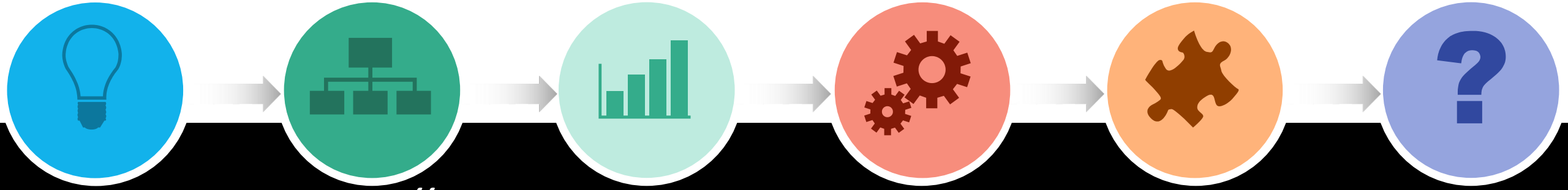
Idea and Inspiration | Data Ingestion and Wrangling | Computation and Analysis | Modeling and Feature Selection | Reporting and Visualization | Conclusion and Q&A



“Should I Stay or Should I Go?”

Applying Machine Learning to the FEVS to Predict Federal Employee Intent to Leave

Chuanfeng “Joy” Wu

Jennifer Ziegler

Georgetown University Data Science Certificate Cohort 27

September 24, 2022

Idea and Inspiration

Project Aspirations

- Our group initially had three members
  - Two of worked for the federal government
    - One for the Department of Labor
  - Third had expertise in organizational communication
- We all wanted to work with a tabular data set
- We all had a goal to learn Python

→ We chose to work with data from the Federal Employees Viewpoint Survey (FEVS)

"FEVS"

"an organizational climate survey that assesses how employees jointly experience the policies, practices, and procedures characteristic of their agency and its leadership" -OPM

- Conducted annually by the Office of Personnel Management (OPM)
- On average, each spring from 2016-2020:
  - About 80 federal agencies participated
  - About 1.25M employees were invited to participate
  - About 100 questions were asked
  - About 500K responded, for a 44% response rate
- Governmentwide and agency reports issued each fall
  - Except for 2020 when survey distribution and tabulation were delayed 6 months due to the pandemic

We focused on one question that asked about an employee's intent to leave within the next year:

91. Are you considering leaving your organization within the next year, and if so, why?

- ❍ No
- ❍ Yes, to retire
- ❍ Yes, to take another job within the Federal Government
- ❍ Yes, to take another job outside the Federal Government
- ❍ Yes, other

Initial Hypothesis: By analyzing the FEVS data using Machine Learning, we will identify the factors that can predict someone answering "Yes, to take another job outside the Federal Government" to this question (i.e., "Leave").

Some of us also remembered a similarly titled Clash song from the 80s…

"Should I Stay or Should I Go, Now?" – The Clash

➔ **And so we became "Team Stay or Go"**

The OPM collected certain questions into "Indexes" and tracked trends each year

Employee Engagement Index (EEI):

- 15 questions
- Three categories

Employee Engagement Index–Governmentwide

Index	2016	2017	2018	2019	2020
Overall Employee Engagement	65	67	68	68	72
Leaders Lead	53	55	56	57	62
Supervisors	72	74	75	76	80
Intrinsic Work Experience	70	71	72	72	76

Global Satisfaction Index (GSI):

- 4 questions

Global Satisfaction Index

The Global Satisfaction Index is an average of the scores of the four items below:

Job Satisfaction	Pay Satisfaction	Organizational Satisfaction	Recommend Organization
Considering everything, how satisfied are you with your job? (Q. 69)	Considering everything, how satisfied are you with your pay? (Q. 70)	Considering everything, how satisfied are you with your organization? (Q. 71)	I recommend my organization as a good place to work. (Q. 40)

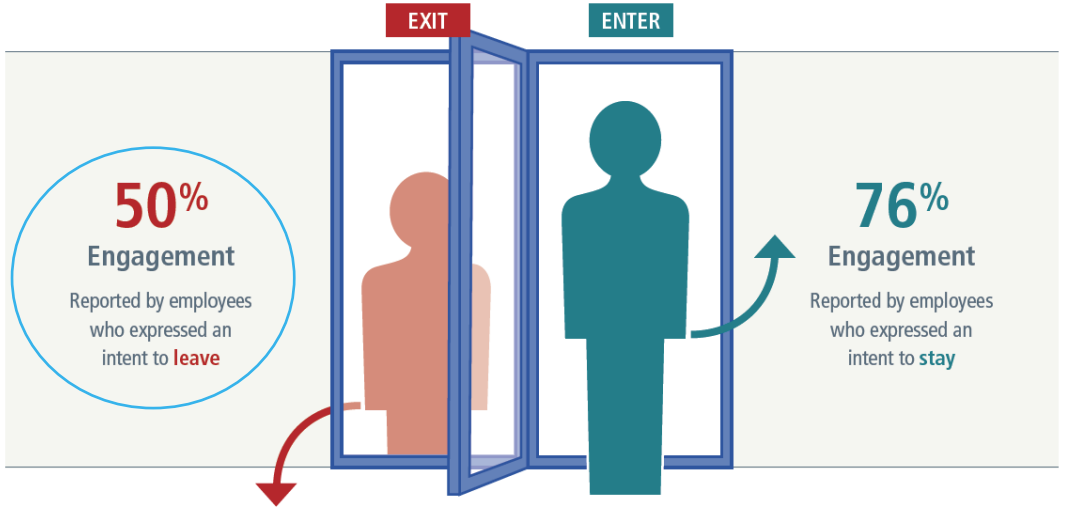
Also:

- **NIQ** (New Inclusion Quotient)
- **AES** (Annual Employee Survey mandated by Congress)

OPM twice used the Indexes to analyze Intent to Leave:

Once in 2016 (EEI):

And again in 2018 (EEI):



(In 2021, after the time period for our analysis, the OPM also associated Intent to Leave with low GSI scores.)

"What proportion of people answered 'Leave'?"

3.6%

According to our analysis of the survey data

"Is that a lot or a little?"

We compared Intent to Leave to the BLS "Quits" Rate:

At the Height of the Great Resignation (Dec 2021):

All Sectors	**2.9%**
Federal Government Employees	**0.7%**

During the time period of the survey:

	All	FGE
May 2016	2.0%	**0.4%**
May 2017	2.2%	**0.4%**
May 2018	2.5%	**0.5%**
May 2019	2.3%	**0.5%**
Sept 2020	2.3%	**0.6%**

Source: Bureau of Labor Statistics Job Openings and Labor Turnover Survey "JOLTS" at https://www.bls.gov/news.release/archives/jolts_02012022.pdf

Project Goals

Primary:

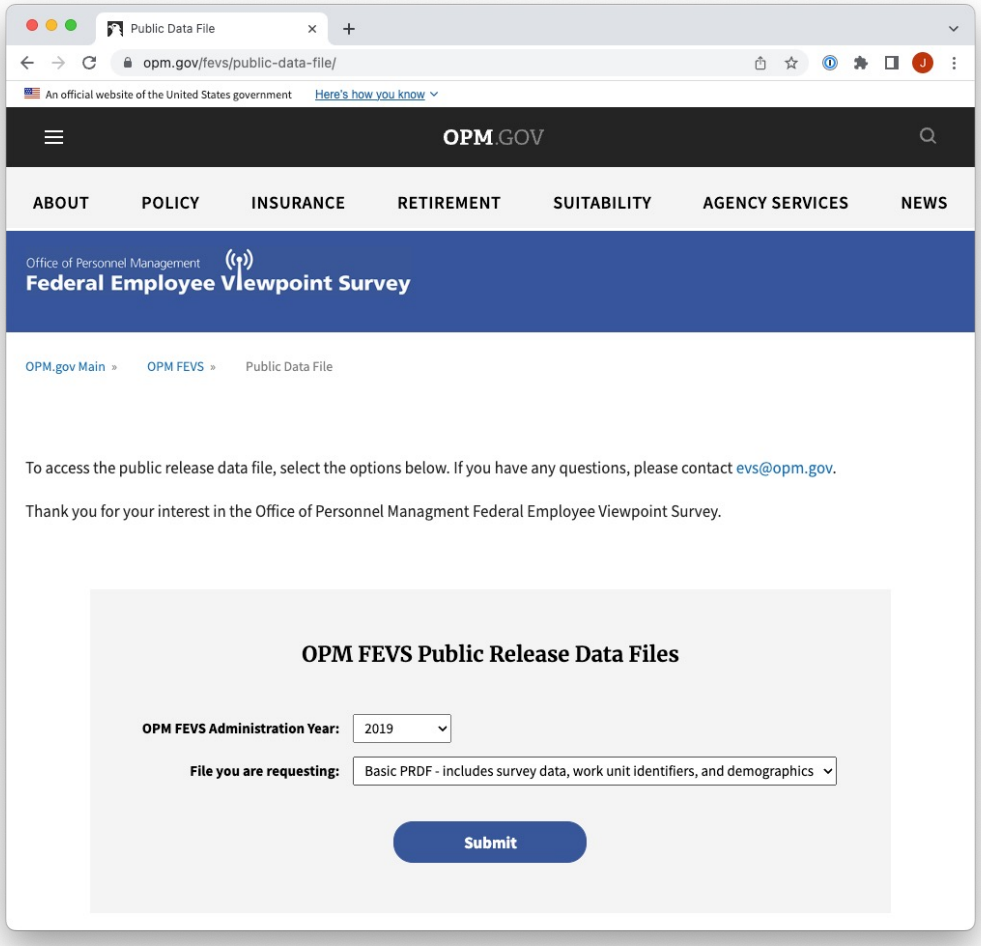
- Use Machine Learning to analyze the FEVS and identify the factors that predict an employee choosing “Leave” on the Intent to Leave Question

Secondary:

- Validate (clarify/debunk) any assumed relationship between the Indexes and answers on Intent to Leave question
- Develop a Web Survey or App that could allow federal managers to predict workforce availability (sooner), or even intervene

Data Ingestion and Wrangling

We downloaded Public Data files (.csv) for 2016, 2017, 2018, 2019, and 2020



- 2016
  - 2016 PRDF Codebook.xlsx
  - 2016 Survey Instrument.pdf
  - 2016_fevs_gwide_final_report.pdf
  - 2016_fevs_technical_report.pdf
  - fevs_2016_prdf.csv
- 2017
  - 2017 PRDF Codebook.xlsx
  - 2017 Survey Instrument.pdf
  - 2017_fevs_gwide_final_report.pdf
  - 2017-technical-report.pdf
  - FEVS_2017_PRDF.csv
- 2018
  - 2018 PRDF Codebook.xlsx
  - 2018 Survey Instrument.pdf
  - 2018_fevs_gwide_final_report.pdf
  - 2018_opm_technical_report.pdf
  - FEVS_2018_PRDF.csv
- 2019
  - 2019 PRDF Codebook.xlsx
  - 2019 Survey Instrument.pdf
  - 2019_fevs_gwide_final_report.pdf
  - 2019-fevs-technical-report.pdf
  - FEVS_2019_PRDF_Revised_2020-04-27.csv
- 2020
  - 2020 PRDF Codebook.xlsx
  - 2020 Survey Instrument.pdf
  - 2020-fevs-gwide-final-report.pdf
  - 2020-opm-fevs-technicalreport-508.pdf
  - 2020-opmfevs-govwide-report-by-agency.xlsx
  - FEVS_2020_PRDF.csv

We also downloaded government-wide reports, technical reports, and codebooks to add context.

Data Features

Public Data File	2016	2017	2018	2019	2020	Totals
File Size (MB)	75.5	83.9	111.8	128.8	177.4	577.4
Core Questions	71	71	71	71	38	
Demographics Questions	5	6	6	6	11	
Other Questions / Variables	4	3	3	8	87	
Total Variables (Columns)	80	80	80	85	136	188 unique
Total Respondents (Rows)	407,789	486,105	598,003	615,395	624,800	2,732,092
Expected Data Points (CxR)	32,623,120	38,888,400	47,840,240	52,308,575	84,972,800	256,633,135
Total Responses (non-Null Data Points)	30,282,319	36,293,100	44,890,253	51,345,364	76,952,601	239,763,637
Data Availability Rate	92.82%	93.33%	93.83%	98.16%	90.56%	93.43%
Cells with Null Data	2,340,801	2,595,300	2,949,987	963,211	8,020,199	16,869,498

Early on, we decided to combine all questions from all years

```
In [6]: 1 FEVS5year.columns

Out[6]: Index(['response_id', 'year', 'agency_id', 'Q234', 'Q226', 'Q228', 'Q225',
               'Q236', 'Q237', 'Q102',
               ...
               'Q277', 'Q278', 'Q279', 'Q280', 'Q281', 'Q282', 'Q283', 'Q284', 'Q285',
               'StayorGo'],
              dtype='object', length=189)
```

```
In [7]: 1 FEVS5year.head()
        2
```

Out[7]:

	response_id	year	agency_id	Q234	Q226	Q228	Q225	Q236	Q237	Q102	...	Q277	Q278	Q279	Q280	Q281	Q282	Q283	Q284	Q285	StayorGo
0	2016000002	2016	TR	TR93	A	B	B	A	A	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay
1	2016000003	2016	AF	AF1C	A	A	B	B	A	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay
2	2016000004	2016	TR	TRAD	A	A	B	B	A	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay
3	2016000005	2016	TR	TR93	A	A	B	B	D	3.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Retire
4	2016000006	2016	HE	HE09	B	B	B	B	A	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay

5 rows × 189 columns

```
In [8]: 1 FEVS5year.tail()
        2
```

Out[8]:

	response_id	year	agency_id	Q234	Q226	Q228	Q225	Q236	Q237	Q102	...	Q277	Q278	Q279	Q280	Q281	Q282	Q283	Q284	Q285	StayorGo
624795	2020624797	2020	HS	NaN	A	B	B	A	NaN	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Transfer
624796	2020624798	2020	HS	NaN	A	A	B	A	NaN	5.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay
624797	2020624799	2020	HS	NaN	A	B	NaN	NaN	NaN	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Transfer
624798	2020624800	2020	HS	NaN	A	B	A	B	NaN	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Retire
624799	2020624801	2020	HS	NaN	B	B	A	B	NaN	4.0	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Stay

5 rows × 189 columns

We collected all Intent to Leave responses in one column: 'Stay or Go': Stay, Transfer, Retire, or Leave.

This brought our column total to 189.

We created a Snowflake SQL database for WORM storage

snowflake

Databases > STAYORGO

Tables | Views | Schemas | Stages | File Formats | Sequences | Pipes

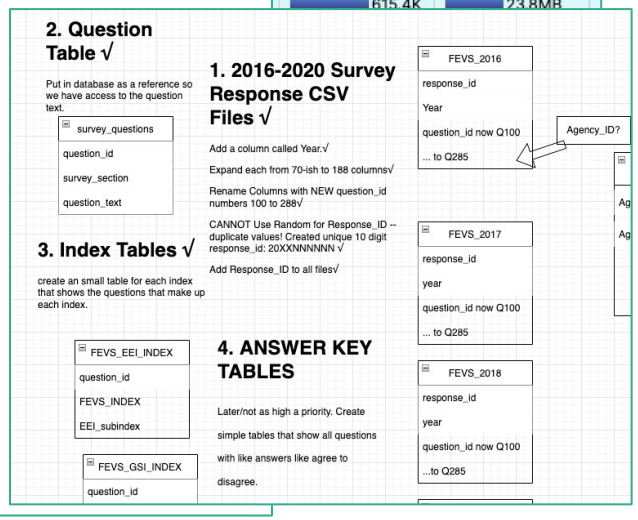
Create... | Create Like... | Clone... | Load Data... | Drop... | Transfer Ownership

Table Name	Schema	Creation Time	Owner	Rows	Size
AGENCY	PUBLIC	12:28:51 PM	ACCOUNTADMIN	54	2.5KB
FEVS_2016	PUBLIC	6:03:12 PM	ACCOUNTADMIN	407.8K	13.9MB
FEVS_2017	PUBLIC	6:07:04 PM	ACCOUNTADMIN	486.1K	17.3MB
FEVS_2018				598.0K	19.7MB
FEVS_2019				615.4K	23.8MB
FEVS_2020					
FEVS_EEI_INDEX					
FEVS_GSI_INDEX					
FEVS_NIQ_INDEX					
SURVEY_QUESTIONS					

```
alter table fevs_2016_B add column StayOrGo varchar(20);

update fevs_2016_b
Set StayOrGo = 'Stay'
where Q237 = 'A';
Set StayOrGo = 'Transfer'
where Q237 = 'B';
Set StayOrGo = 'Leave'
where Q237 = 'C';
Set StayOrGo = 'Retire'
where Q237 = 'D';

update fevs_2020_b
Set StayOrGo = 'Stay'
where Q231 = 'A';
Set StayOrGo = 'Transfer'
where Q231 = 'C';
Set StayOrGo = 'Leave'
where Q231 = 'D';
Set StayOrGo = 'Retire'
where Q231 = 'B';
```



But after our columns were standardized, we mainly worked with CSV files.

