DATA144 FINAL PROJECT (fall2021): *FINANCIAL ANALYSIS ON GATES FOUNDATION ANNUAL GRANTS*

By: Mingyuan Ma, Yolanda Wang, Angel Xu, Yuqi Ye, Lucy Zhang

RESEARCH TOPIC:
GATES FOUNDATION ANNUAL DONATION

We found dataset of the Gates Foundation annual donation towards education thru Gates Website

OUR WHY

BECAUSE...

- Our interdisciplinary team has backgrounds in: CS, Stats, Finance, Education
- We all care about social goods: education as UN's SDG
- We think it would be challenging but rewarding if this project can help us peak into Gates' Foundation's decision making process with limited information we got (reverse engineer their grant decision matrix)
- The results of our prediction could be helpful to other investment agency, education institutes and education practitioners to gauge where the opportunity lies

OUR GOAL

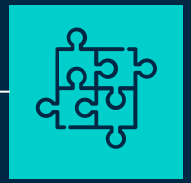
ANALYZE

We want to first analyze the past 10 years of donation data thru methods like TSNE Dimension Reduction, Tableau, geographical analysis and many more.

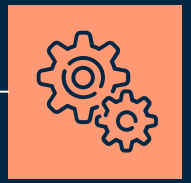
PREDICT

After thoroughly understanding and cleaning the data, we wish to predict and detect the future trends & targets of Gates Foundation Donations

TABLE OF CONTENTS



01 GOAL

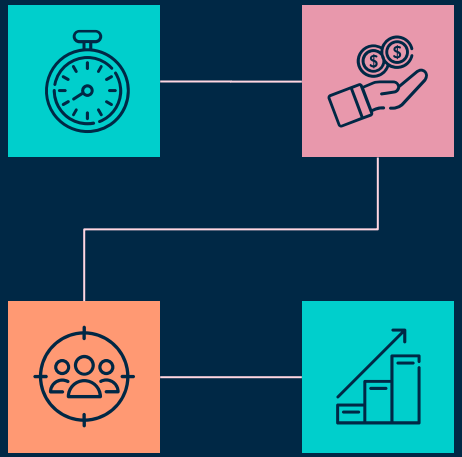


02 OUR PROCESS

03 RESULT

HOW WE DID IT

DATA CLEANING



PRELIMINARY OBSERVATIONS

TABLEAU

ANALYZATION

RANDOM FOREST, ADABOOST TSNE, REGRESSION, etc

PREDICTION

STEPS

Deciding on dataset

—Finding Data

Can we know the grants type from its description?
- Gensim, tsne & clustering

—Data Cleaning & Feature Engineering

Dataframe manipulation, plot, tableau; To inform modeling in the next step

—Preliminary Observation

what will grants be distributed in 2019? (OLS & Neural Network)

—Regression Analysis

which company will be favoured and get multiple donation in the future? (Randomforest, Adaboost)

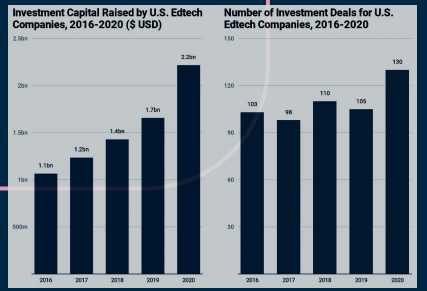
—Exploration Question

Identifying topics & features

—Topic Classifications

CONTEXT

- The $2.2 billion marks the highest investment total in a single year for the U.S. edtech industry.
- Total grant payments since inception (through Q4 2020): $60.1 billion
- Total 2020 Direct Grantee Support: $5.8 billion
- Total 2019 Direct Grantee Support: $5.1 billion
- Foundation Trust Endowment: $49.9 billion
- Over the past 20 years, the Bill and Melinda Gates Foundation has spent $53.8 billion on issues ranging from public health to economic development. Some **16%** of these funds have been spent on the foundation's U.S. programs, which focus on **education.**
- Gates Foundation has funded full college scholarships to 20,000 students of color, Melinda concedes that this is just a small percentage of the tens of millions of students who have attended U.S. public schools since the scholarship's inception 16 years ago.

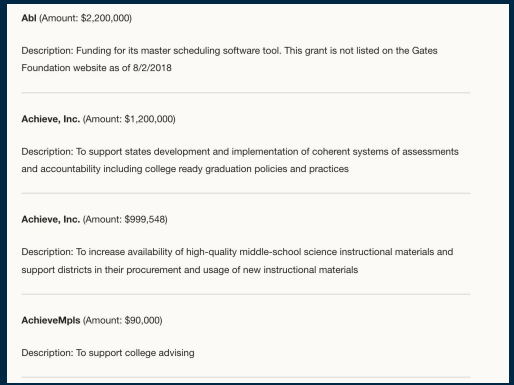


DATA CLEANING 01

DATA COLLECTION

- Original Dataset: (Cr. Hack Education Data/ Github)
- Gates Foundation (education) grants 1998–2018
- Challenge: only few features – name, year, amount, description of grant
- Opportunities: longitudinal analysis of 8 year grant history, textual analysis of description

- New Dataset!: (Cr. Gates Foundation Official)
  - More features: location, topic, purpose etc.



