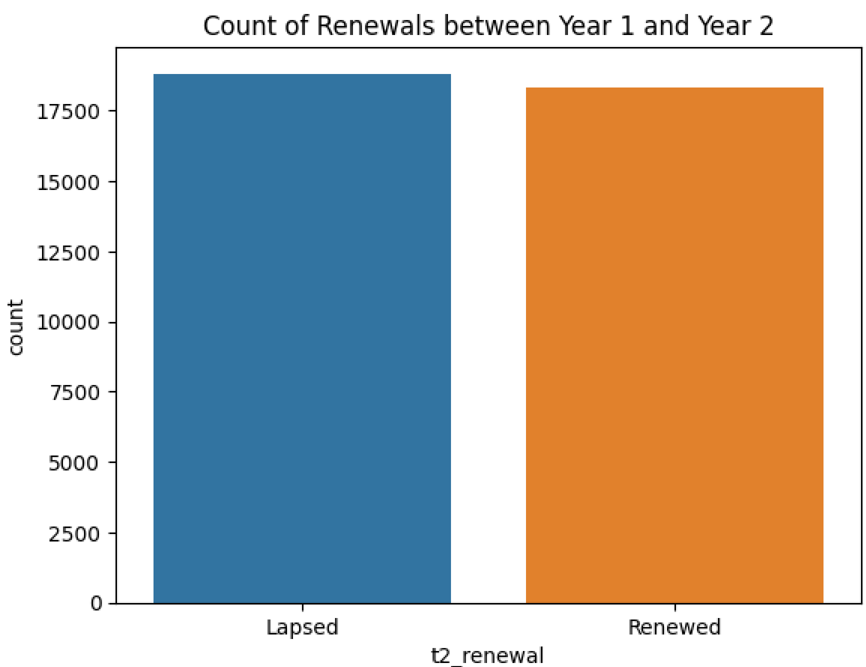
Out[5]:

						t2_renewal_indicator
homeowner	**married**	**inc_range**	**age_range**	**metropolitan_area**	**gender**	
Owner	**Married**	**High**	**40_to_59**	**Metropolitan area**	**Female**	977
					Male	996
					Non-binary/Additional responses	43
				Rural	**Female**	13
					Male	4
...	**...**	**...**	**...**	**...**	**...**	...
Renter	**Unmarried**	**Medium**	**Under_40**	**Metropolitan area**	**Male**	29
					Non-binary/Additional responses	0
				Rural	**Female**	4
					Male	1
					Non-binary/Additional responses	1

196 rows × 1 columns

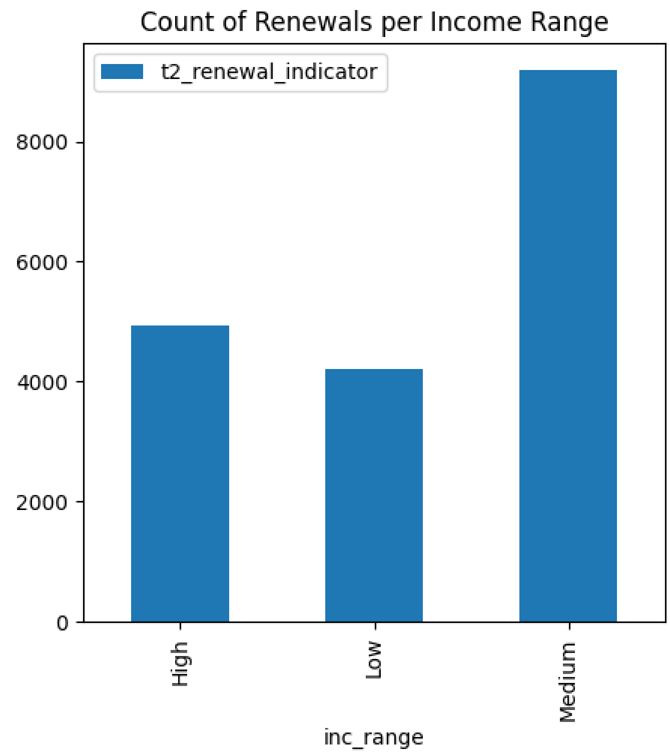
In [6]:
```python
# counting renewals: slightly more lapses than renewals,
# but overall negligible difference
sns.countplot(x = 't2_renewal', data = df
             ).set(title = 'Count of Renewals between Year 1 and Year 2')
```