1. Read 2016 to 2020 FEVS survey data as Dataframes.

In [2]:

```
# need to give the path of these file on you own
df2016 = pd.read_csv('Standardized_Columns_fevs_2016_prdf.csv', low_memory=False)
df2017 = pd.read_csv('Standardized_Columns_fevs_2017_prdf.csv', low_memory=False)
df2018 = pd.read_csv('Standardized_Columns_fevs_2018_prdf.csv', low_memory=False)
df2019 = pd.read_csv('Standardized_Columns_fevs_2019_prdf.csv', low_memory=False)
df2020 = pd.read_csv('Standardized_Columns_fevs_2020_prdf.csv', low_memory=False)

```

2. Combine 5 years of FEVS survey data into one dataframe.

call the new file FEVS5year

In [3]:

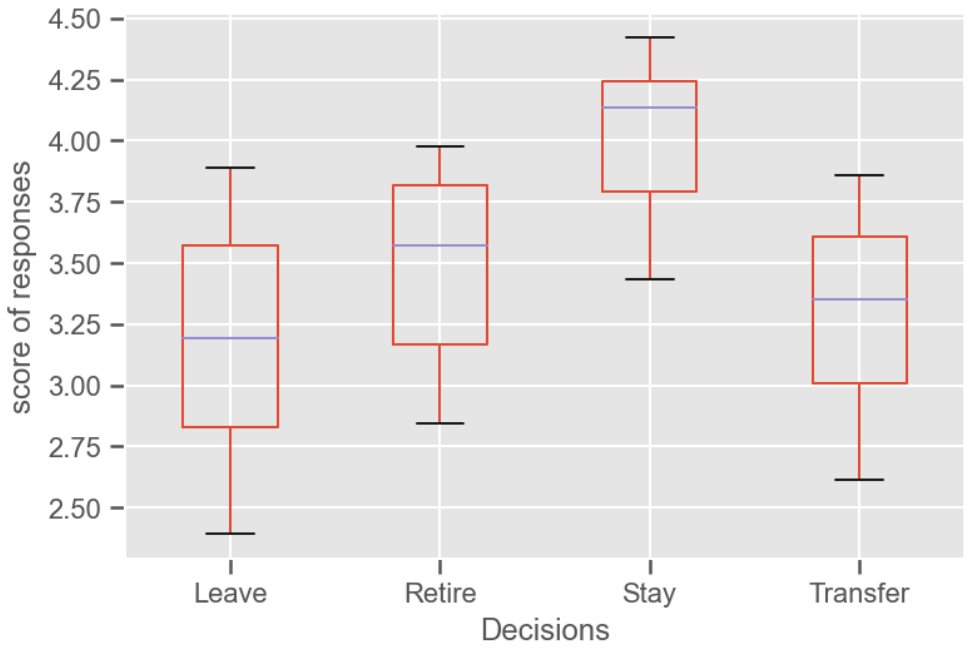
```
# combine 5 years of survey data into one datafram
frames = [df2016, df2017, df2018, df2019, df2020]
FEVS5year= pd.concat(frames)

```

Computation and Analysis

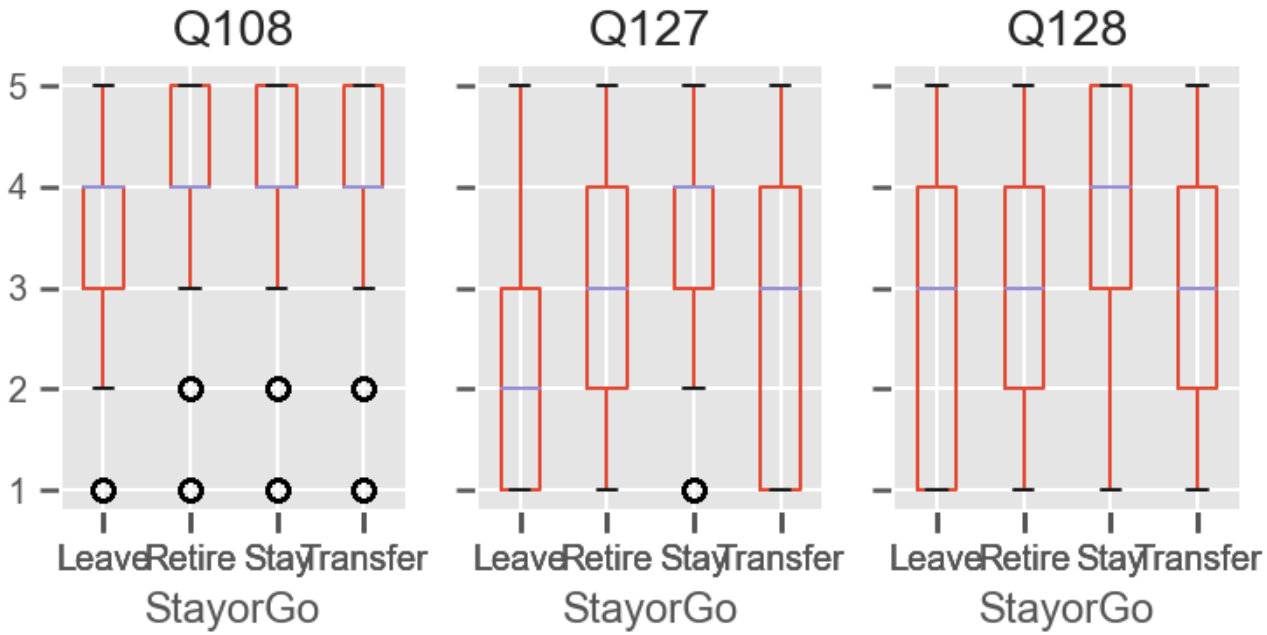
In early exploratory data analysis (EDA), we found “Leave” to be correlated with low EEI scores

On average, employees expressing “Leave” scored lower on EEI questions than those intending “Stay.”



EEI= Employee Engagement Index

“Leave” responses to Q108, Q127, and Q128 were especially striking.

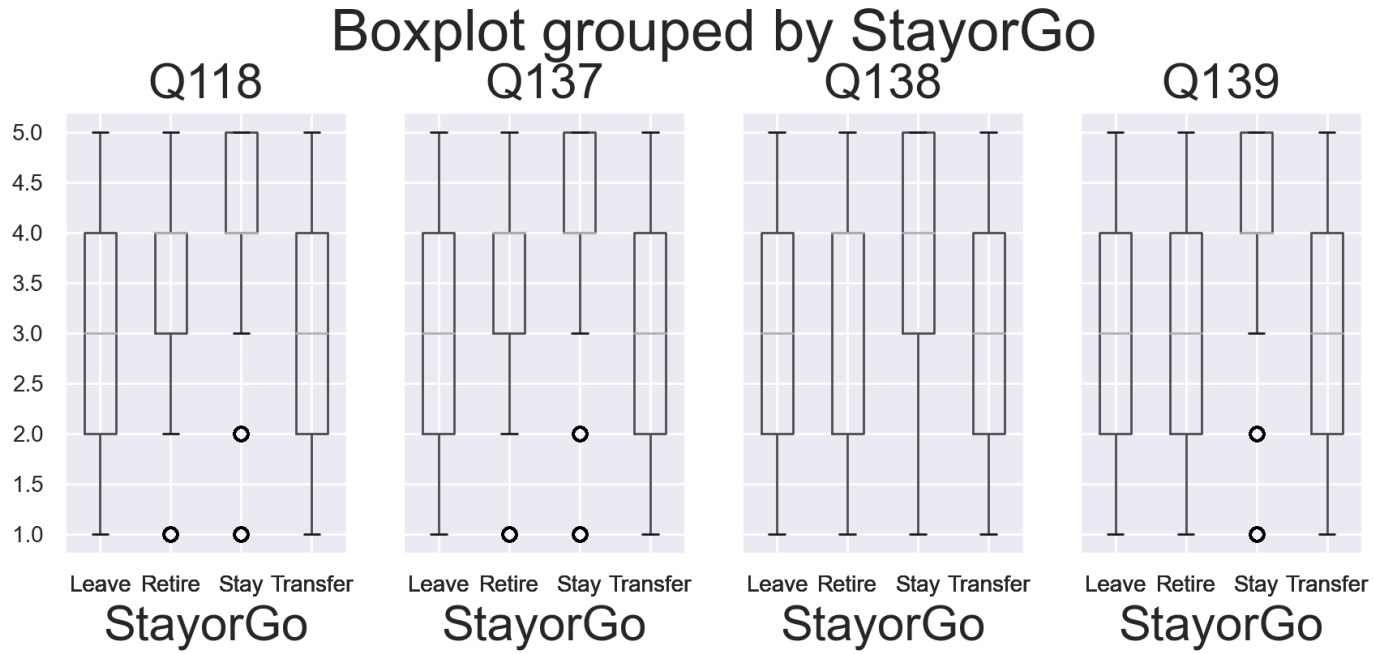


Q108 = I know how my work relates to the agency's goals.
Q127 = In my organization_ senior leaders generate high levels of motivation and commitment in the workforce.
Q128 = My organization's senior leaders maintain high standards of honesty and integrity.

In early EDA, we also found “Leave” to be correlated with low Global Satisfaction Index (GSI) scores

Those answering “Leave” rated answers to the GSI questions lowest of all.

Leave and Transfer answered similarly on GSI questions, but Stay responses looked dramatically different.



Q118 = I recommend my organization as a good place to work.
Q137 = Considering everything, how satisfied are you with your job?
Q138 = Considering everything, how satisfied are you with your pay?
Q139 = Considering everything, how satisfied are you with your organization?

Before we could start with Machine Learning we still needed to transform our data

Back to Data Wrangling…

1. Some of the data were already numeric (no transformation required)

My work gives me a feeling of personal accomplishment.		
	5	Strongly Agree
	4	Agree
	3	Neither Agree nor Disagree
	2	Disagree
	1	Strongly Disagree

```
In [13]:  1  FEVS5yearT2['Q104']

Out[13]: 0          4.0
         1          3.0
         2          5.0
         3          4.0
         4          5.0
                   ...
         2732087    4.0
         2732088    4.0
         2732089    3.0
         2732090    4.0
         2732091    5.0
         Name: Q104, Length: 2732092, dtype: float64
```

2. We “coerced” almost-numeric data to numeric using Pandas

Q119=

My talents are used well in the workplace.		
	5	Strongly Agree
	4	Agree
	3	Neither Agree nor Disagree
	2	Disagree
	1	Strongly Disagree
	X	Do Not Know

```
In [22]: 1 print(FEVS5year['Q119'].unique())

['5' '2' '4' '3' 'X' nan '1']
```

```
In [35]: 1 FEVS5year3[['Q103','Q104','Q106','Q107','Q113','Q118',
         2              'Q119','Q126','Q133','Q136','Q137','Q138','Q139',]]= FEVS5year3[
         3              ['Q103','Q104','Q106','Q107','Q113','Q118','Q119',
         4               'Q126','Q133','Q136','Q137','Q138','Q139',]].apply(pd.to_numeric,errors='coerce')
         5 FEVS5year3.Q119.unique()
         6
         7

Out[35]: array([ 5.,  2.,  4.,  3., nan,  1.])
```

3. We used OneHotEncoding to convert categorical data to binary

Q226=

What is your supervisory status?		
	A	Non-Supervisor/Team Leader
	B	Supervisor/Manager/Executive

In [26]:
```
1 FEVS5year3 = pd.get_dummies(FEVS5year2, columns = ['Q226','Q227','Q228','
2 print(FEVS5year3)
```

```
        agency_id  Q103  Q104  Q106  Q107  Q113  Q118  Q119  Q126  Q133  ...  \
0              TR   2.0   4.0     4     4     2   5.0     5   4.0     3  ...
1              AF   3.0   3.0     4     2     2   2.0     2   3.0     X  ...
2              TR   5.0   5.0     5   NaN     4   5.0     4   4.0   NaN  ...
3              TR   4.0   4.0     4     4     3   4.0     3   5.0     3  ...
4              HE   5.0   5.0     4     5     4   5.0     4   5.0     5  ...
...           ...   ...   ...   ...   ...   ...   ...   ...   ...   ...  ...
624795         HS   5.0   4.0     2     4     2   3.0     2   4.0     2  ...
624796         HS   5.0   4.0     5     5     4   5.0     5   5.0     4  ...
624797         HS   2.0   3.0     3     4     4   1.0     1   5.0     3  ...
624798         HS   3.0   4.0     4     3     3   4.0     3   5.0     4  ...
624799         HS   5.0   5.0     5     5     2   5.0     4   4.0     5  ...

        Q226_A  Q226_B  Q227_A  Q227_B  Q227_C  Q228_A  Q228_B  Q235_A  Q235_B
0            1       0       0       0       0       0       1       0       0
1            1       0       0       0       0       1       0       0       0
2            1       0       0       0       0       1       0       0       0
3            1       0       0       0       0       1       0       0       0
4            0       1       0       0       0       0       1       0       0
```

4. We imputed NaN using the column mean

Q107=

My talents are used well in the workplace.		
	5	Strongly Agree
	4	Agree
	3	Neither Agree nor Disagree
	2	Disagree
	1	Strongly Disagree
	X	Do Not Know

```
In [49]:  1 FEVS5yearT2.Q107.unique()

Out[49]: array([ 4.,  2., nan,  5.,  3.,  1.])
```

```
In [54]:  1 impute = {}
          2
          3 for col in cols:
          4     impute[col] = FEVS5yearT2[col].mean()
```

```
In [55]:  1 FEVS5yearT2.fillna(value=impute, inplace=True)
```

```
In [56]:  1 FEVS5yearT2.Q107.unique()
          2

Out[56]: array([4.        , 2.        , 3.52379154, 5.        , 3.        ,
                1.        ])
```

5. We observed a class imbalance for “Leave” . . .

3.6%

In [8]:
```
# determine the number of NaN cases in Stay or Go
print(FEVS5year['StayorGo'].unique())
FEVS5year['StayorGo'].value_counts(dropna=False)
```

```
['Stay' 'Retire' 'Transfer' nan 'Leave']
```

Out[8]:
```
Stay        1736122
Transfer     442466
Retire       278698
NaN          176148
Leave         98658
Name: StayorGo, dtype: int64
```

```
Stay = 0.6354551750087479
Transfer = 0.1619513544931869
Retire = 0.1020090099454923 2
Leave = 0.03611078982698972 4
NaN = 0.06447367072558317
--
Stay = 0.6354551750087479
any_not_stay = 0.3000711542656689 4
NaN = 0.06447367072558317
```

```
In [10]:  1 df.StayorGo.unique()
```

```
Out[10]: array(['Stay', 'Retire', 'Transfer', 'Leave', 'Go'], dtype=object)
```

```
In [11]:  1 df['StayorGo'].value_counts(dropna=False)
```

```
Out[11]: Stay        1855770
         Transfer     442466
         Retire       278698
         Leave         98658
         Go            56500
         Name: StayorGo, dtype: int64
```

```
In [12]:  1 df=df.replace(to_replace='Retire', value='Go')
          2 df=df.replace(to_replace='Transfer', value='Go')
          3 df=df.replace(to_replace='Leave', value='Go')
```

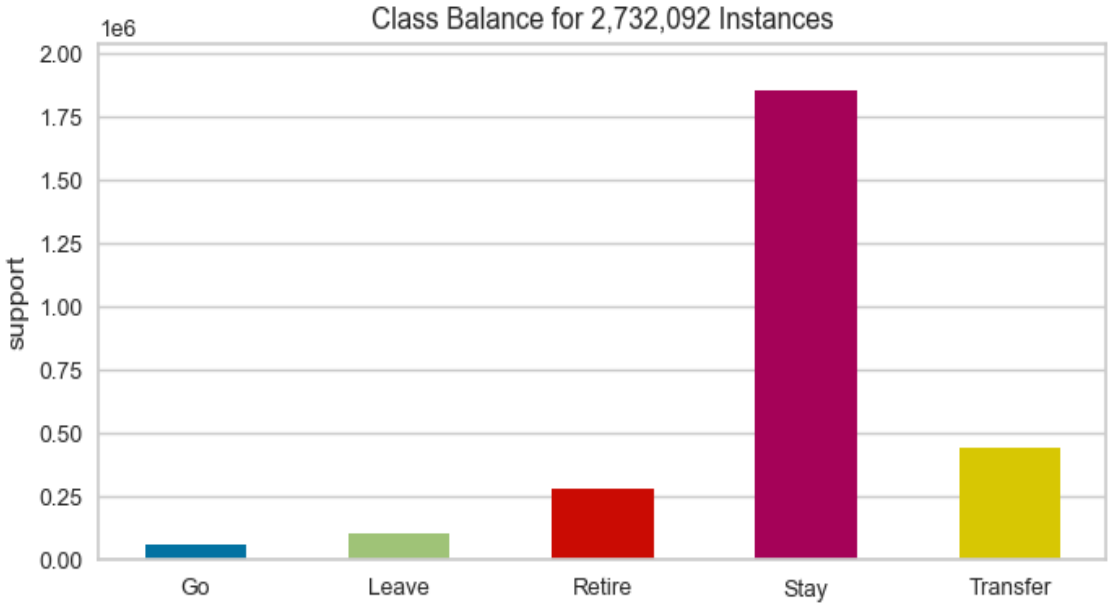
```
In [13]:  1 df['StayorGo'].value_counts(dropna=False)
```

```
Out[13]: Stay    1855770
         Go       876322
         Name: StayorGo, dtype: int64
```