Abl (Amount: $2,200,000)

Description: Funding for its master scheduling software tool. This grant is not listed on the Gates Foundation website as of 8/2/2018

Achieve, Inc. (Amount: $1,200,000)

Description: To support states development and implementation of coherent systems of assessments and accountability including college ready graduation policies and practices

Achieve, Inc. (Amount: $999,548)

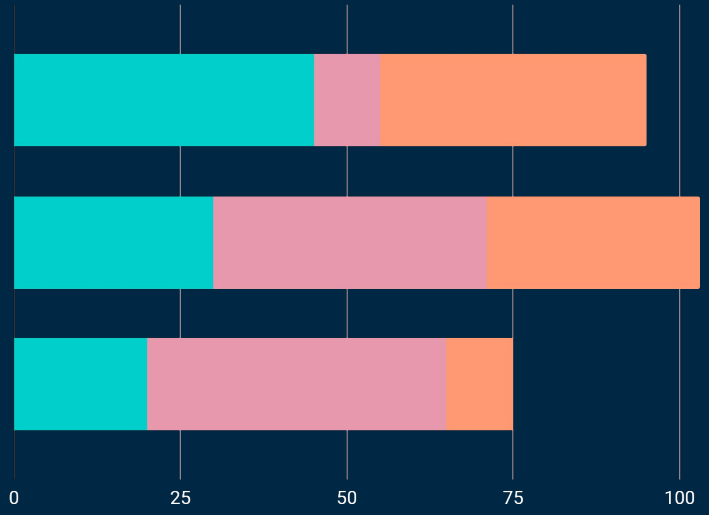
Description: To increase availability of high-quality middle-school science instructional materials and support districts in their procurement and usage of new instructional materials

AchieveMpls (Amount: $90,000)

Description: To support college advising

COMBINE, DROP, & RENAME

- Exploration question: can we know the grants type from its description?
- Gensim, tsne & clustering



COMBINE

We put the data into a dataframe

DROP

We dropped NA columns

RENAME

We renamed unnamed columns to its category (ie. amount)

FEATURE ENGINEERING

Dataframe manipulation

Eliminated

Added new features

Add more features that can inform grant decision:

- # of past investments
- Sum of past investments
- Proportion of annual grant got (avoid inflation)
- Amount of investment / company size (eliminated due to inaccurate data)
- Grant topic/ company type
- Location

Obstacle: There are 33 different types of grants ("Topic"), many of them are repetitive. Even after we manually combined them into 9, they are still highly skewed towards certain categories and not very informative about the grant overall.

Feature engineering (Advanced)

Challenge: There are 33 different types of grants (“Topic”), many of them are repetitive/ overlapping. Even after we manually combined them into 9, they are still highly skewed towards certain categories and not very informative about the grants

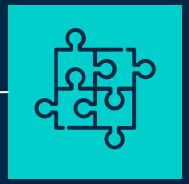
```
z = np.array(['K-12 Education', 'Postsecondary Education',
       'Research and Learning Opportunities', 'Family Planning',
       'Empower Women and Girls',
       'K-12 Education|Research and Learning Opportunities',
       'Global Health and Development Public Awareness and Analysis',
       'Public Awareness and Analysis',
       'Discovery and Translational Sciences',
       'Washington State Education and Human Service Needs',
       'K-12 Education|Postsecondary Education|U.S. Economic Mobility & Opportunity',
       'Global Health and Development Public Awareness and Analysis|K-12 Education',
       'Postsecondary Education|Research and Learning Opportunities',
       'Early Learning', 'K-12 Education|Postsecondary Education',
       'MNCH Discovery and Tools',
       'Pacific Northwest Education and Human Service Needs',
       'Postsecondary Education|Public Awareness and Analysis|Research and Learning Opportunities',
       'K-12 Education|Postsecondary Education|Research and Learning Opportunities',
       'K-12 Education|Postsecondary Education|Public Awareness and Analysis',
       'Financial Services for the Poor',
       'U.S. Education, Poverty, and Family Homelessness Public Awareness and Analysis',
       'Postsecondary Education|Public Awareness and Analysis',
       'K-12 Education|Public Awareness and Analysis',
       'Agricultural Development', 'Water, Sanitation and Hygiene', 'HIV',
       'Global Libraries', 'Community Engagement Grantmaking',
       'U.S. Economic Mobility & Opportunity|Washington State Education and Human Service Needs',
       'Postsecondary Education|U.S. Economic Mobility & Opportunity',
       'Vaccine Development'], dtype=object)
```

```
# logic: pick subtopics that are
more representative and specific
```

Topic	Count
K-12 Education	379
Postsecondary Education	165
Community & Education Service	31
Research & Learning Opportunities	30
Public Awareness & Analysis	25
Global Health & Development	12
Newborn and Child Health & Learning	12
The Underrepresented & Poor	6
U.S. Economic Mobility & Opportunity	1

FEATURE ENGINEERING (Contd)

- *BUT can we get a more informative "Type" feature from grant descriptions?*



SOLUTION

Use NLP and dimension reduction to perform textual analysis on "description" and extract a categorical feature, "Advanced Topic"

Process:

- Reduce redundancy: one-hot & gensim
- Reduce dimension: Word2Vec & TSNE (3D)
- Clustering: Elbow method (5 clusters) & Visualization

FEATURE ENGINEERING (Contd)

Example of Effect

```
[ ] new_df.loc[6,'Description']
```

```
'To support the American Federation of Teachers Innovation Fund and the union's teacher development and evaluation progr
msFor conference support'
```

```
[ ] new_df.loc[6,'Purpose']
```

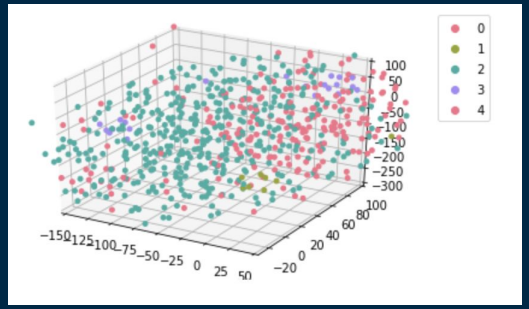
```
'to support the AFT Innovation Fund and work on teacher development and Common Core State Standards'
```