Out[6]: `[Text(0.5, 1.0, 'Count of Renewals between Year 1 and Year 2')]`



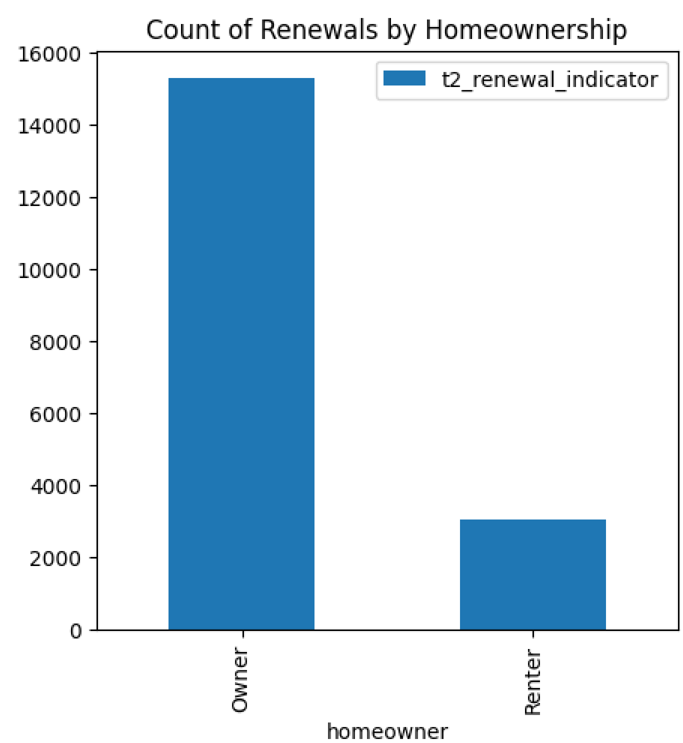
In [7]:
```
# visualizing effect of income range on renewals
pyplot.rcParams["figure.figsize"] = (5,5)
pd.pivot_table(df, index=['inc_range'],
               values='t2_renewal_indicator', aggfunc=np.sum
              ).plot(kind='bar').set(title = 'Count of Renewals per Income Range')
```

Out[7]:
```
[Text(0.5, 1.0, 'Count of Renewals per Income Range')]
```



In [8]:
```
# visualizing effect of homeownership on renewals
pd.pivot_table(df, index=['homeowner'],
               values='t2_renewal_indicator', aggfunc=np.sum
              ).plot(kind='bar').set(title = 'Count of Renewals by Homeownership')
```

Out[8]:
```
[Text(0.5, 1.0, 'Count of Renewals by Homeownership')]
```



In [9]:
```
# building logistic regression model for t2 renewal prediction
# creating dummy variables
homeowner = pd.get_dummies(df['homeowner'], drop_first=True)
married = pd.get_dummies(df['married'], drop_first=True)
gender = pd.get_dummies(df['gender'], drop_first=True)
inc_range = pd.get_dummies(df['inc_range'], drop_first=True)
age_range = pd.get_dummies(df['age_range'], drop_first=True)
metropolitan_area = pd.get_dummies(df['metropolitan_area'], drop_first=True)
df2 = pd.concat([df['t2_renewal_indicator'],
                 homeowner, married, gender,
                 inc_range, age_range, metropolitan_area], axis = 1)

print(df2.columns)
```

```
Index(['t2_renewal_indicator', 'Renter', 'Unmarried', 'Male',
       'Non-binary/Additional responses', 'Low', 'Medium', '60_orgreater',
       'Under_40', 'Rural'],
      dtype='object')
```

In [10]:
```
labels = pd.DataFrame(df2['t2_renewal_indicator'])
df2 = df2.drop(['t2_renewal_indicator'], axis=1)
df2 = df2.apply(pd.to_numeric)
labels = labels.apply(pd.to_numeric)
```

In [11]:
```
# checking dummy variables
df2.head(5)
```

Out[11]:

	Renter	Unmarried	Male	Non-binary/Additional responses	Low	Medium	60_orgreater	Under_40	Rural
0	0	1	0	0	1	0	1	0	1
1	1	1	1	0	0	0	1	0	1
2	0	0	0	0	0	1	0	0	1
3	1	1	1	0	0	0	1	0	1
4	0	1	1	0	1	0	1	0	1

In [12]:
```
# splitting data
x_train, x_test, y_train, y_test = train_test_split(df2, labels,
                                                    test_size = 0.3,
                                                    random_state = 0)
```

In [13]:
```
# Fitting logistic regression model
df2_lr = LogisticRegression()
df2_lr.fit(x_train, y_train.values.ravel())
predictions = df2_lr.predict(x_test)
```

In [14]:
```
# We are able to predict with a 73% precision rate whether
# a donor will renew or lapse between year 1 and year 2.
print(classification_report(y_test, predictions))
```