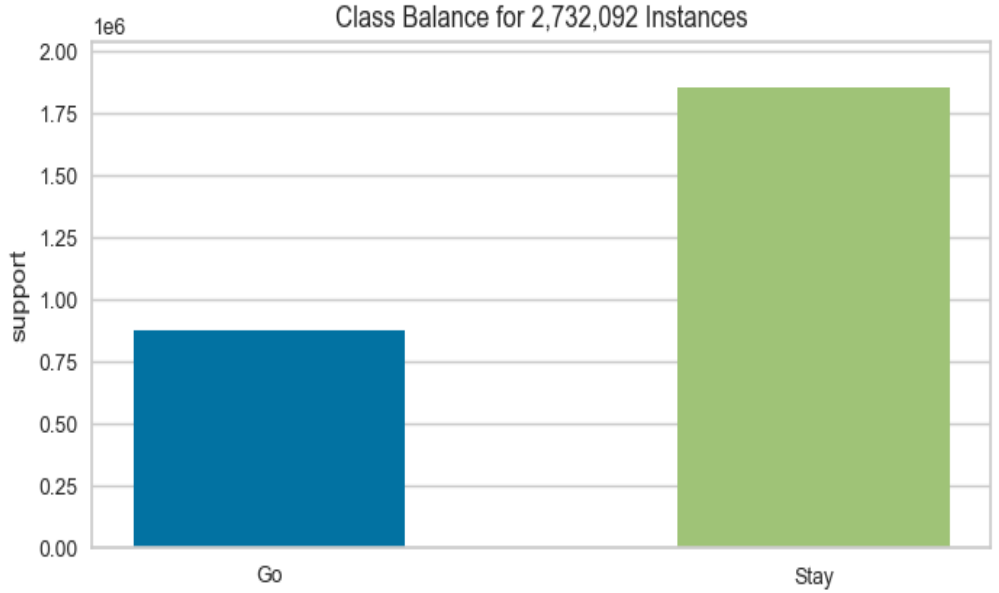
. . . So we transformed StayorGo to binary (‘Stay’, ‘Go’) and filled NaNs proportionally

This resulted in better balanced classes going into Machine Learning

Multi Class



Binary



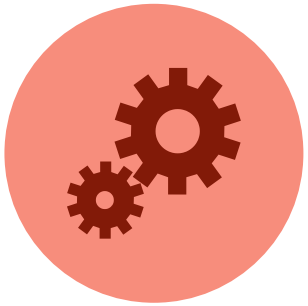
6. Finally, we dropped an additional 29 non-relevant columns, and entered ML with a 171 feature numeric dataframe

```
In [62]: 1 FEVS5year4.describe()
```

Out[62]:

	year	Q102	Q238	Q103	Q104	Q239	Q105	Q240	Q241	Q242	...	
count	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	2.732092e+06	...	2.732
mean	2.018206e+03	3.710247e+00	3.756616e+00	3.626316e+00	3.892961e+00	4.165207e+00	4.039856e+00	4.585513e+00	4.390788e+00	3.109714e+00	...	2.036
std	1.368376e+00	1.115781e+00	8.891551e-01	1.187627e+00	1.071978e+00	7.877853e-01	9.511967e-01	5.682929e-01	6.314818e-01	1.117031e+00	...	4.022
min	2.016000e+03	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	...	0.000
25%	2.017000e+03	3.000000e+00	3.756616e+00	3.000000e+00	3.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00	2.000000e+00	...	0.000
50%	2.018000e+03	4.000000e+00	4.000000e+00	4.000000e+00	4.000000e+00	4.165207e+00	4.000000e+00	4.585513e+00	4.390788e+00	3.109714e+00	...	0.000
75%	2.019000e+03	4.000000e+00	4.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	4.000000e+00	...	0.000
max	2.020000e+03	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	5.000000e+00	...	1.000

8 rows × 172 columns



Modeling and Feature Selection

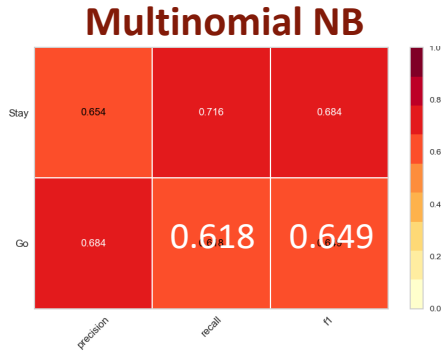
We evaluated five model families with Binary Classification Algorithms

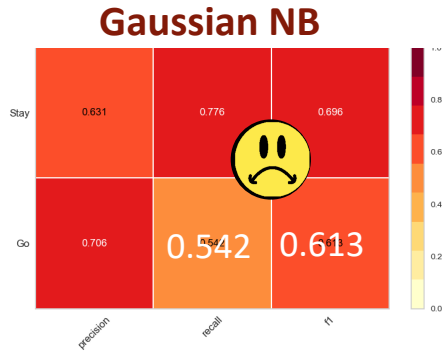
- **Naive Bayes**
  - Multinomial NB ()
  - Gaussian NB ()
- **Support Vector Classifier**
  - SVC (gamma='auto’)
  - Nu SVC (gamma='auto'),
  - Linear SVC()
- **Logistic Regression**
  - Logistic Regression (solver='lbfgs’)
  - Logistic Regression CV (cv=3)
- **K-Nearest Neighbors**
  - K Neighbors Classifier()
- **Ensemble Classifiers**
  - SGD Classifier (max_iter = 100, tol=1e-3)
  - Bagging Classifier()
  - Extra Trees Classifier (n_estimators = 100),
  - Random Forest Classifier (n_estimators = 100)

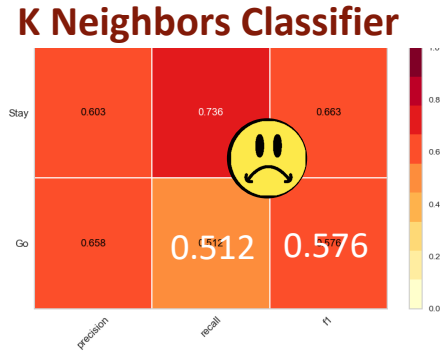
We initially sampled 20,000 cases from the data

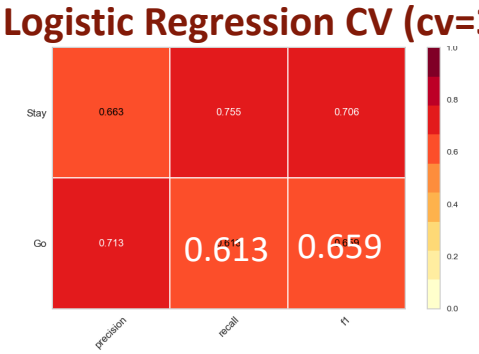
Modeling and Feature Selection

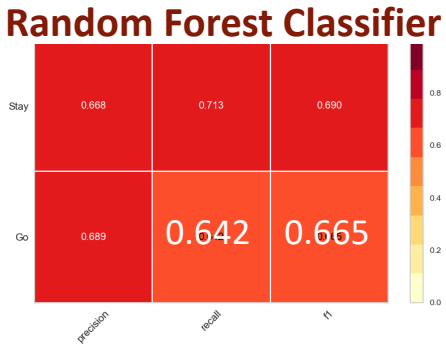
We examined Classification Reports to compare “Go” Recall Scores and F1 Scores

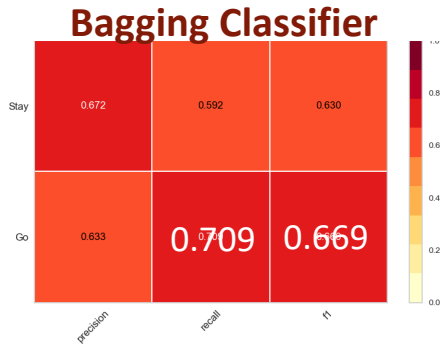


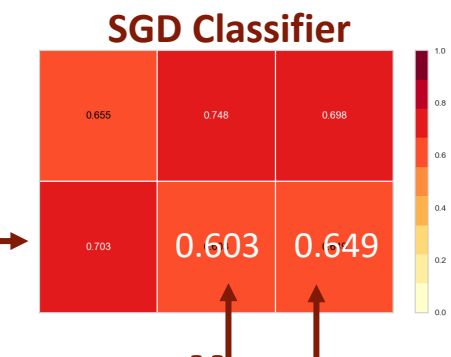


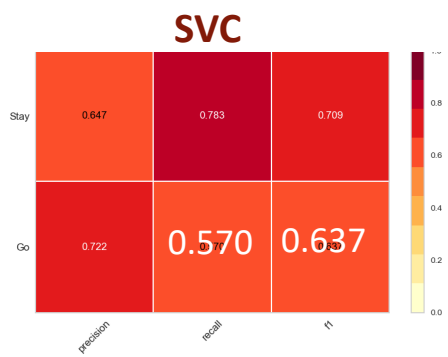


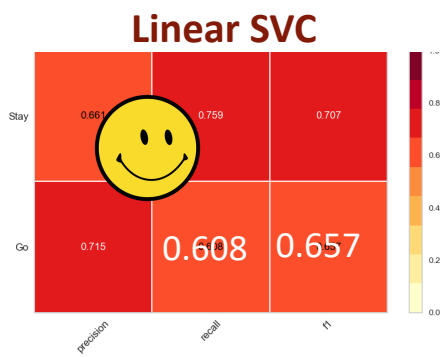


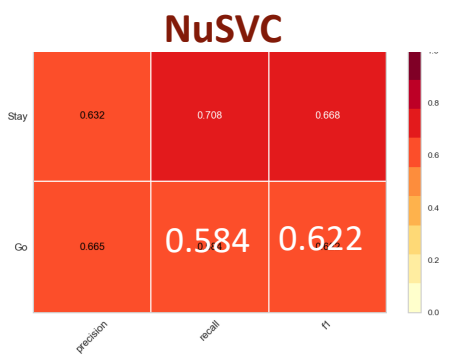










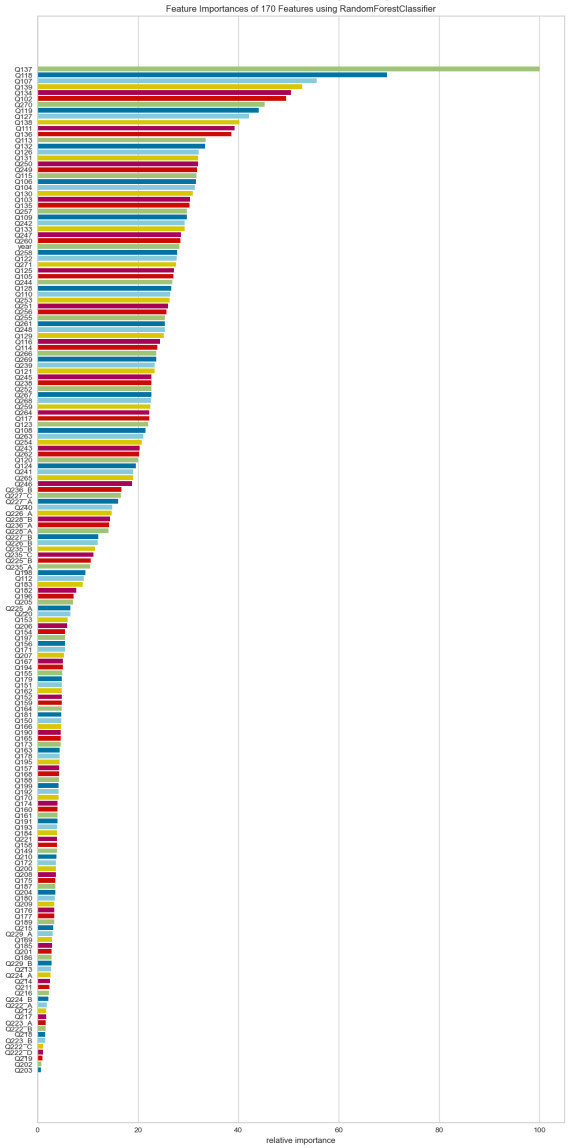


Go →

Recall **F1**

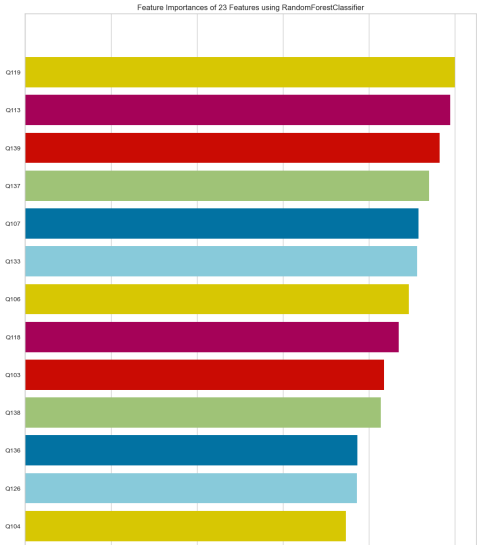
We gave stronger models more weight in feature reduction

Before

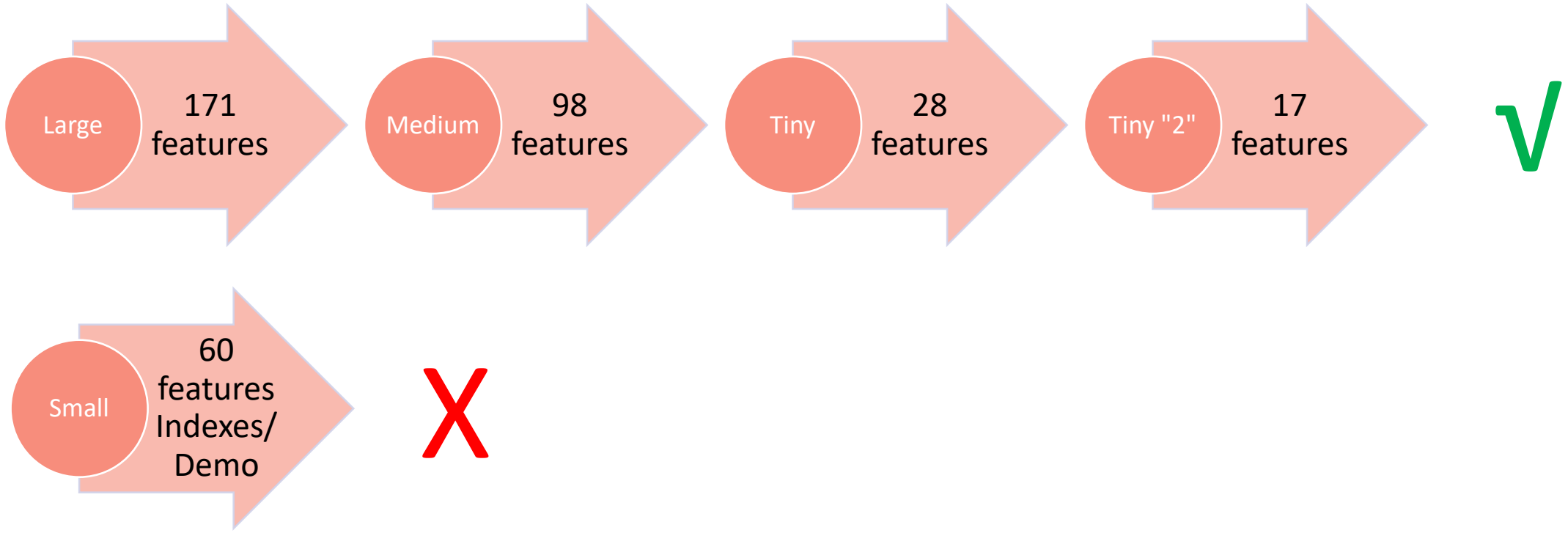


Label	question_id	Linear SVC	SGD Classifier	Logistic Regression	Logistic RegressionCV	Extra Trees	Random Forest
SatisfJob	Q137	1	1	1	1	1	4
TalentsUsed	Q107	4	6	4	4	4	5
RecommendOrg	Q118	2	4	2	2	7	8
SatisfOrg	Q139	6	5	7	7	12	3
SatisfPay	Q138	7	7	6	6	10	10
SurveyBetter	Q119	11	9	16	16	2	1
SrLSupportWL	Q133	16	21	13	13	6	6
PerfDiffRecog	Q113	18	13	19	19	3	2
GoodJobBySup	Q126	10	20	10	10	11	12
EducBach	Q235_B	8	12	8	8	15	15
SatisfRecog	Q136	14	15	14	14	8	11
WorkloadReas	Q106	17	11	18	18	5	7
PersAccomp	Q104	15	16	11	11	13	13
NonSupervisor	Q226_A	3	2	3	3	22	22
Female	Q228_B	9	14	9	9	18	17
Supervisor	Q226_B	5	3	5	5	23	23
EducMoreBach	Q235_C	12	8	15	15	14	14
EducLessBach	Q235_A	13	10	12	12	16	19
EncourBetter	Q103	21	18	21	21	9	9
Male	Q228_A	20	19	20	20	17	16
11to20Yr	Q227_B	19	17	17	17	19	20
10orLessYr	Q227_A	22	22	22	22	20	18
21orMoreYr	Q227_C	23	23	23	23	21	21
	StayOrGo						
	agency_id						