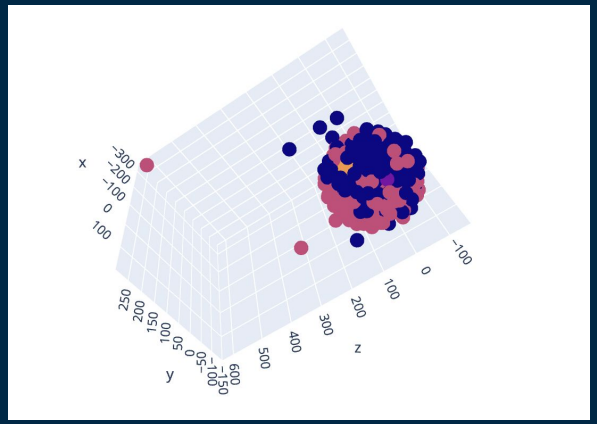
```
[ ] new_df.loc[6,'new_description']
```

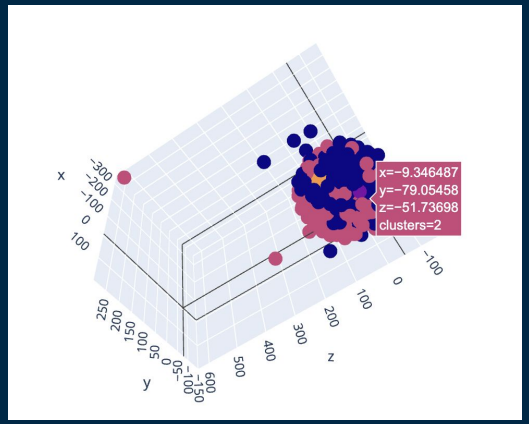
```
'american federation teachers innovation fund teacher evaluation conference work on common core state standards'
```

Feature engineering (Advanced) (conti.)

- Clustering: Elbow method (5 clusters) & Visualization







- Observation: clusters are better in distinguishing each grants → new feature “cluster” (informative “topic”)

PRELIMINARY OBSERVATION

02

What we have so far...

- Name
- Year of investment
- Amount
- Description
- # of past investments
- Sum of past investment amount
- Proportion of annual grant got (avoid inflation)
- Grant topic (manual)
- Location
- Cluster (Grant topic generated by ML)

Original

Data Manipulation

Advanced Feature

PRELIMINARY OBSERVATION

- We use matplot, seaborn and tableau to visualize our data, so as to:
  - Identify some common trends/ changes and analyze Gates' foundation **grant preference**
  - Find **influential features**, inform hyperparameter tuning in the modeling part

DATAFRAME MANIPULATION

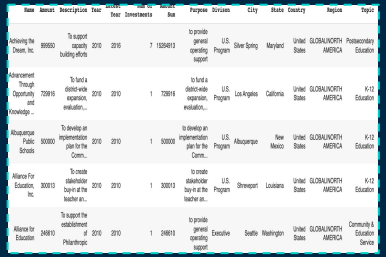
01

02

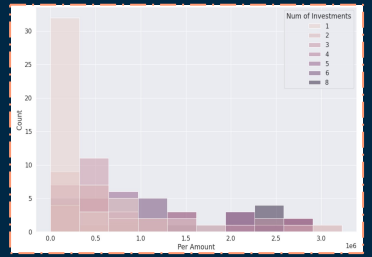
PLOT

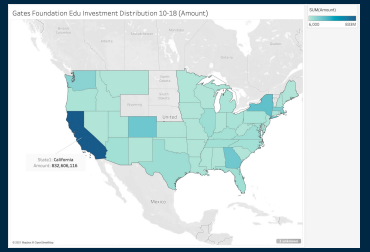
TABLEAU

03



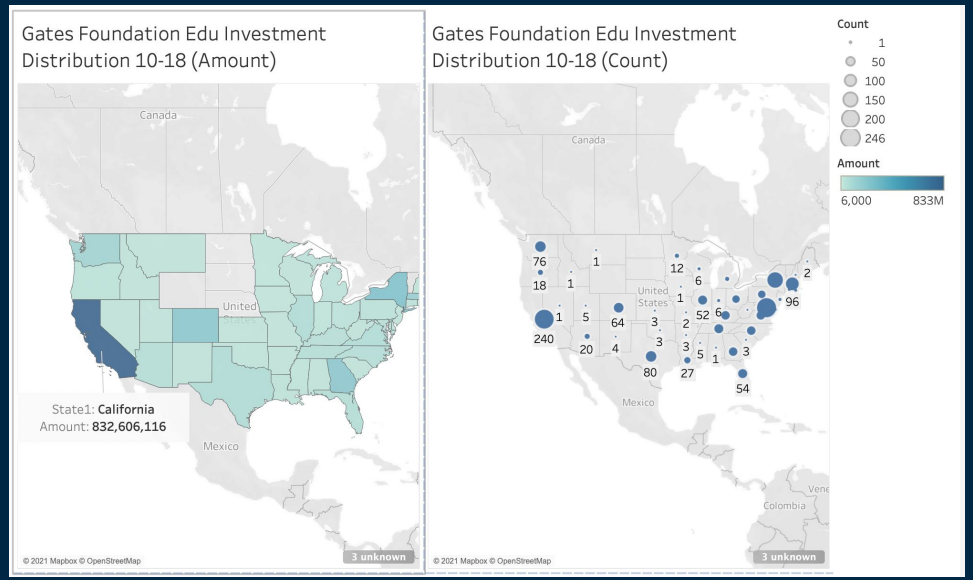
Name	Amount	Description	Year	Earliest Year	Num of Investments	Amount Bin	Purpose	Division	City	State	Country	Region	Topic
Achieving the Dream, Inc	999550	To support capacity building efforts	2010	2010	7	15294813	to provide general operating support	U.S. Program	Silver Spring	Maryland	United States	GLOBAL\NORTH AMERICA	Postsecondary Education
Advancement through Opportunity and Knowledge ...	720816	To fund a district-wide expansion, evaluation...	2010	2010	1	720816	to fund a district-wide expansion, evaluation...	U.S. Program	Los Angeles	California	United States	GLOBAL\NORTH AMERICA	K-12 Education
Albuquerque Public Schools	500000	To develop an implementation plan for the Comm...	2010	2010	1	500000	to develop an implementation plan for the Comm...	U.S. Program	Albuquerque	New Mexico	United States	GLOBAL\NORTH AMERICA	K-12 Education
Alliance For Education, Inc.	300813	To create stakeholder buy-in at the teacher an...	2010	2010	1	300813	to create stakeholder buy-in at the teacher an...	U.S. Program	Shreveport	Louisiana	United States	GLOBAL\NORTH AMERICA	K-12 Education
Alliance for Education	249810	To support the establishment of Philanthropic	2010	2010	1	249810	to provide general operating support	Executive	Seattle	Washington	United States	GLOBAL\NORTH AMERICA	Community & Education Service