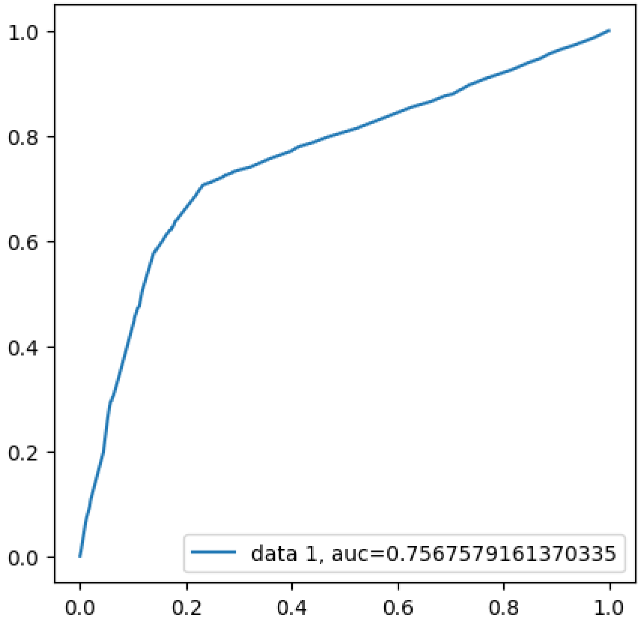
```
              precision    recall  f1-score   support

           0       0.73      0.73      0.73      5662
           1       0.72      0.72      0.72      5480

    accuracy                           0.73     11142
   macro avg       0.73      0.73      0.73     11142
weighted avg       0.73      0.73      0.73     11142
```

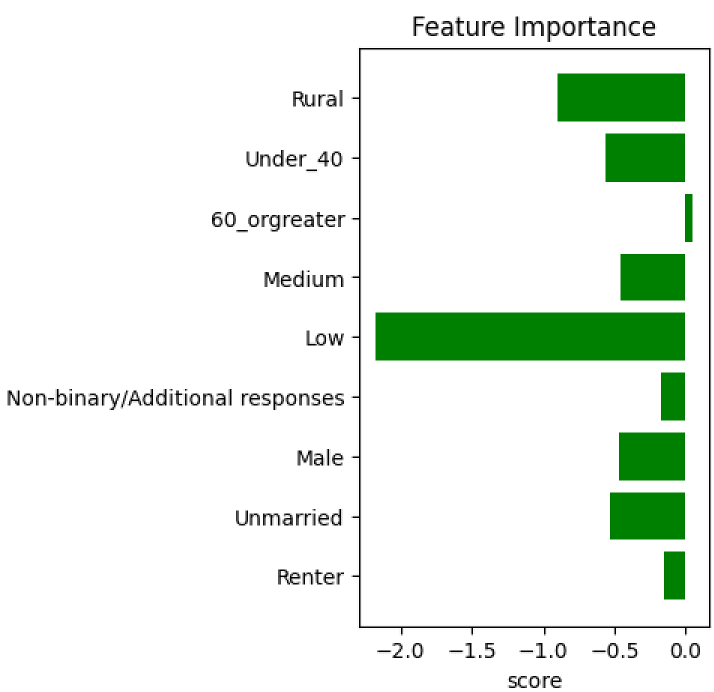
In [15]:
```
pyplot.rcParams["figure.figsize"] = (5,5)
y_pred_proba = df2_lr.predict_proba(x_test)[::,1]
fpr, tpr, _ = roc_curve(y_test,  y_pred_proba)
auc = roc_auc_score(y_test, y_pred_proba)
pyplot.plot(fpr,tpr,label="data 1, auc="+str(auc))
pyplot.legend(loc=4)
pyplot.show()

#AUC score of 0.756
```



In [16]:
```
importance = df2_lr.coef_.flatten()
```

In [17]:
```
pyplot.rcParams["figure.figsize"] = (3,5)
pyplot.barh(df2.columns, importance, color = 'g')
pyplot.title("Feature Importance")
pyplot.xlabel("score")
pyplot.show()
```



In [18]:
```python
# quick visualization of total amount of funds donated
# over 10 years per giving interest category, religion is the clear leader
df['t_total'] = df[['t1','t2','t3', 't4', 't5',
                    't6', 't7', 't8', 't9', 't10']].sum(axis=1)
df[['t_total','giving_interest']].groupby(by='giving_interest'
                                        ).sum().plot.bar().set(title =
                                                               'Total amount given per interest group',
                                                               ylabel =
                                                               'Dollars (one hundred million)')
df[['t_total','giving_interest']].groupby(by='giving_interest').sum()
```

Out[18]:

giving_interest	t_total
Art	14417879.0
Combined_Purposes	56650025.0
Education	22720166.0
Environment	4130223.0
Health	30171612.0
Human_Needs	31286065.0
International	2104519.0
Neighborhood_Giving	26951226.0
Other	11357545.0
Religion	186787727.0
Youth	10700863.0