After



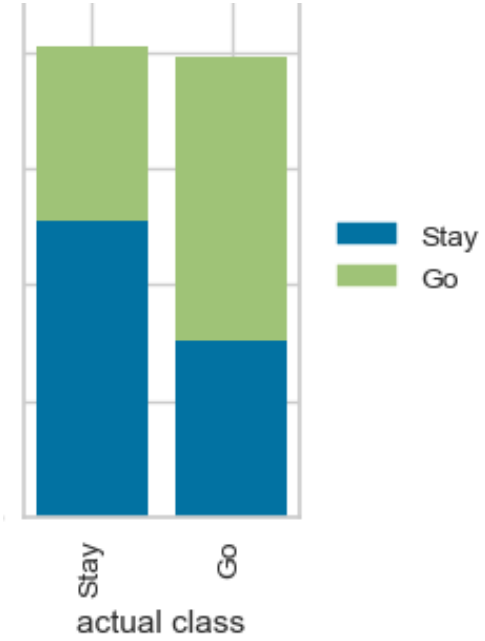
We reduced features from 171 to 17 through four iterations



(Ask us about this detour)

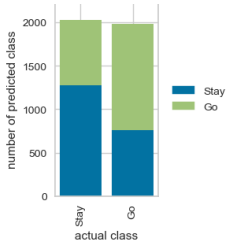
We compared Class Prediction Errors and Cross Validation Scores

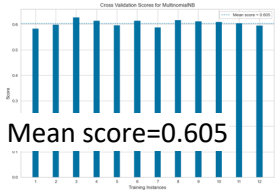
ClassPredictionError()



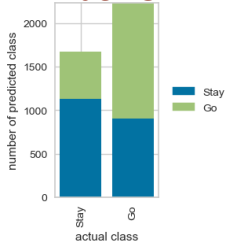
StratifiedKFold(K=12)
scoring='f1_weighted'

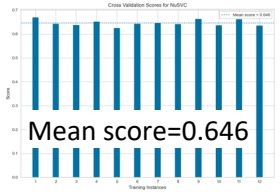
Multinomial NB



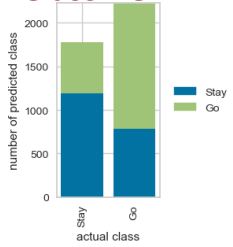


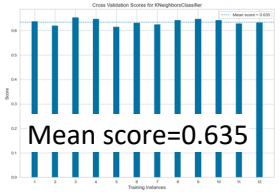
NuSVC



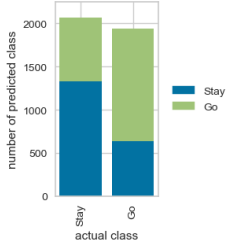


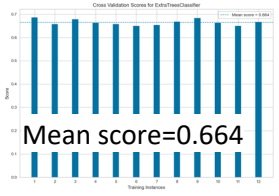
KNeighbors Classifier



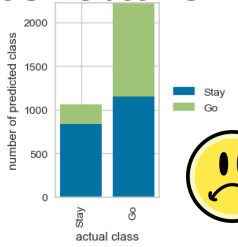


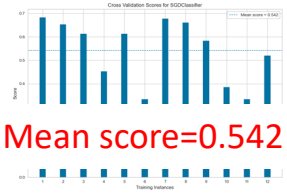
Extra Trees Classifier



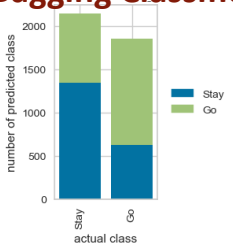


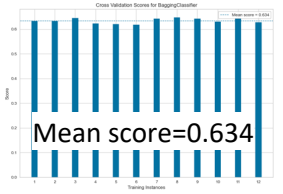
SGD Classifier



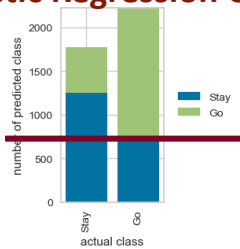


Bagging Classifier

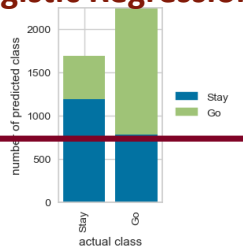




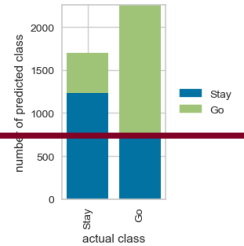
Logistic Regression CV

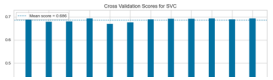


Logistic Regression

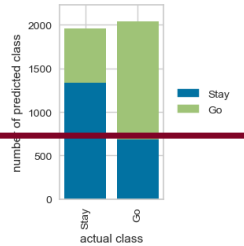


SVC

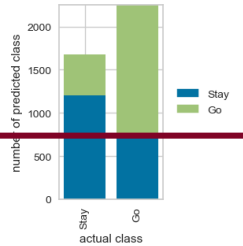


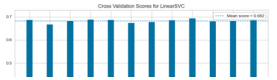


Random Forest Classifier

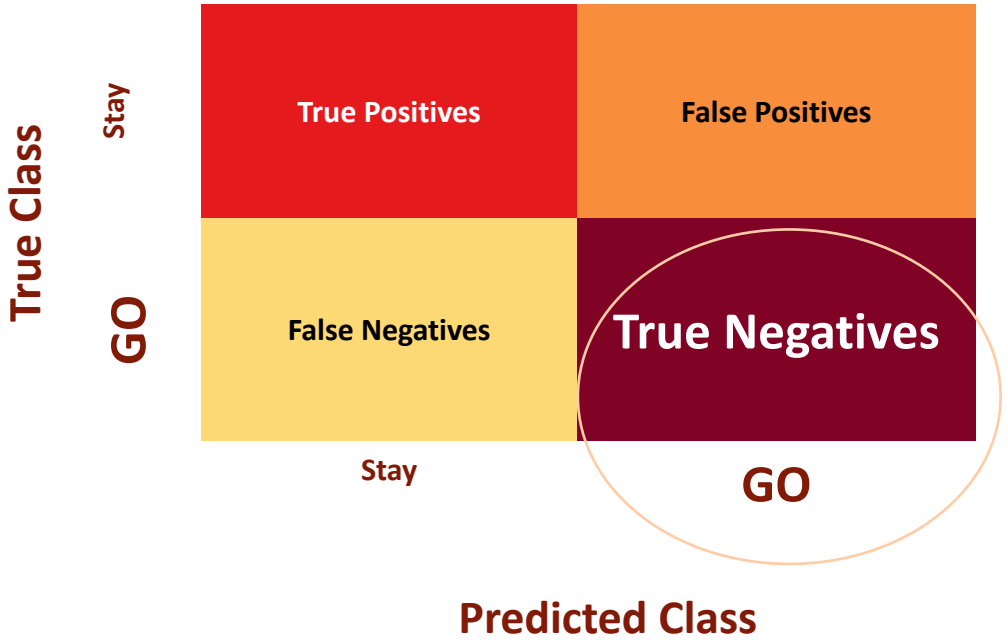


Linear SVC

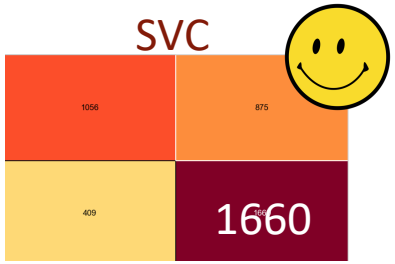




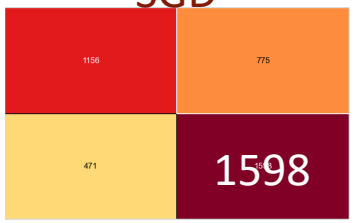
We examined confusion matrices to compare the models' abilities to accurately predict "GO"



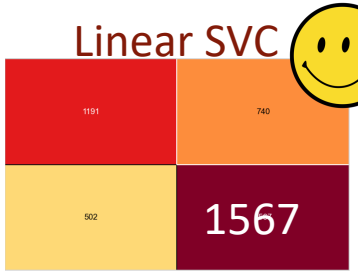
SVC



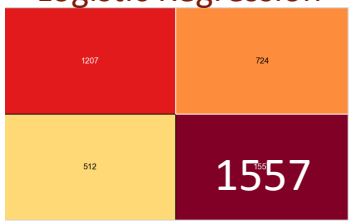
SGD



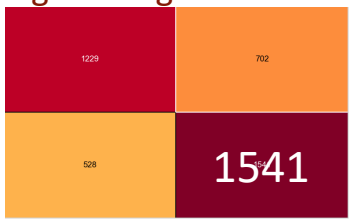
Linear SVC



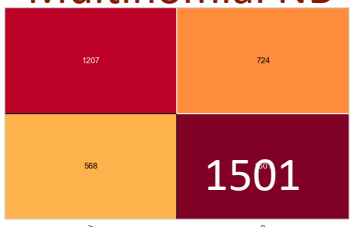
Logistic Regression



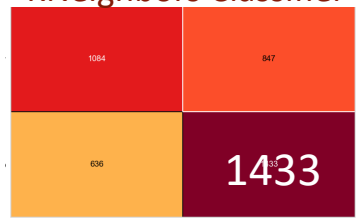
Logistic Regression CV



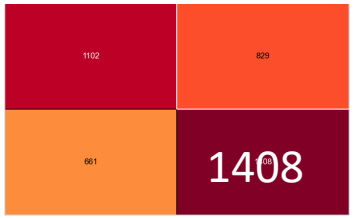
Multinomial NB



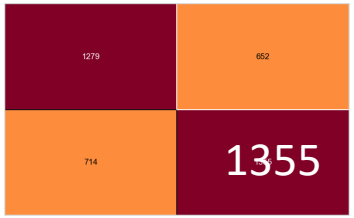
KNeighbors Classifier



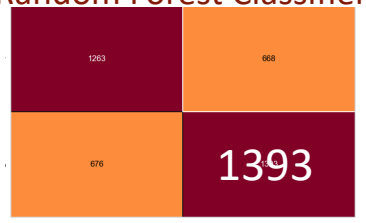
NuSVC



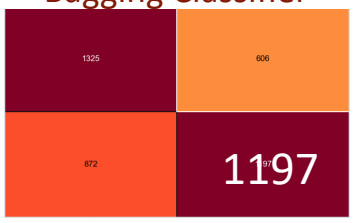
Extra Trees Classifier



Random Forest Classifier

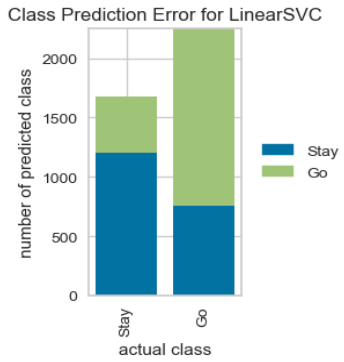


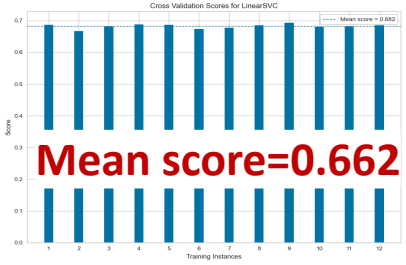
Bagging Classifier



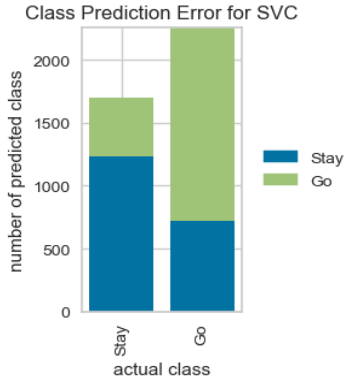
Two models performed strongest: Linear SVC and SVC

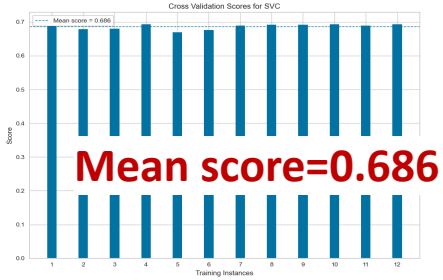
Linear SVC



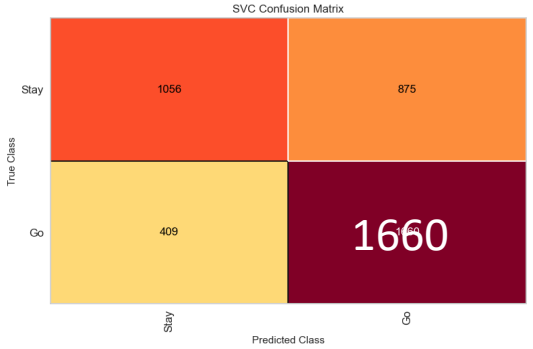


SVC

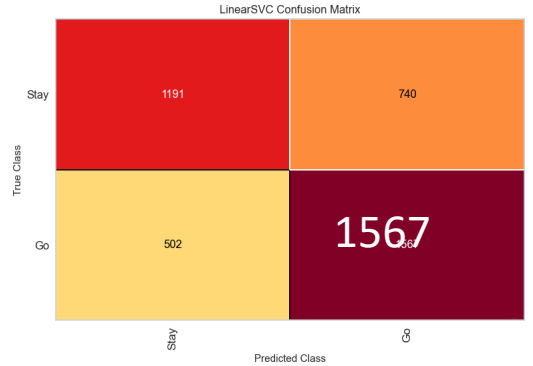




SVC



Linear SVC



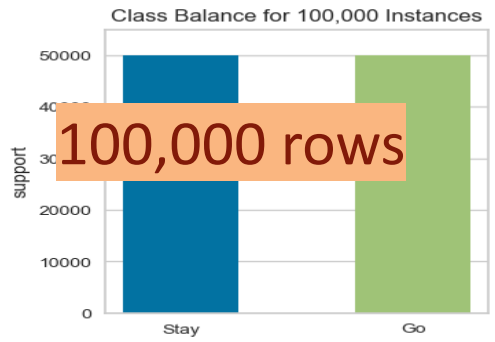
SVC, or Support Vector Classifier, is **a supervised machine learning algorithm typically used for classification tasks**. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.