OBSERVATION (Geographic)

- Question: How are grants distributed during 10 years?
  - Table manipulation & Tableau



Top 3 states w/ most grant & most #grant:

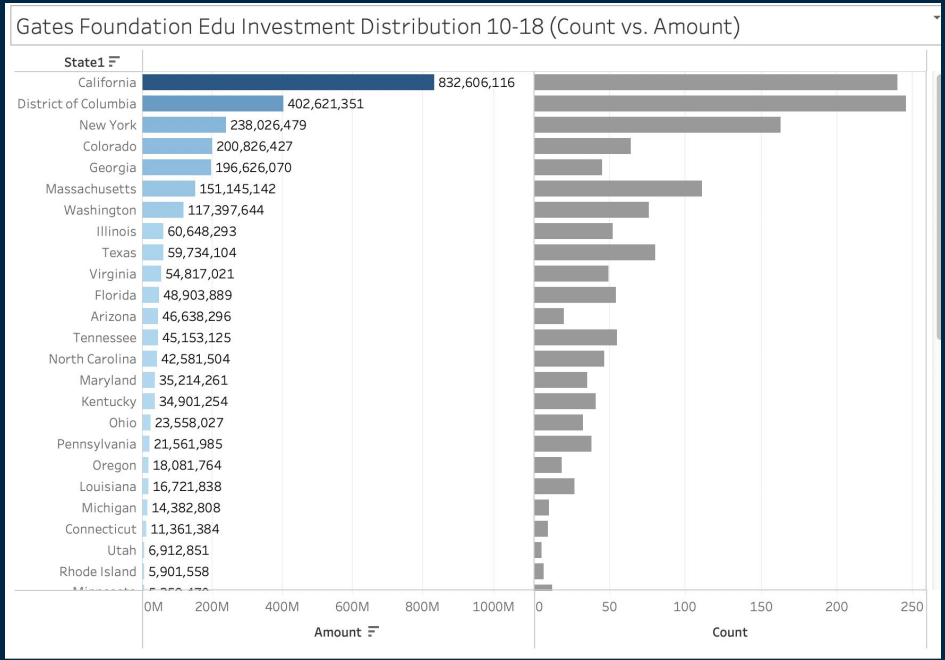
- CA, District of Columbia, NY

Followed by:

- Colorado, Georgia, Massachusetts

Coasts are more favored, some central states don't have any

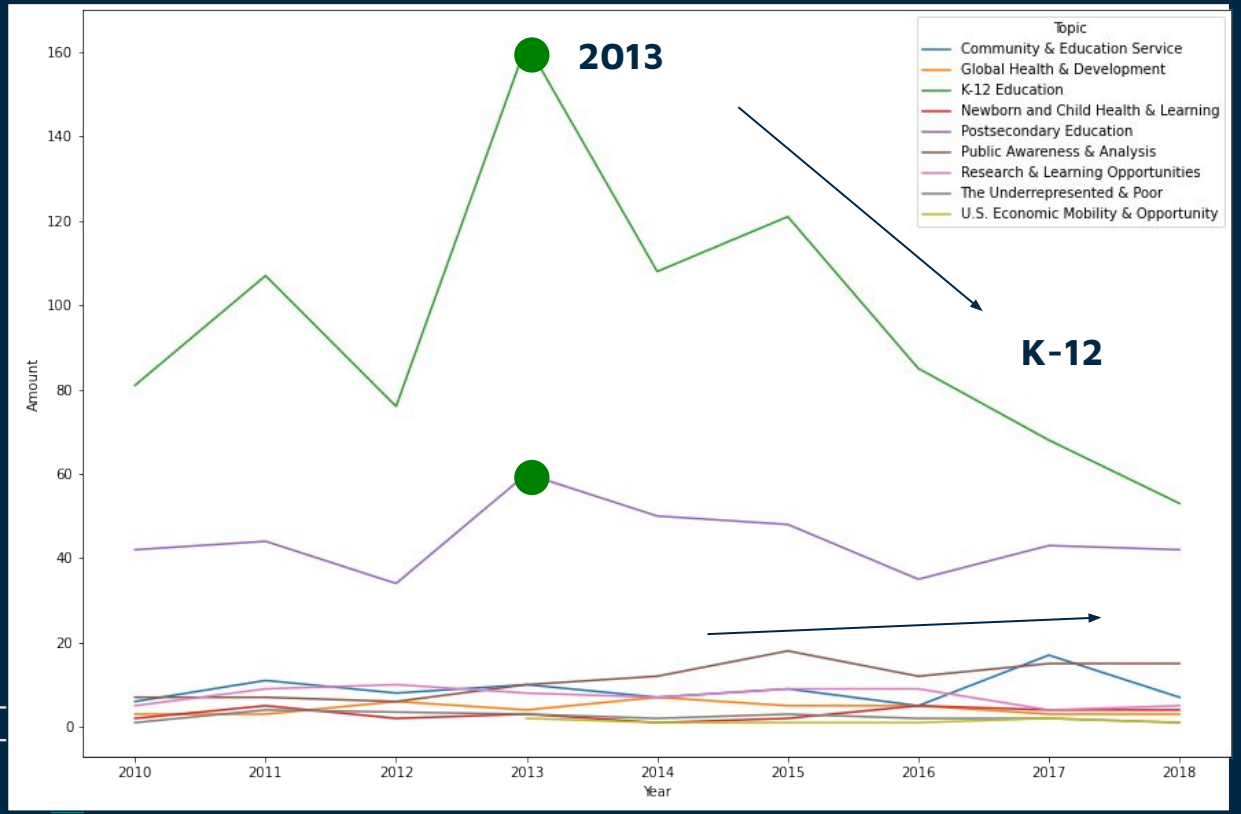
OBSERVATION (Geographic)



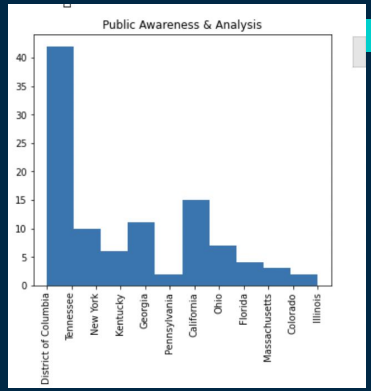
Most surprising:

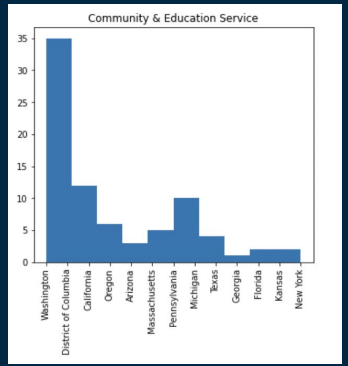
- With the similar #grant as District of Columbia, California got 2x grants
- Disparity

OBSERVATION (Topic)



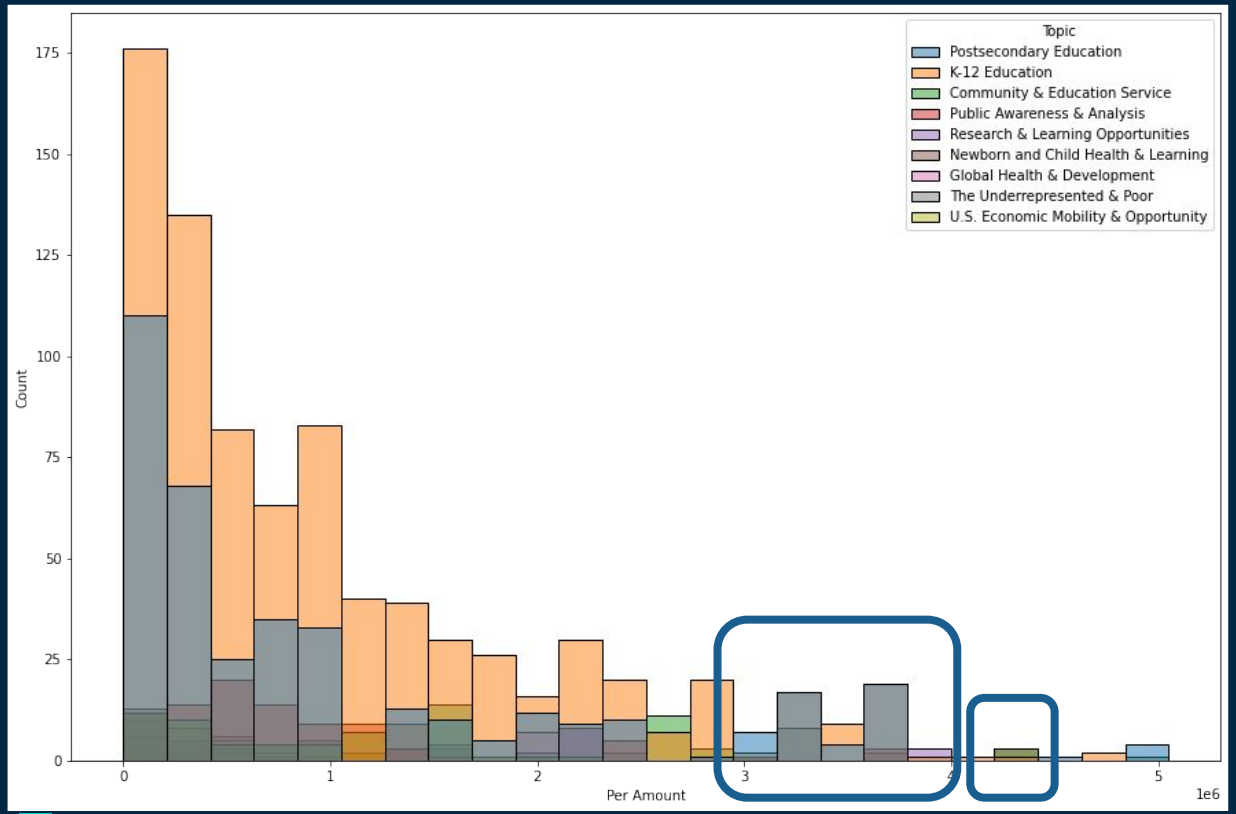
Public Awareness & Analysis





Community & Education Service

OBSERVATION (Topic)