Reporting and Visualization

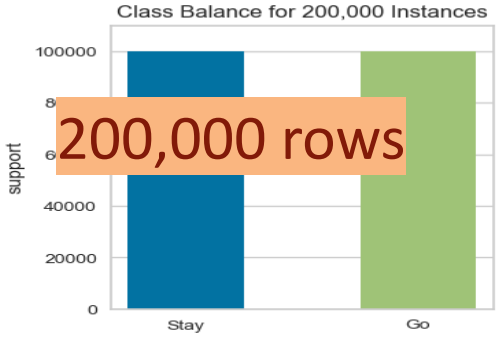
Linear SVC as Our Ideal Model

Linear SVC

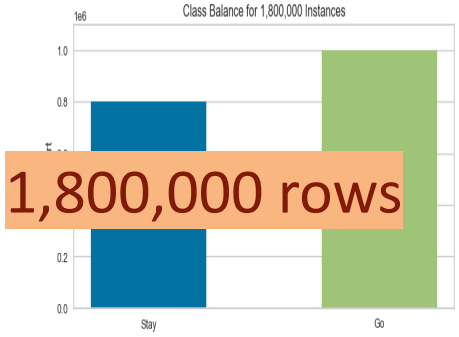
Data Sampled



100,000 rows

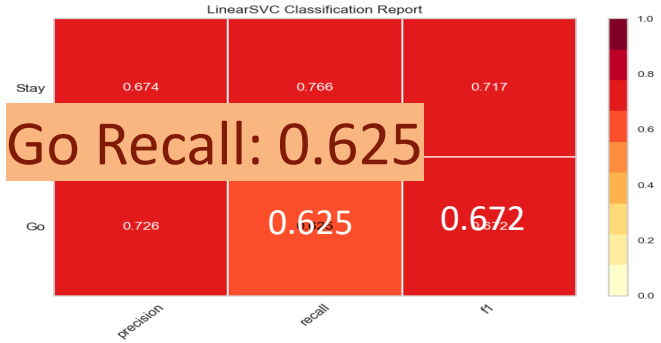


200,000 rows

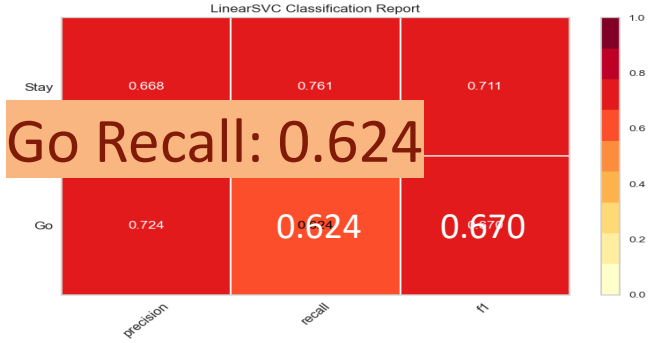


1,800,000 rows

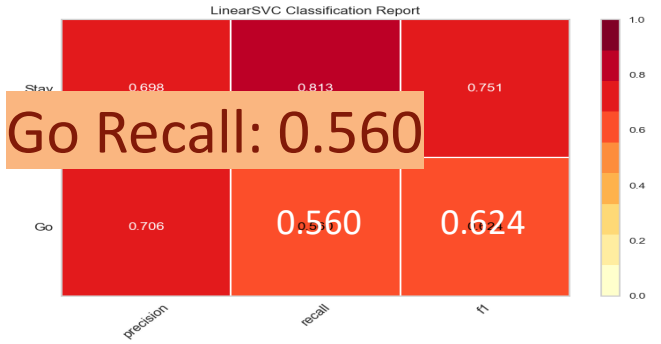
Classification Reports



Go Recall: 0.625

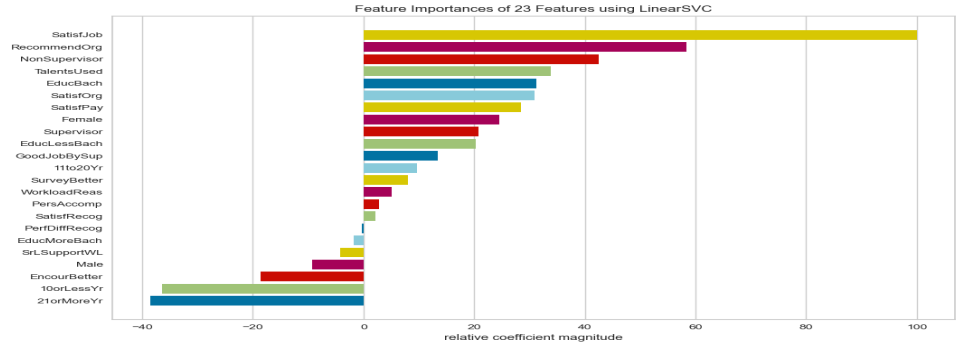


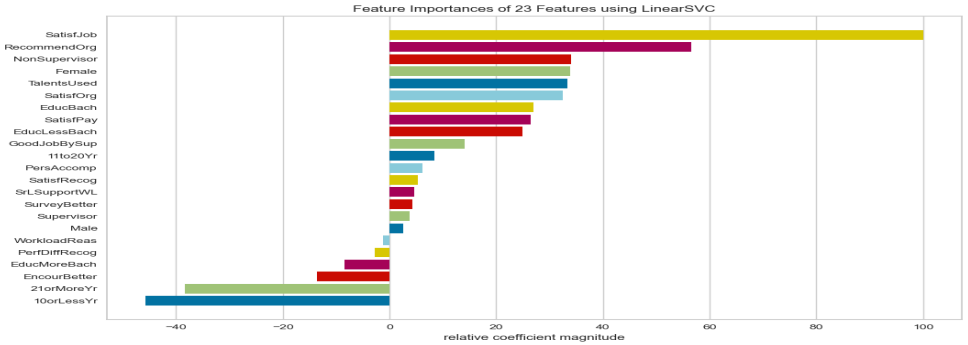
Go Recall: 0.624

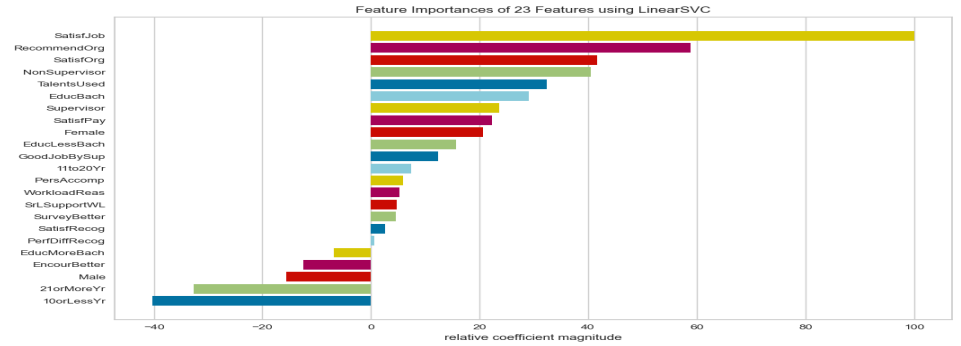


Go Recall: 0.560

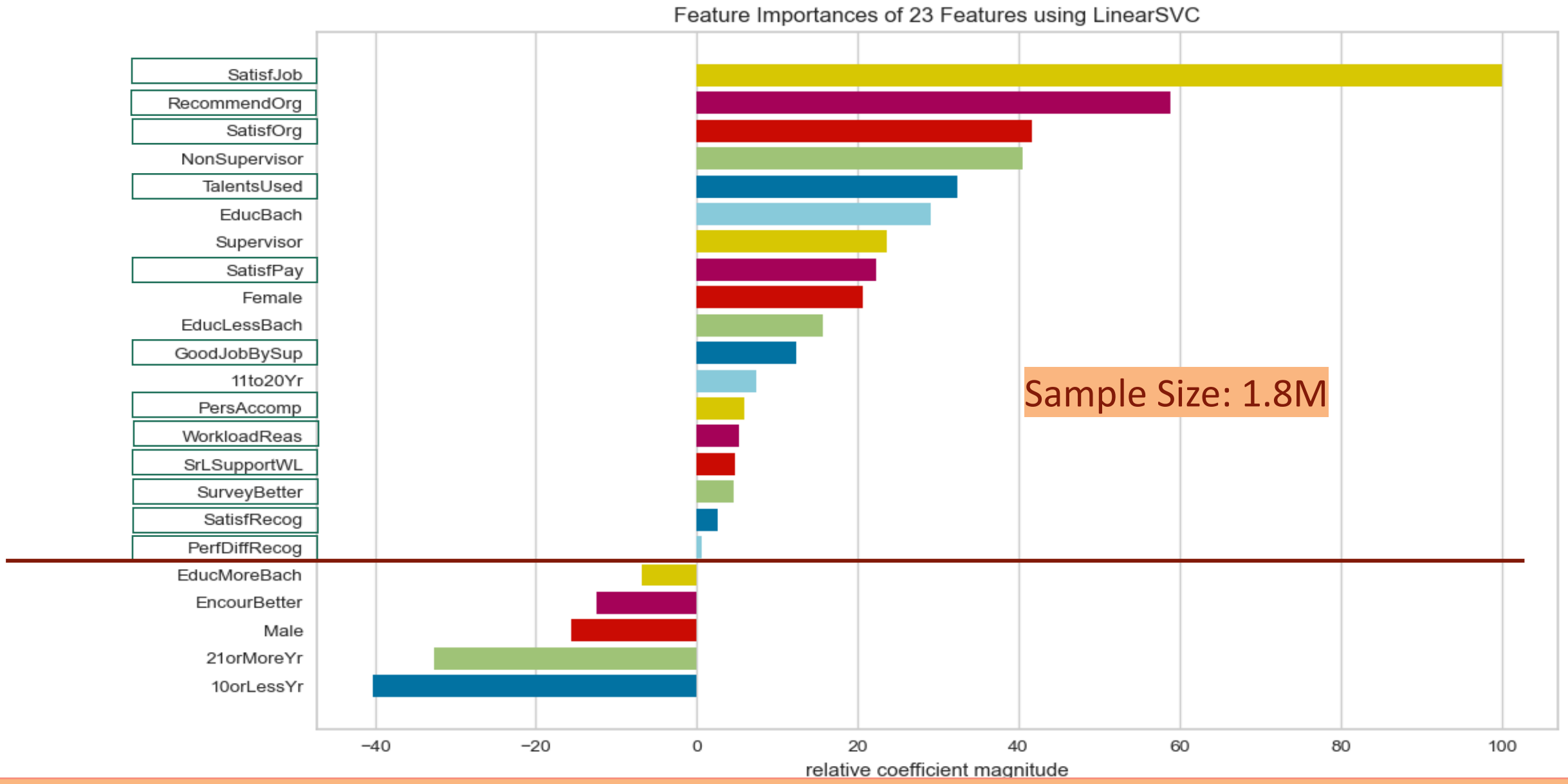
Feature Importances







Reporting and Visualization

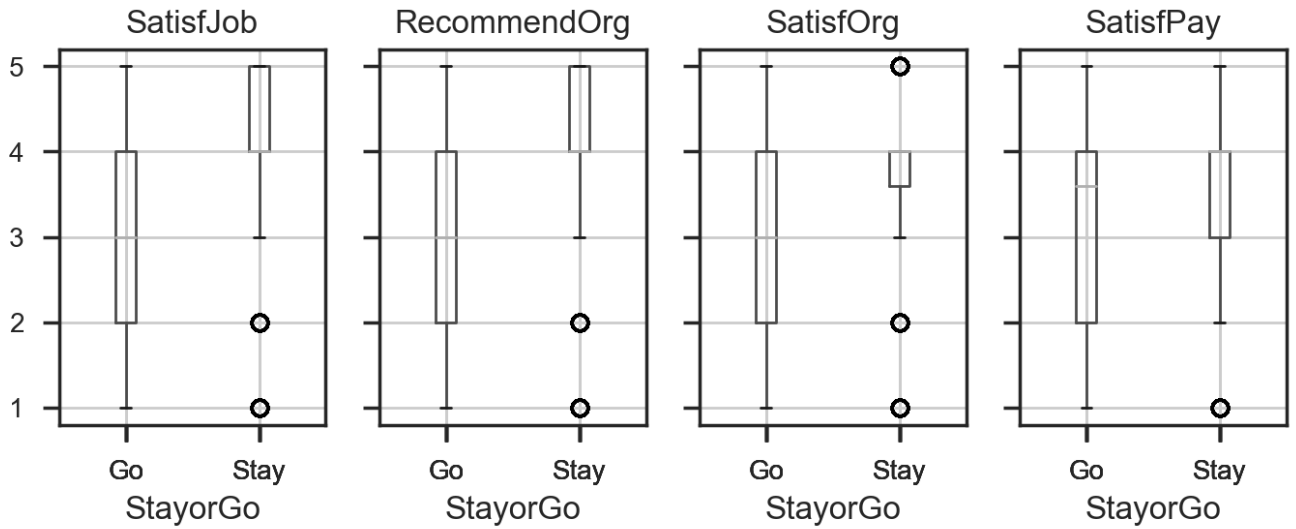


Back to Computation and Analysis…

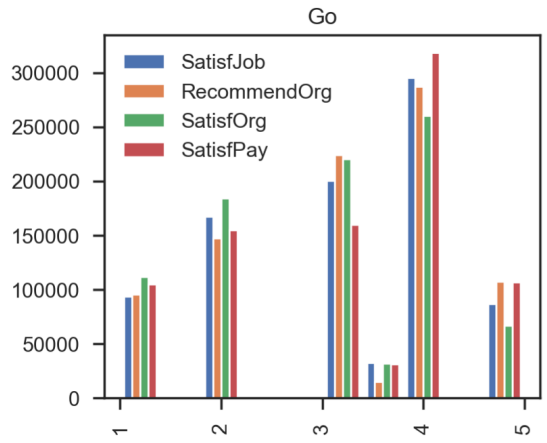
Employees who rate low on Global Satisfaction questions are more likely to indicate GO

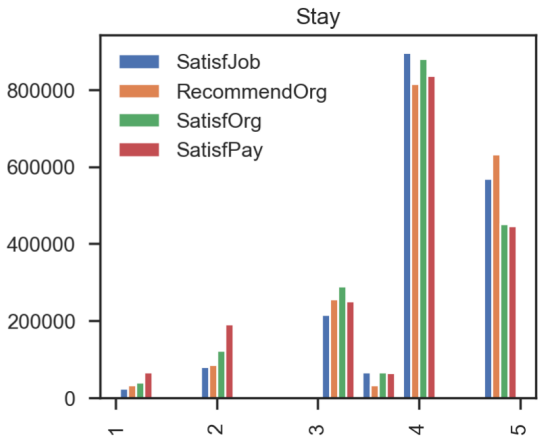
The Global Satisfaction Index is an average of the scores of the four items below:

Job Satisfaction	Pay Satisfaction	Organizational Satisfaction	Recommend Organization
Considering everything, how satisfied are you with your job? (Q. 69)	Considering everything, how satisfied are you with your pay? (Q. 70)	Considering everything, how satisfied are you with your organization? (Q. 71)	I recommend my organization as a good place to work. (Q. 40)



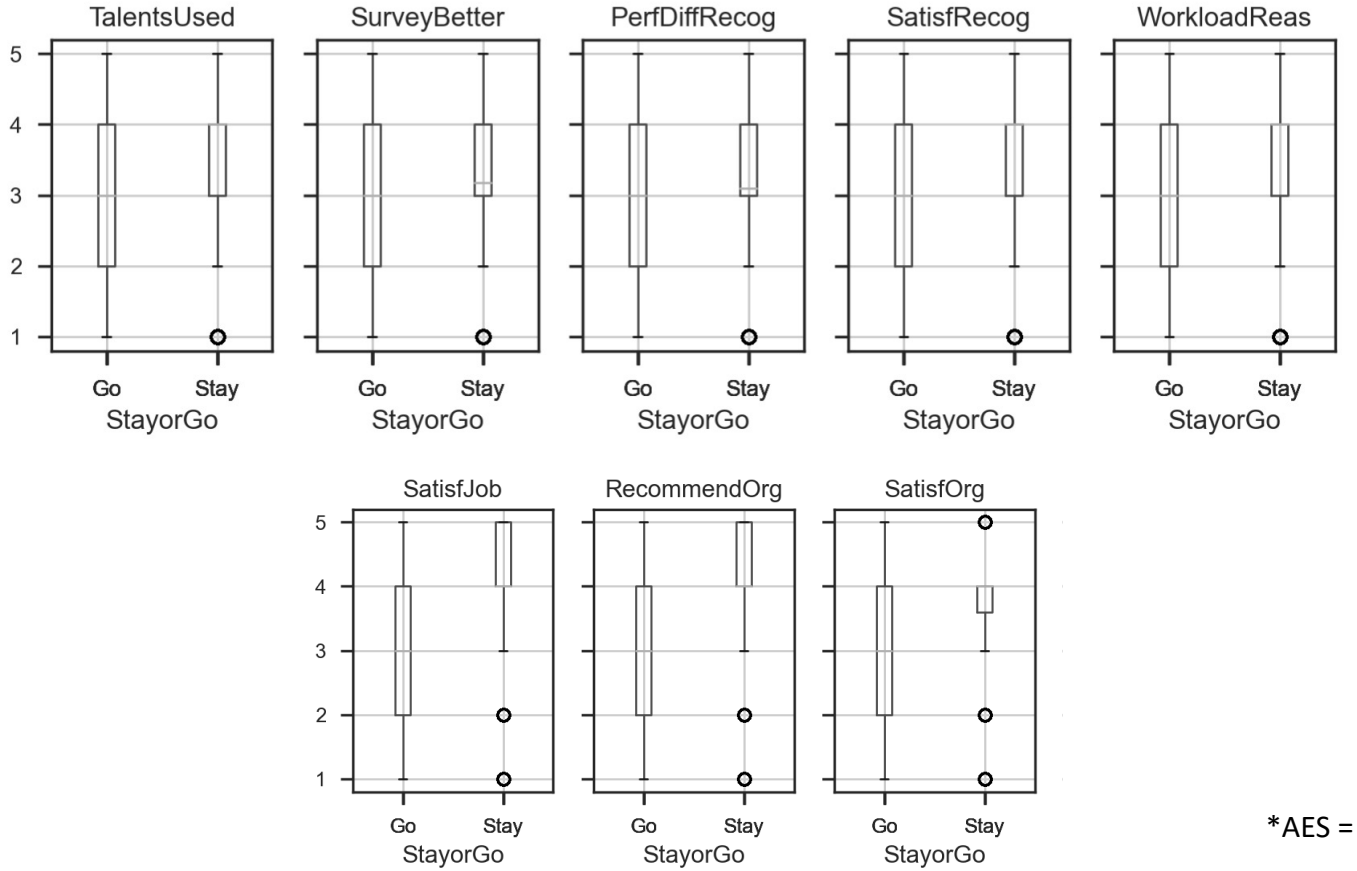
GSI, AES | GSI, AES | GSI, AES | GSI only



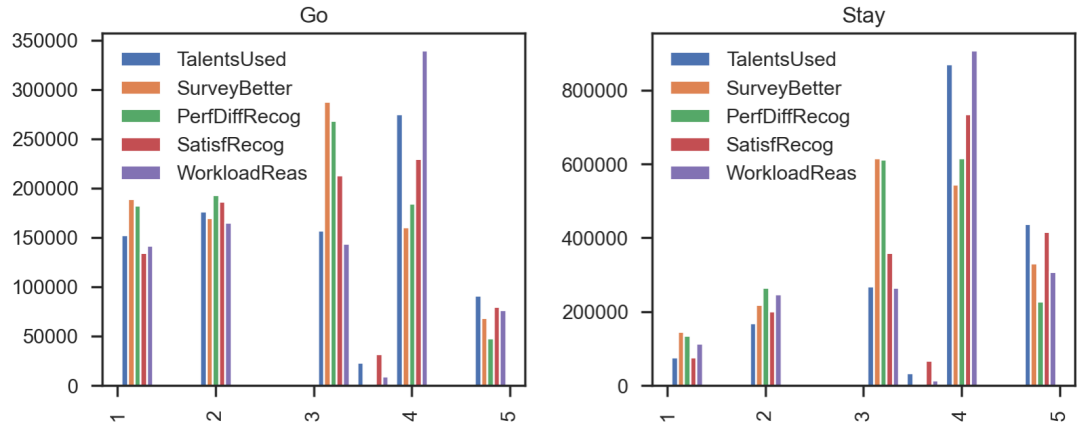


(All 4 GSI Questions, n = 2.7M)

Eight out of 16 AES* questions contribute to our ML model in predicting intent to say GO

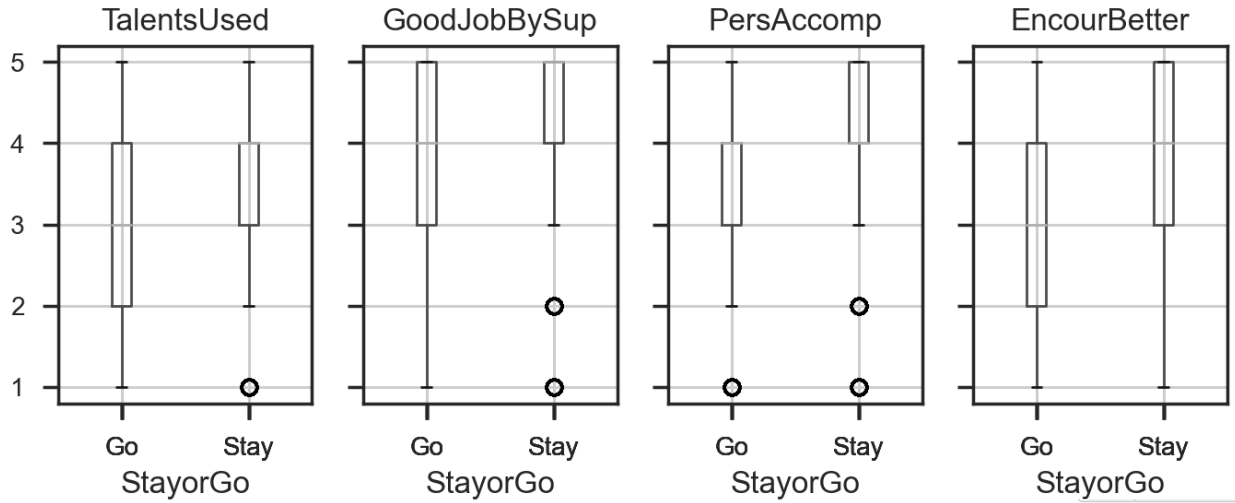


- My talents are used well in the workplace
- I believe the results of this survey will be used to make my agency a better place to work
- In my unit, differences in performance are recognized in a meaningful way.
- How satisfied are you with the recognition you receive for doing a good job?
- My workload is reasonable.