K-12

- High number of grant

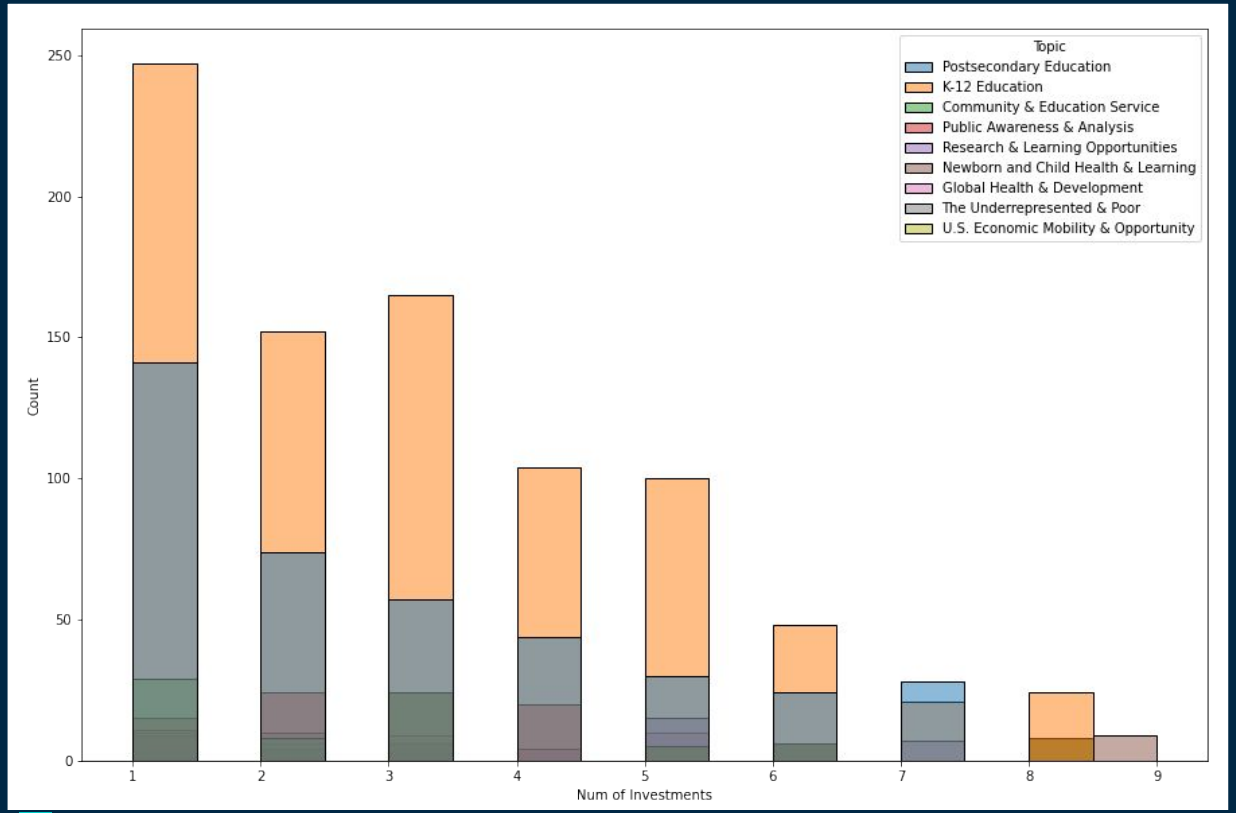
Postsecondary

- High amount per grants

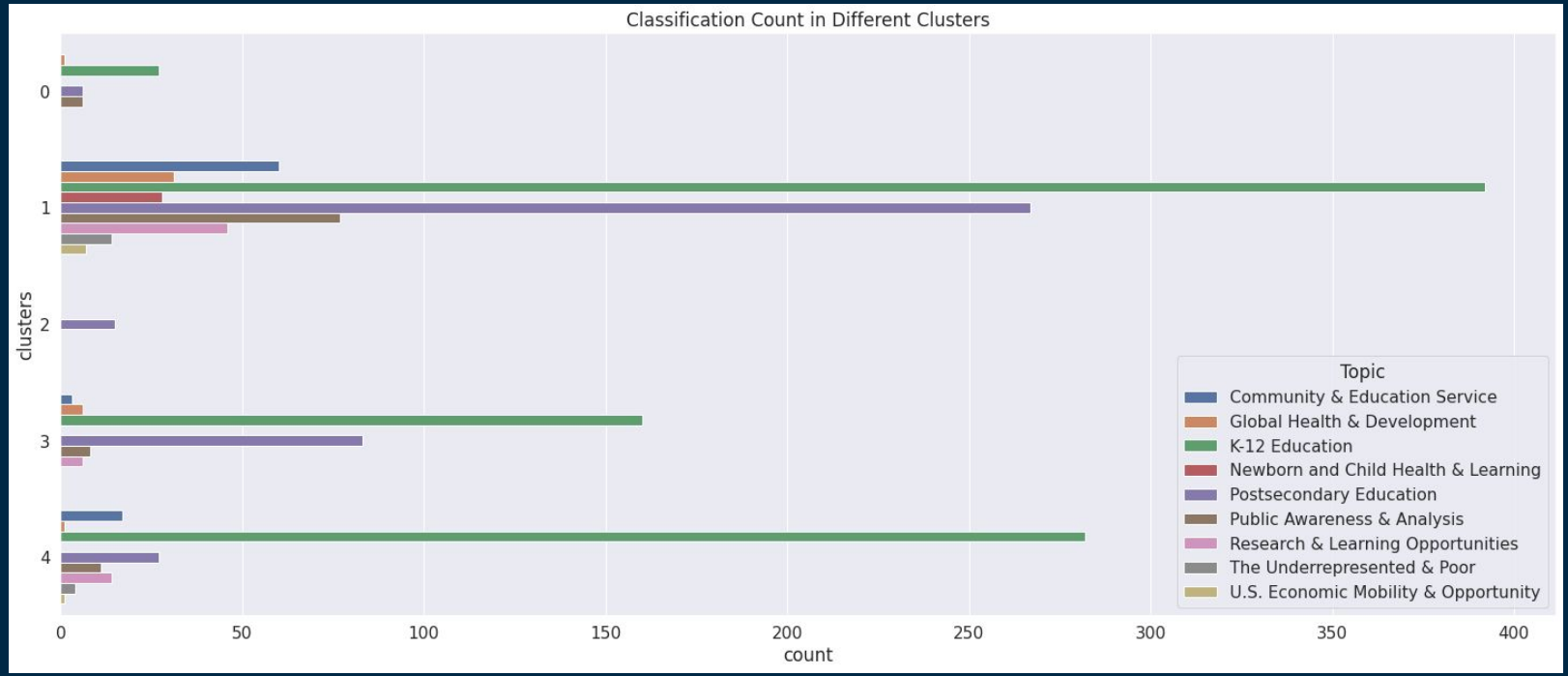
The Unexpected

- The Underrepresented & Poor have relatively high amount per grants

ANALYSIS (Per Company)