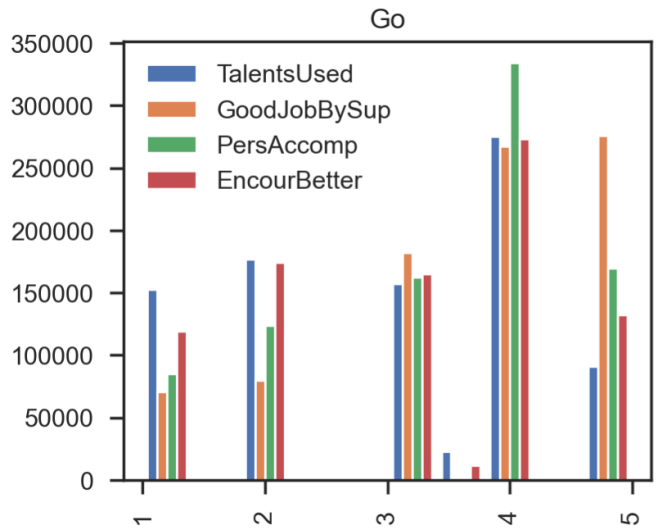


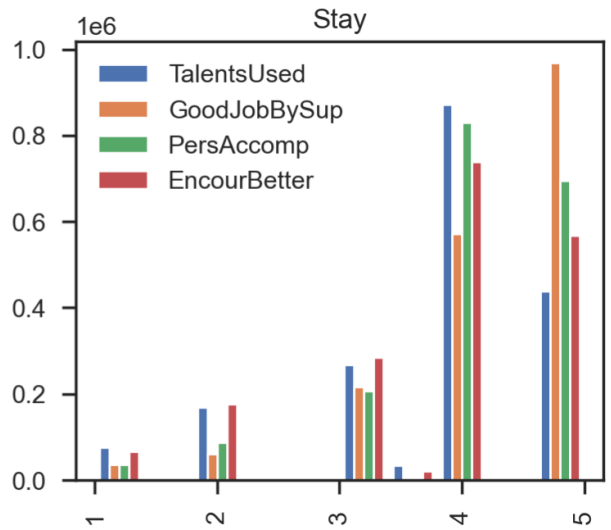
*AES = Annual Employee Survey mandated by Congress

Individuals who score low on certain other factors are likely to indicate GO

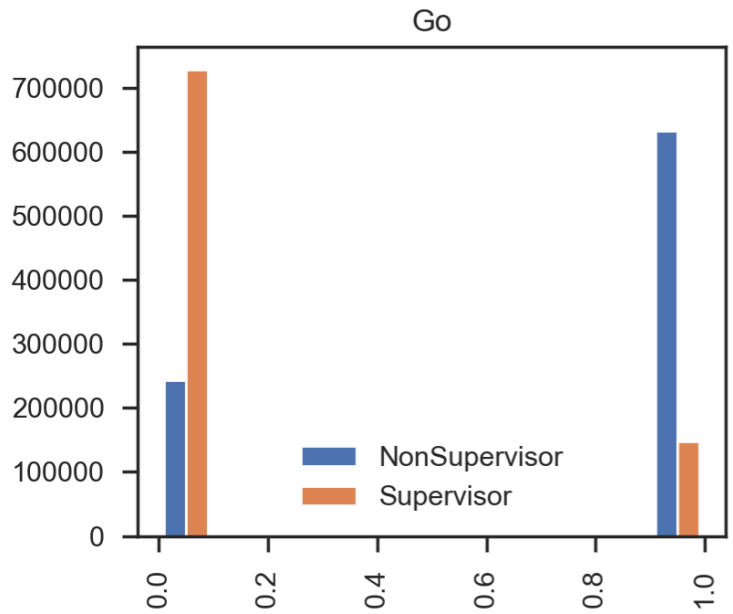


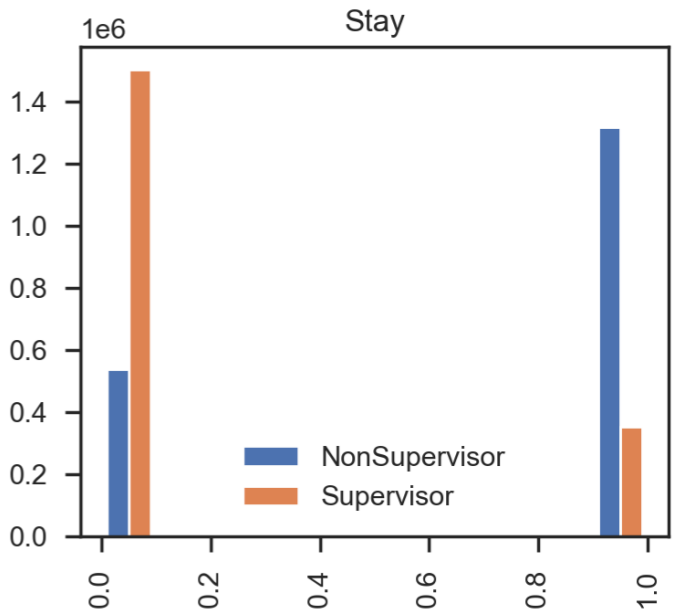
- Overall, how good a job do you feel is being done by your immediate supervisor?
- My work gives me a feeling of personal accomplishment
- I feel encouraged to come up with new and better ways of doing things.



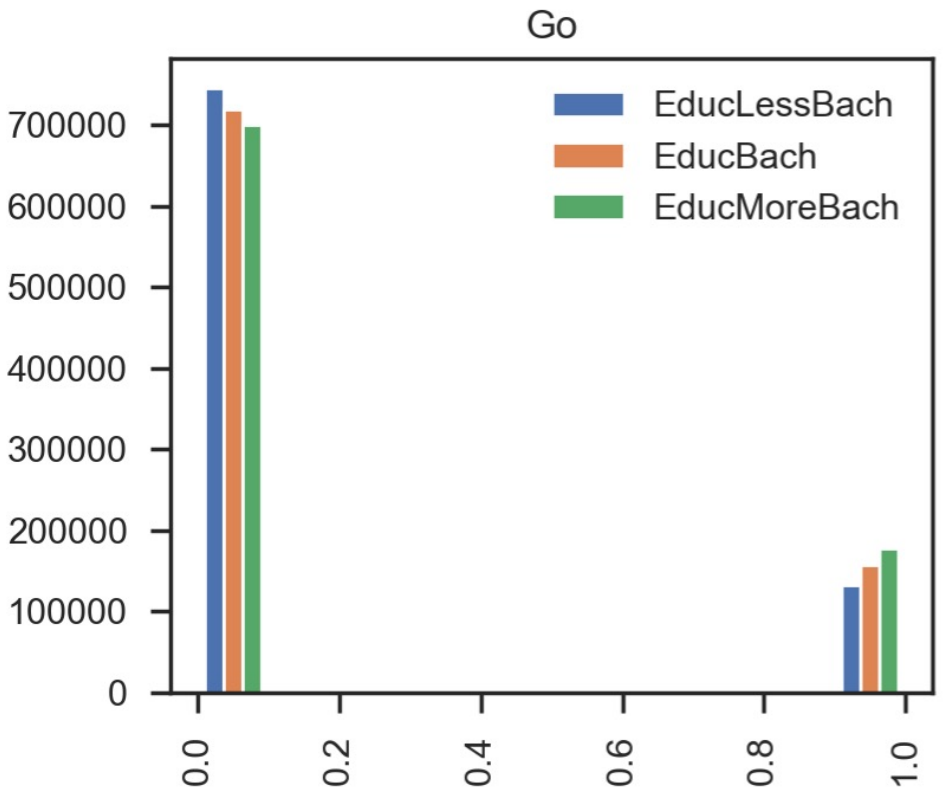


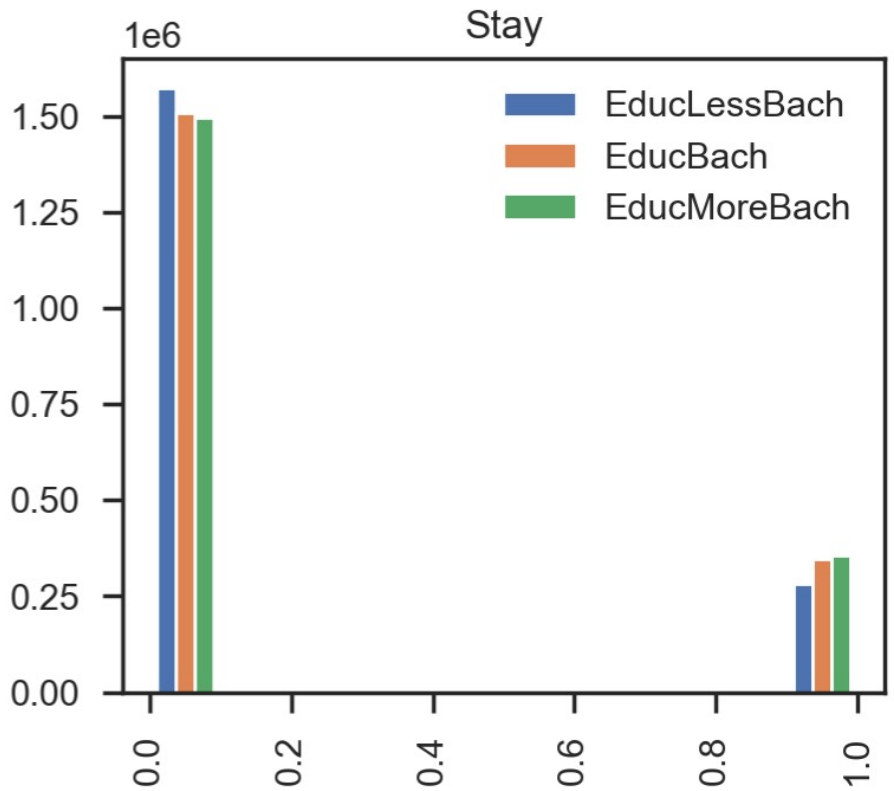
One's Supervisor status does not seem to play a critical role in predicting GO or Stay





Education may or may not contribute to whether employee will choose GO or Stay





Project Results

- **Primary:**
  - Use Machine Learning to analyze the FEVS and identify the factors that predict an employee choosing “Leave” on the Intent to Leave Question

Initial Hypothesis: By analyzing the FEVS data using Machine Learning, we will identify the factors that can predict someone answering “Yes, to take another job outside the Federal Government” to this question (i.e., “Leave”).

Outcome: **Yes**, by analyzing the FEVS data using Machine Learning, we identified the most important factors that can predict an employee’s intent to leave.

Project Results

- **Secondary:**
  - Validate (clarify/debunk) any assumed relationship between the Indexes and answers on Intent to Leave question
    - **Validated** GSI (100%) and (AES 50%) as predictors of Intent to Leave
    - **No support for** EEI's presumed relationship to Intent to Leave
  - Develop a Web Survey or App that could allow federal managers to predict workforce availability (sooner), or even intervene
    - (Not started)

Practical Application

- Best options for for predicting whether employees will say they intend to leave:
  - ✓ Global Satisfaction Index and the answers to all 4 questions
  - ✓ Answers to 8 out of the 16 Annual Employee Survey (AES) questions
- The Employee Engagement Index (EEI) is not as useful for predicting intent to leave
  - ✗ Only four of the 15 questions made it into our top 13, and even those were in our bottom 8.

Limitations

- Worked with sampled data
  - Up to 1.8M/2.7M rows
- Collapsing StayorGo to binary might have diluted our focus “Leave”
  - “Not-Stay” may need to be compared to a different turnover metric (e.g., Total Separations)
- Results are valid for federal government as a whole
  - We did not use weighting factors supplied in the Codebooks to break out results for individual agencies

Future Opportunities

- Analyze additional “Leave” questions asked during:
  - Shutdown in 2019
  - Covid in 2020
- Use the weighting variables to compare results across different agencies

Lessons Learned

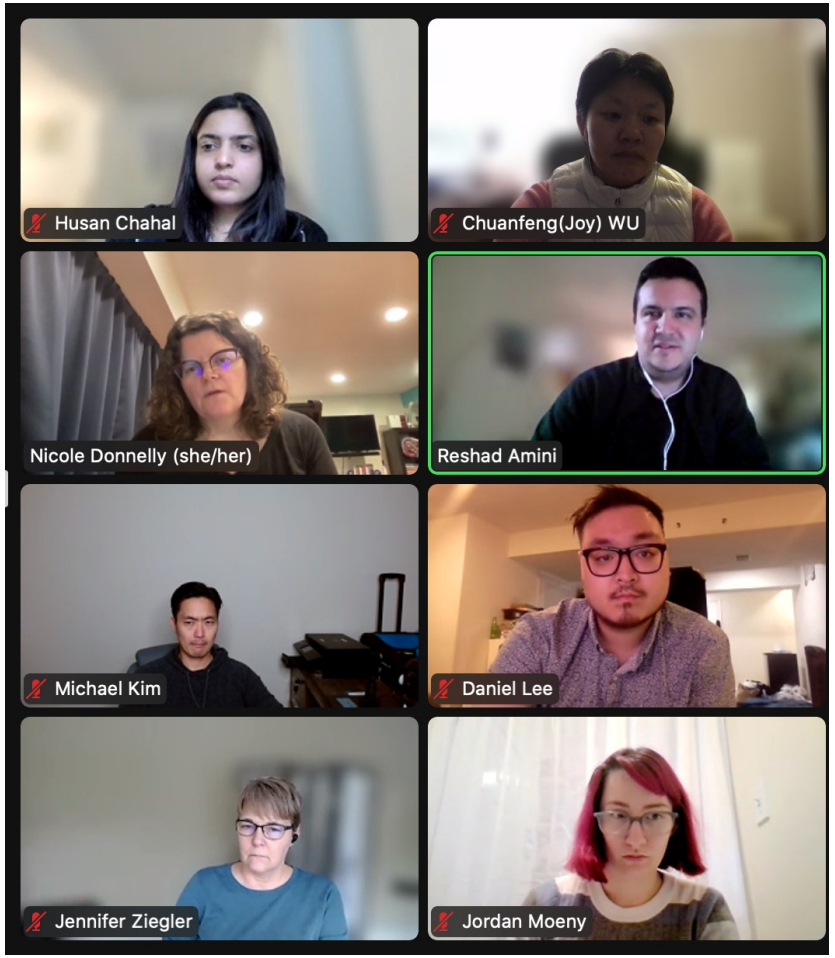
Related to Project Goals

- Try harder to obtain the raw data set to view finer demographics
- Setting up a SQL database was a good experience but we didn't really use it

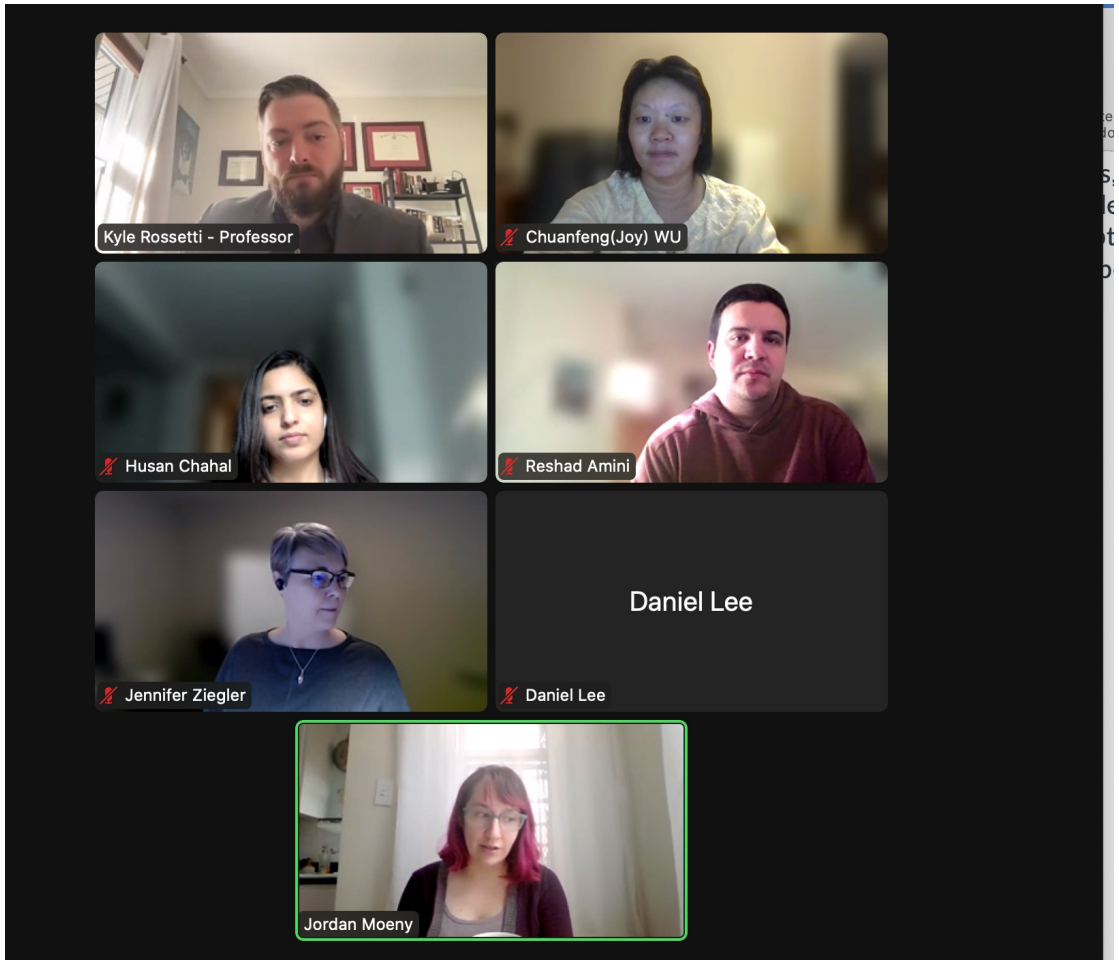
Related to Data Science

- Data wrangling takes longer than we expected
- Data balance impacts Machine Learning model performance
- Feature selection and analysis are iterative with Machine Learning
- To evaluate a Machine Learning model, multiple factors need to be considered (e.g., F1 scores, cross validation scores, confusion matrices, etc.)
- Data Science is a community: Ask for help, Google it, and rely on your colleagues and instructors for help

Great to meet you all! Please STAY in touch!



XBUS-500



XBUS-507

Questions?

Idea and Inspiration

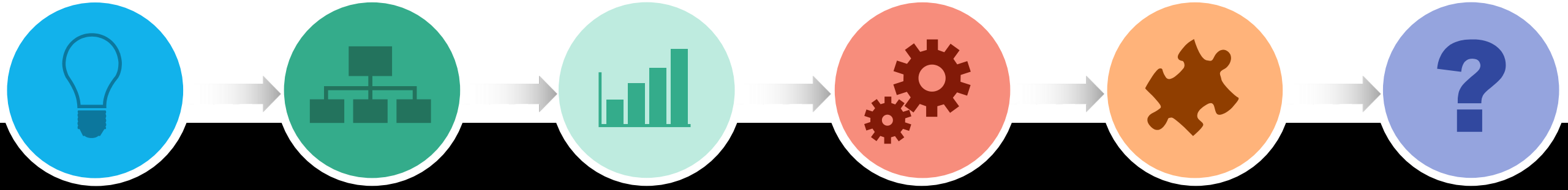
Data Ingestion and Wrangling

Computation and Analysis

Modeling and Feature Selection

Reporting and Visualization

Conclusion and Q&A



Thank you to Adam, Daniel, our classmates, and all of our Cohort 27 faculty!

Chuanfeng “Joy” Wu
Jennifer Ziegler

Georgetown University Data Science Certificate Cohort 27

September 24, 2022

Appendix
Additional Slides if Needed for Discussion

Feature Importances Questions by Index

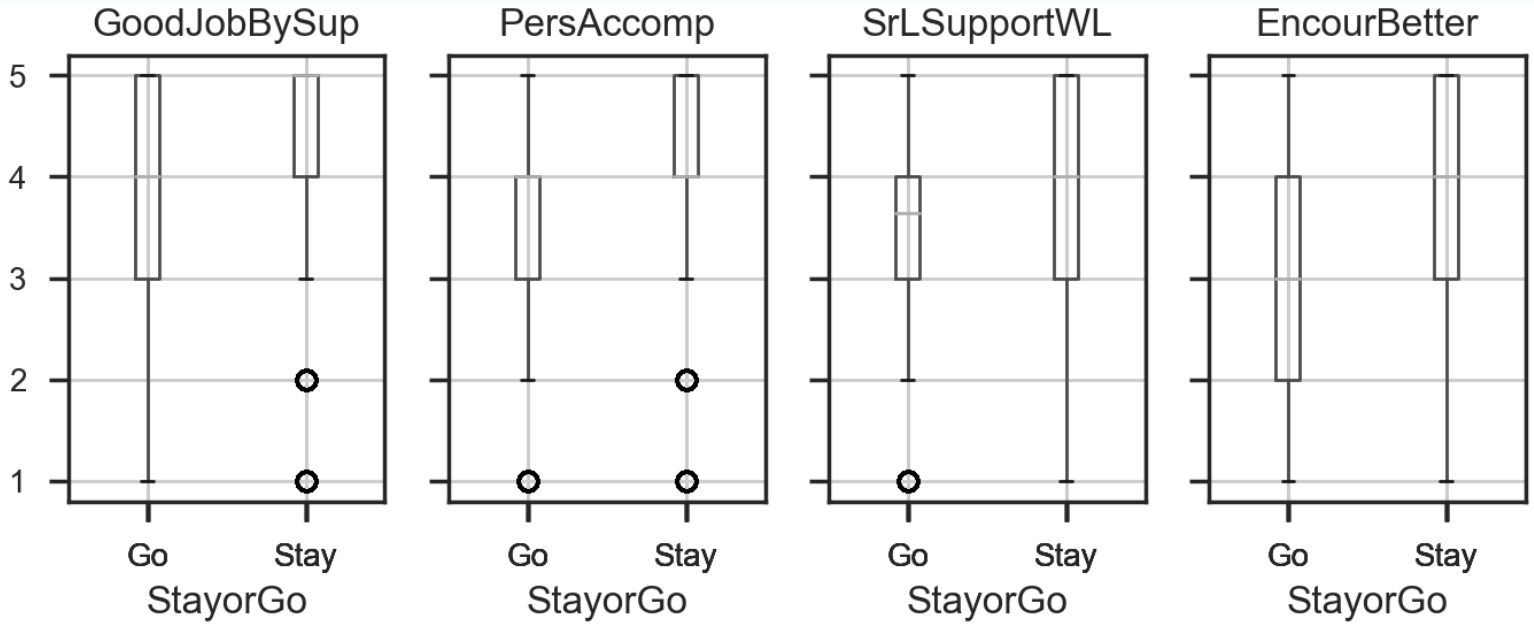
AES							
AES		GSI			SatisfJob	1	Considering everything, how satisfied are you with your job? 5 = Very Satisfied … 1 = Very Dissatisfied
AES		GSI			RecommendOrg	2	I recommend my organization as a good place to work. 5 = Strongly Agree … 1 = Strongly Disagree
AES		GSI			SatisfOrg	3	Considering everything, how satisfied are you with your organization? 5 = Very Satisfied … 1 = Very Dissatisfied
AES	IWE		EMPO		TalentsUsed	5	My talents are used well in the workplace. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
AES					WorkloadReas	14	My workload is reasonable. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
AES					SurveyBetter	16	I believe the results of this survey will be used to make my agency a better place to work. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
AES					SatisfRecog	17	How satisfied are you with the recognition you receive for doing a good job? 5 = Very Satisfied … 1 = Very Dissatisfied
AES			FAIR		PerfDiffRecog	18	In my work unit, differences in performance are recognized in a meaningful way. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
	(EEI)						
AES	IWE		EMPO		TalentsUsed	5	My talents are used well in the workplace. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
	SUPE				GoodJobBySup	11	Overall, how good a job do you feel is being done by your immediate supervisor? 5 = Very Good … 1 = Very Poor
	IWE				PersAccomp	13	My work gives me a feeling of personal accomplishment. 5 = Strongly Agree … 1 = Strongly Disagree
	IWE		EMPO		EncourBetter	20	I feel encouraged to come up with new and better ways of doing things. 5 = Strongly Agree … 1 = Strongly Disagree
		GSI					
AES		GSI			SatisfJob	1	Considering everything, how satisfied are you with your job? 5 = Very Satisfied … 1 = Very Dissatisfied
AES		GSI			RecommendOrg	2	I recommend my organization as a good place to work. 5 = Strongly Agree … 1 = Strongly Disagree
AES		GSI			SatisfOrg	3	Considering everything, how satisfied are you with your organization? 5 = Very Satisfied … 1 = Very Dissatisfied
		GSI			SatisfPay	8	Considering everything, how satisfied are you with your pay? 5 = Very Satisfied … 1 = Very Dissatisfied
			(**NIQ**)				
AES	IWE		EMPO		TalentsUsed	5	My talents are used well in the workplace. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
AES			FAIR		PerfDiffRecog	18	In my work unit, differences in performance are recognized in a meaningful way. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
	IWE		EMPO		EncourBetter	20	I feel encouraged to come up with new and better ways of doing things. 5 = Strongly Agree … 1 = Strongly Disagree

The Entire EEI – the original 15 questions (with the 4 reflected in Feature Importances in **bold**)

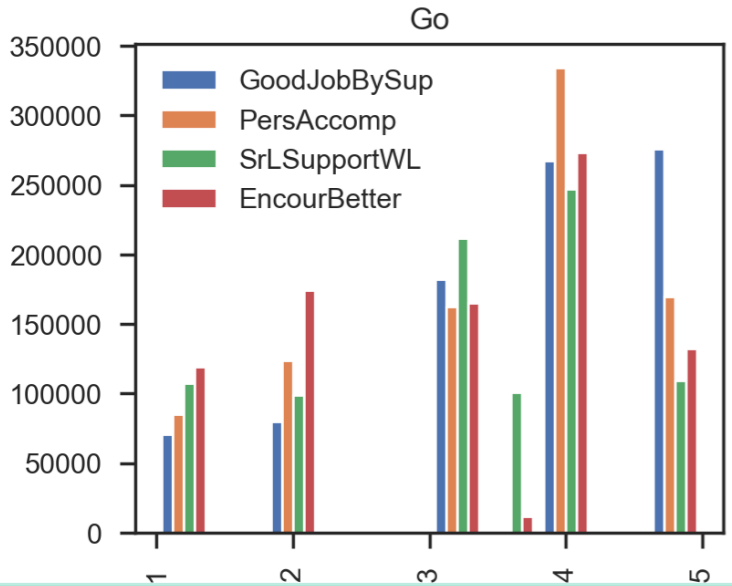
I feel encouraged to come up with new and better ways of doing things.
My work gives me a feeling of personal accomplishment.
I know what is expected of me on the job.
My talents are used well in the workplace.
I know how my work relates to the agency's goals.
In my organization, senior leaders generate high levels of motivation and commitment in the workforce.
My organization's senior leaders maintain high standards of honesty and integrity.
Managers communicate the goals of the organization.
Overall, how good a job do you feel is being done by the manager directly above your immediate supervisor?
I have a high level of respect for my organization's senior leaders.
Supervisors in my work unit support employee development.
My supervisor listens to what I have to say.
My supervisor treats me with respect.
I have trust and confidence in my supervisor.
Overall, how good a job do you feel is being done by your immediate supervisor?

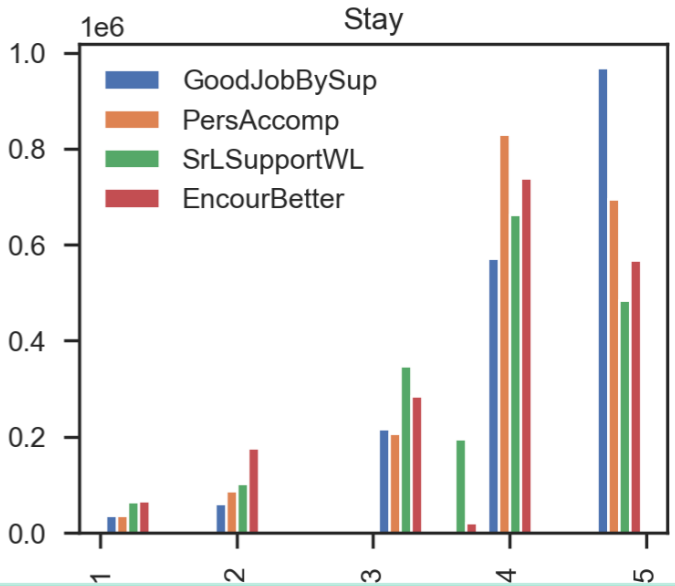
The entire AES – the original 16 questions (with the 8 reflected in feature importances in **bold**)

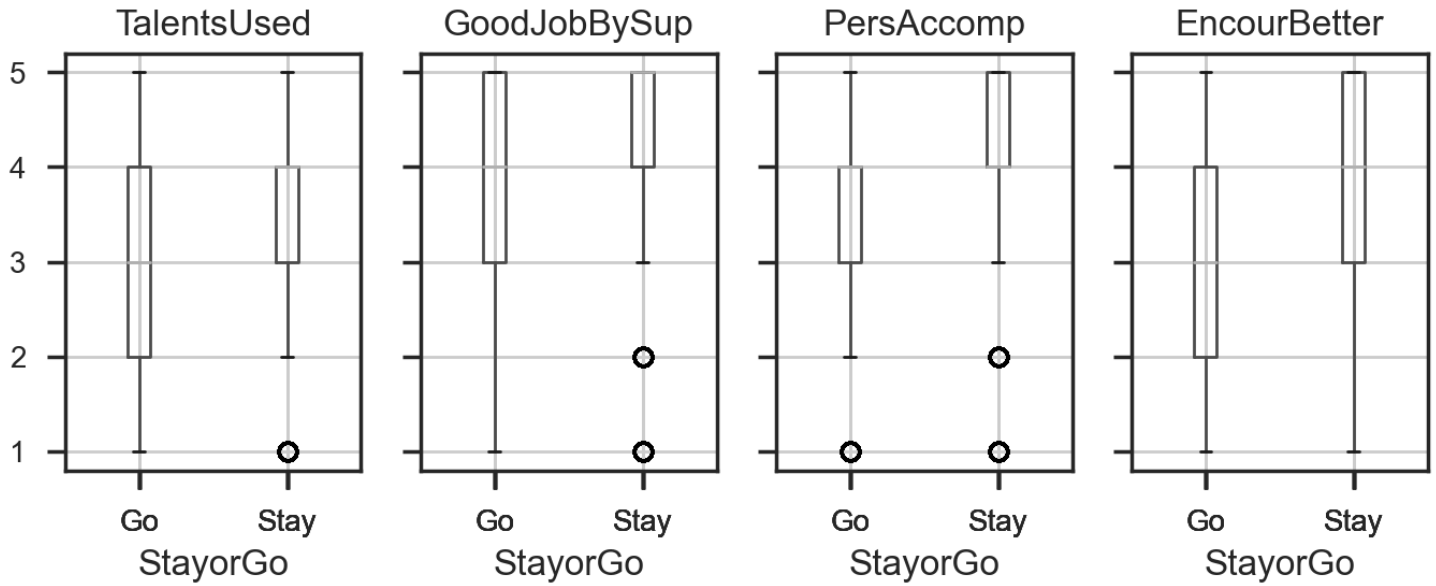
Considering everything, how satisfied are you with your job? 5 = Very Satisfied … 1 = Very Dissatisfied
I recommend my organization as a good place to work. 5 = Strongly Agree … 1 = Strongly Disagree
My talents are used well in the workplace. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
Considering everything, how satisfied are you with your organization? 5 = Very Satisfied … 1 = Very Dissatisfied
I believe the results of this survey will be used to make my agency a better place to work. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
I am given a real opportunity to improve my skills in my organization. 5 = Strongly Agree … 1 = Strongly Disagree
How satisfied are you with your involvement in decisions that affect your work? 5 = Very Satisfied … 1 = Very Dissatisfied
How satisfied are you with the recognition you receive for doing a good job? 5 = Very Satisfied … 1 = Very Dissatisfied
The people I work with cooperate to get the job done. 5 = Strongly Agree … 1 = Strongly Disagree
My workload is reasonable. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
How satisfied are you with the information you receive from management on what's going on in your organization? 5 = Very Satisfied … 1 = Very Dissatisfied
I can disclose a suspected violation of any law, rule or regulation without fear of reprisal. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
In my work unit, differences in performance are recognized in a meaningful way. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
Managers communicate the goals of the organization. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
My work unit has the job-relevant knowledge and skills necessary to accomplish organizational goals. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know
I know how my work relates to the agency's goals. 5 = Strongly Agree … 1 = Strongly Disagree; X = Do Not Know