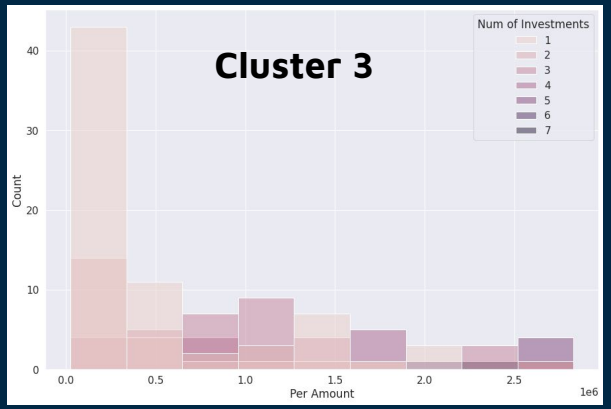


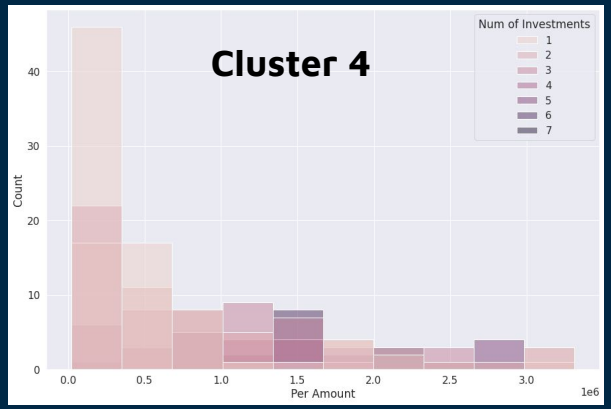
OBSERVATION (Cluster Label)

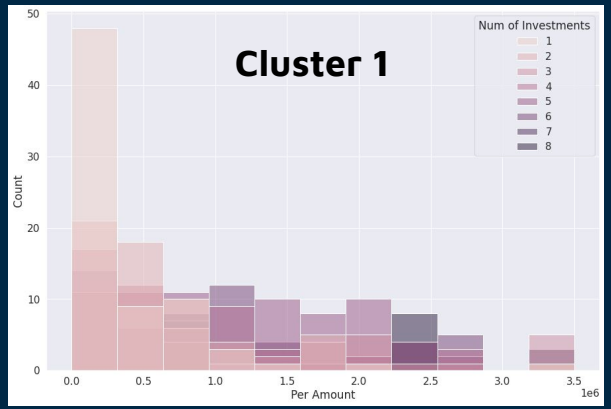


OBSERVATION (Cluster Label) (Conti.)

Common trend: More #grants, higher $grant each time (trust building)







Cluster insights

More companies with multiple grants

More favored companies

PRELIMINARY ANALYSIS

- Identify some common trends/ changes and analyze Gates' foundation **grant preference**
  - Gates Foundation values trust building
  - Original "Topic", though skewed still show some trends
    - Over time - 2010-2018
    - State x Topic preference
    - Certain topic tends to get high average grant
- Find **influential features**, inform hyperparameter tuning in the modeling part
  - **Geographic data** play a big role in grant decision (0 or 800M!)
  - **Cluster label** learned from ML has association with trust/preference
  - **Topic** should be included as a supplementary feature

MODELING & RESULTS

03

PREDICTION - OLS

- **Goal**: Predict the percentage of grants for each topic in 2019
- **Method**:
  - Independent variables: State and clusters
    - Use One Hot Encoding to change the categorical variables to numerical
  - Dependent variable:
    - Group by Topic
    - Use aggregate function to get the percentage of grants in each topic
- **Result**:
  - MSE: 0.00032
  - R squared: 38.43%
- **Further Improvement**: Add more independent variables, but be cautious with overfitting

Classification RANDOM FOREST

Goal: Given a company/ institution, to determine whether or not Gates Foundation to its donation would be durable, being donated more than once

Input: Everything excludes features related to num_investment, including `Year`, `Location`, `Average Amount,` etc

Output: 0 or 1

How to build up the model: Use Grid Search with cross-validation to try different parameters combinations, and choose the combination with best performance.

Performance:

Training_accuracy: 1
Tests_accuracy: 0.9425

Classification ADABOOST (similar to RF)

Goal: Given a company/ institution, to determine whether or not Gates Foundation to its donation would be durable, being donated more than once

Input: Everything excludes features related to num_investment, including `Year`, `Location`, `Average Amount,` etc

Output: 0 or 1

How to build up the model: Use Grid Search with cross-validation to try different parameters combinations, and choose the combination with best performance.

Performance:

Training_accuracy: 0.995
Tests_accuracy: 0.978

Why Ensemble Method

From the EDA process and domain knowledge, we believe this characteristics, `Duration`, could be classified by known features with decision boundary. We do not want to use Neural Net to achieve it as it would make features become more complex and may be overfitting and the final features under NN would not be easily interpreted or making sense.

However, we need to keep the mind that when using those methods, we may encounter the problem of overfitting. Hence, we should not be surprise and excited when seeing the training accuracy is close to 1.

The training and test set are splitted by all data set with tests_sizes = 0.25. The reason we split the data as whole from 2010-2018, instead of using the latest 2018 as test, is because different years may have different policy or other confounding factors. Hence, with ensemble method and such way splitting data would help us reduce the effects of those confounding factors.

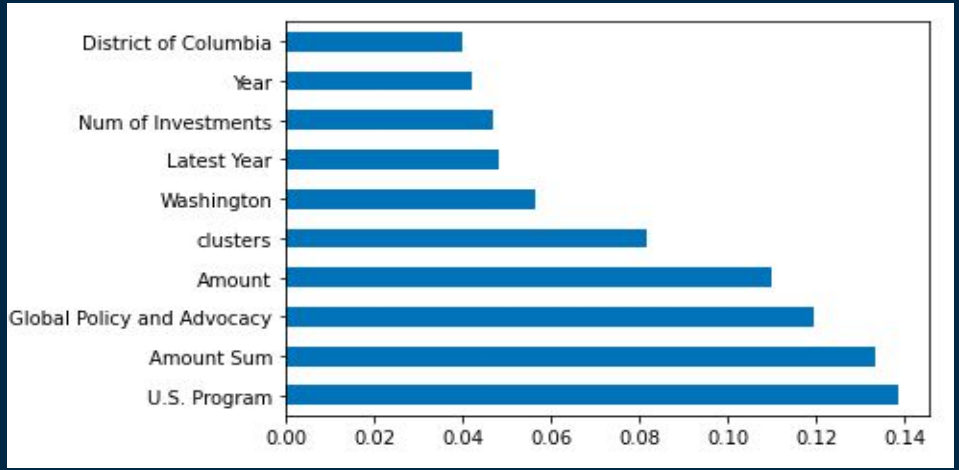
TOPIC CLASSIFICATION - Random Forest

Goal: Based on the information of a company, classify the topic of the grant it received

Observation: 20 unique topics, 10 combined topics (dropped)

Features:

- Division (one-hot encoded)
- State (one-hot encoded)
- Cluster Label (from KMeans)
- Number of Grants Received
- Amount Sum
- Amount
- Latest Year
- Year



Performance: Random Forest Classifier with max_depth = 10

Accuracy on the training set is 0.8929, on the test set is 0.7

TOPIC CLASSIFICATION - RNN

Goal: Based on the new description we have created for each grant, classify the topic

Model: Keras

- Maximum sequence length in the list of sentences: 899
- Number of unique tokens found: 1828

Layers & Hyperparameters:

- Embedding: number_of_words = 2000, input_length = max_seq_length
- LSTM: memory_cell = 100
- Dropout: 0.2
- Dense: activation_function = 'softmax' (multi-class classification)
- Loss= 'categorical_crossentropy'; Optimizer= 'adam'; Epochs = 10 ...

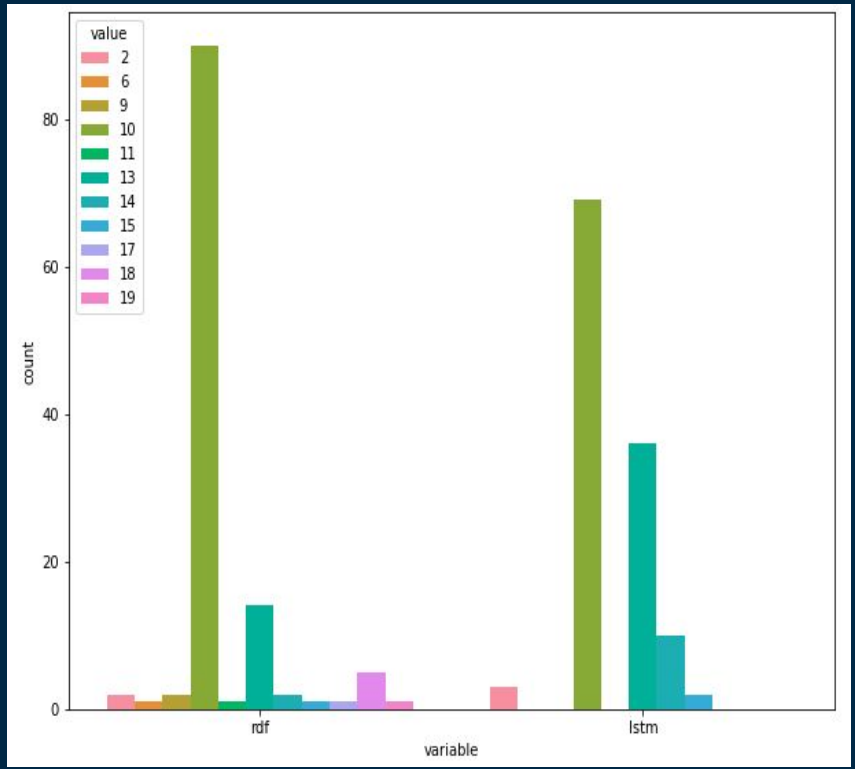
Performance:

```
test
  Loss: 1.438
  Accuracy: 0.742
```

- Slightly better than RDF
- Adding layers didn't help improve the accuracy
- Validation accuracy jumped back and forth between 0.67 and 0.71

Further Exploration

- Try more hyperparameter tuning
- Examine the grant that has different predicted topic (what makes the difference?)
- Examine the grant that has the same predicted topic (what contributes to the consistency?)
- Combine the two models using ensemble method
- Expand classification to the grants with more than one topic
- Topic prediction for a new company

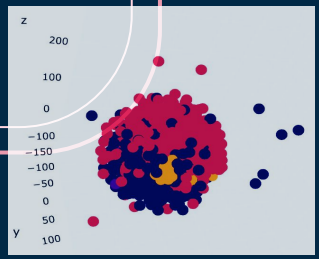


CONCLUSION

04

CONCLUSION

- We found that the more frequent a company gets donations from Gates the higher the amount of the donation it likely to be
  - Which shows Gates foundation is more likely to invest into companies that they trust
- We observed that majority of the donations go towards K12 education and postsecondary education (as shown on bottom right picture)
- Geographically, California gets the highest amount of donations followed by District of Columbia and New York
- Hence, **we predict that the company with a focus on K12 education that is based in California and have received donations from Gates in the past will be more likely to receive future donations**



THANK YOU