Misc. Questions n=2.7M not in any other index except EEI (3Q); see next

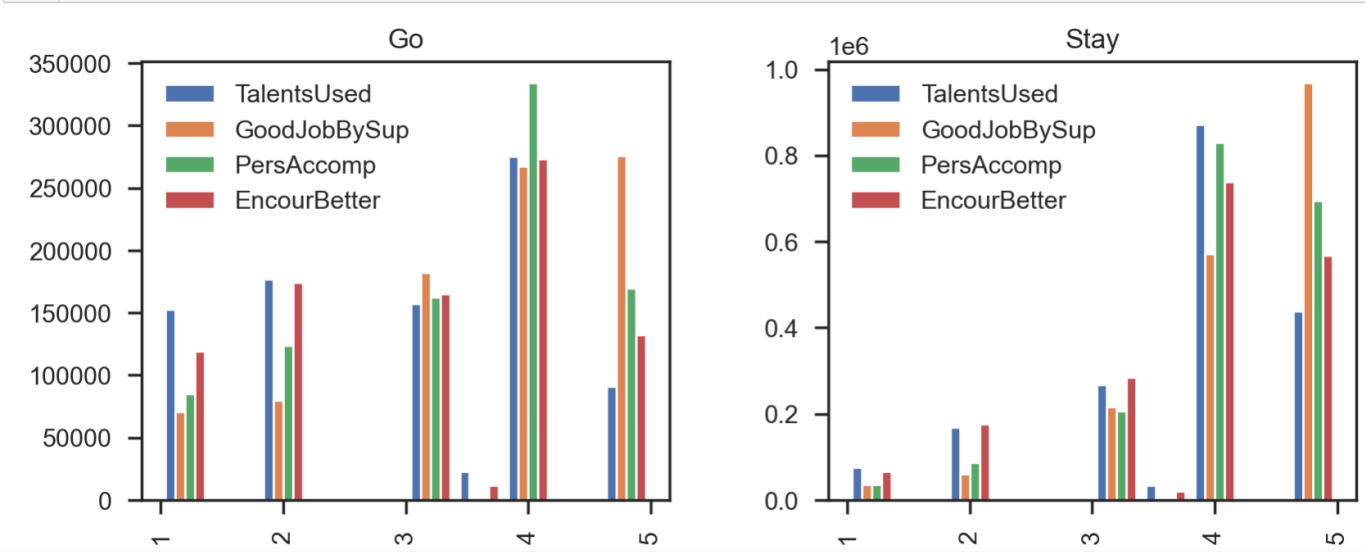


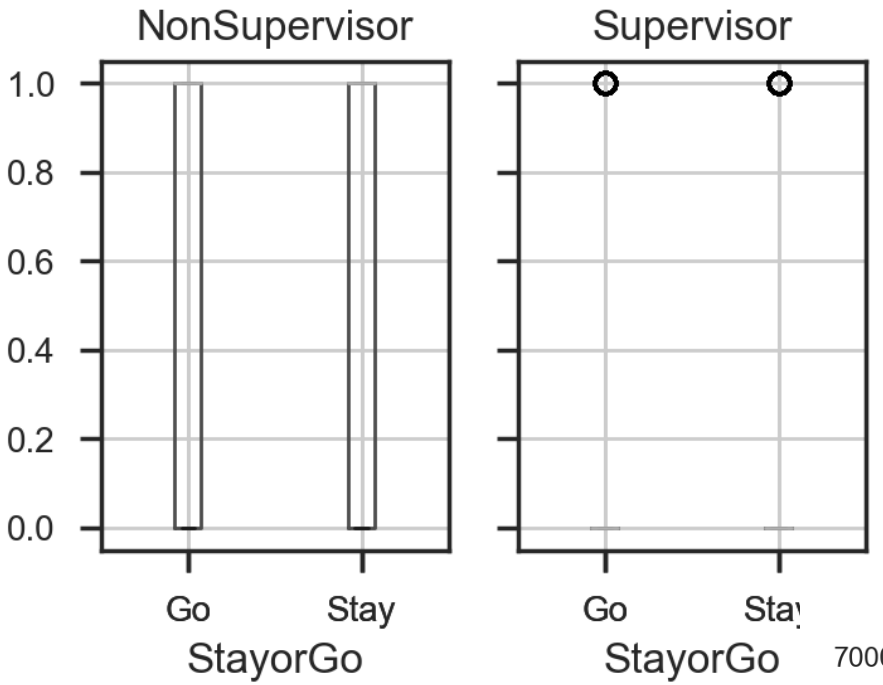




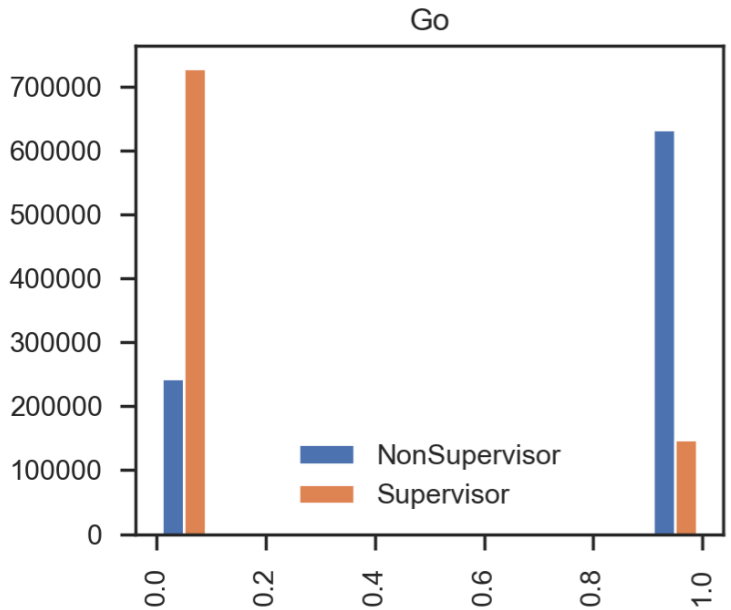
Only four out of 15 EEI questions made it into our top 13
n=2.7M

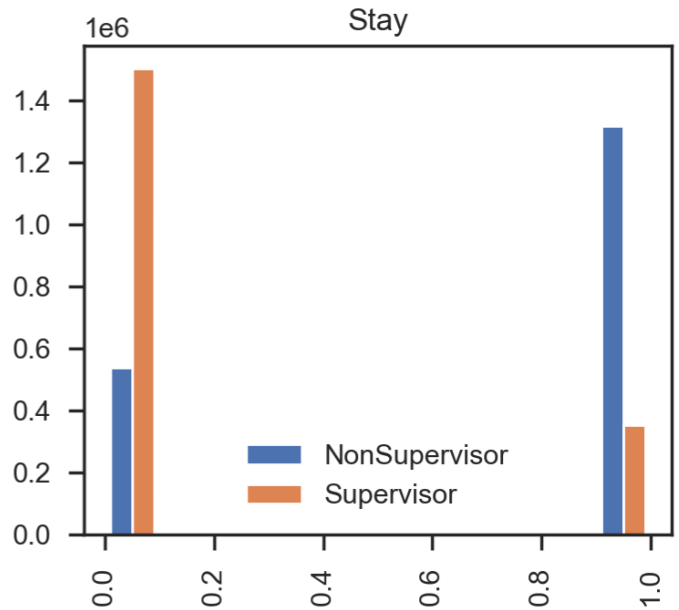
- There are 15 total EEI questions.
- Our features had a total of 4 AES questions (27%) in our top 13; none in the top four.
- Yet the only reporting on the Stay or Go question during the time period for our survey was in terms of how it correlated with EEI.
- (This changed in 2021 when OPM compared Leave to GSI scores.)

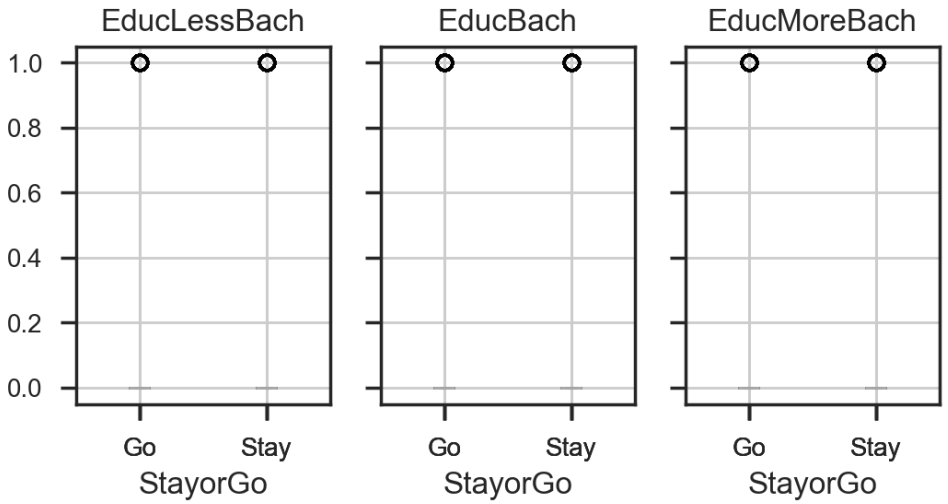




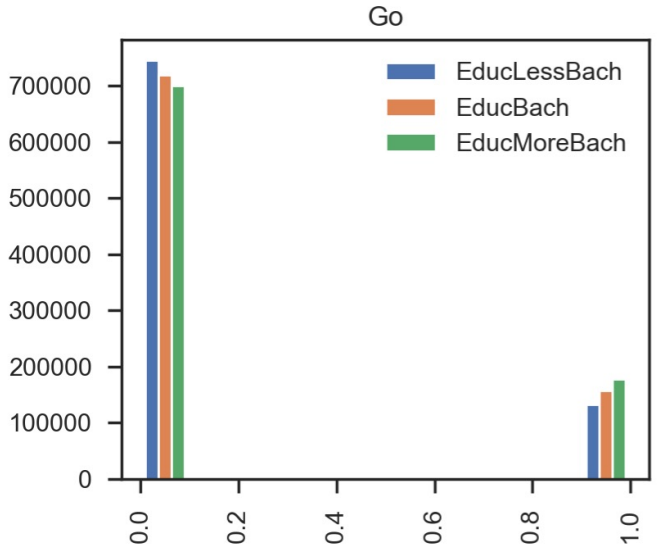
Supervisor vs. Non-Supervisor n=2.7M

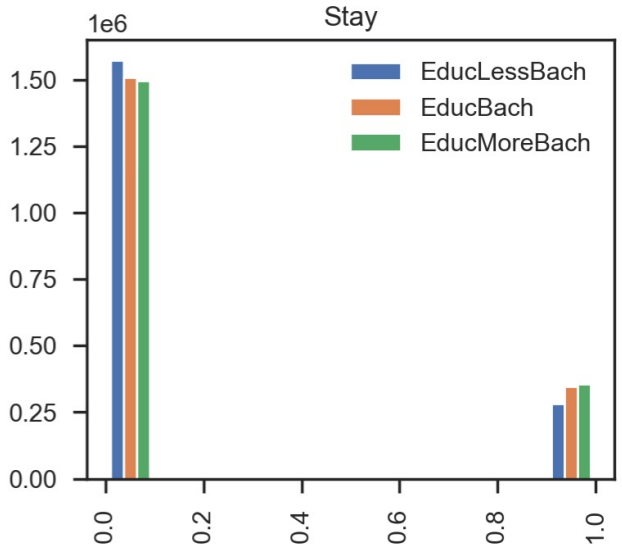


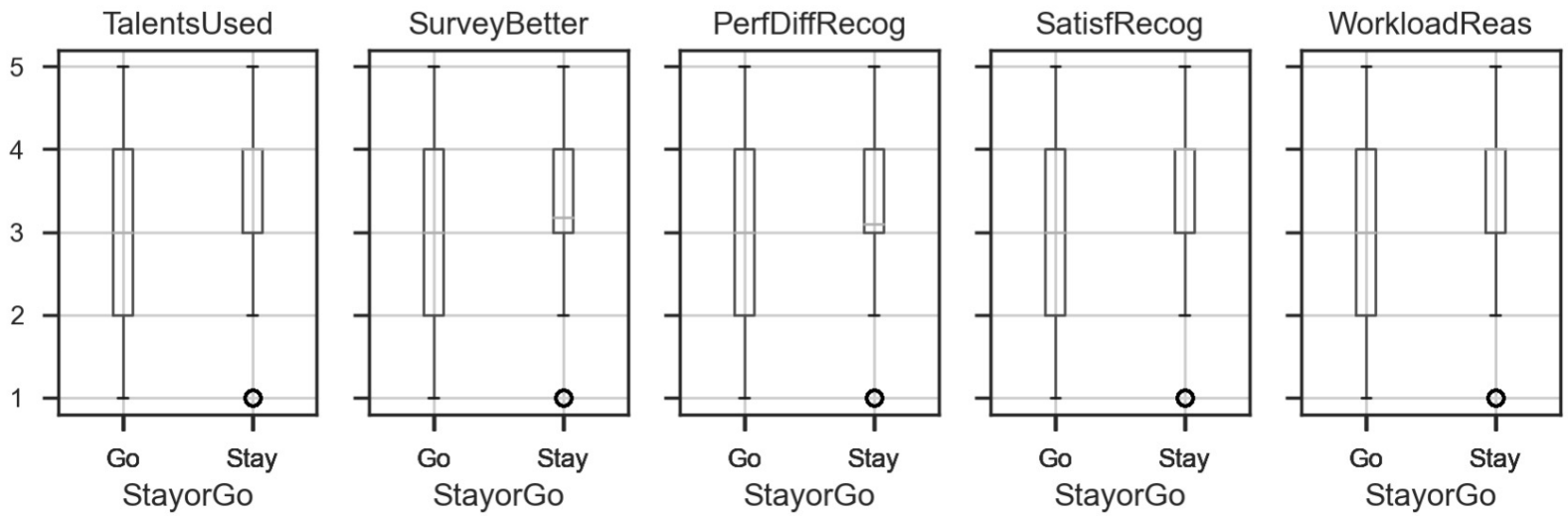




# Education Levels
n= 2.7M

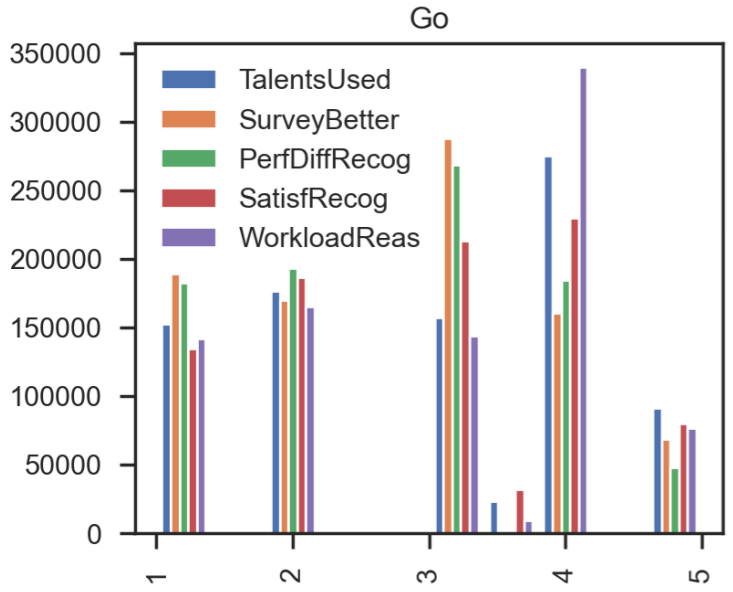


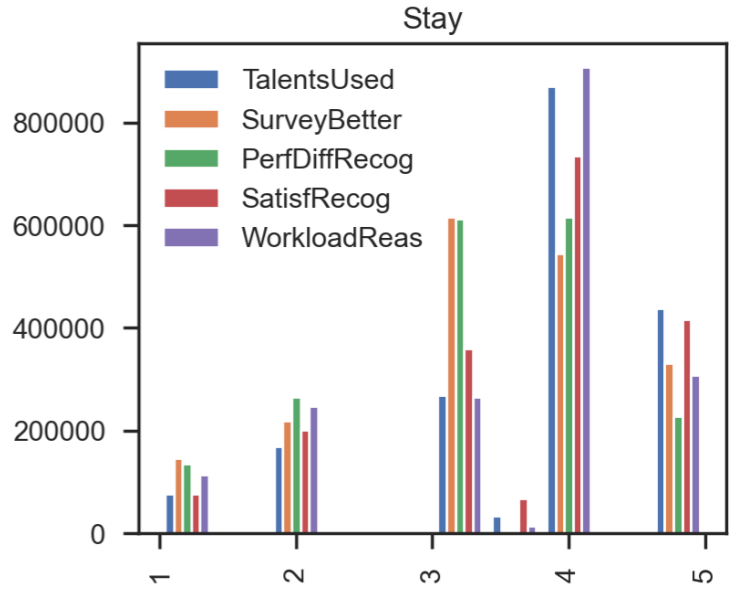




Other Questions in AES, n= 2.7M

There are 16 total AES questions. Our features had a total of 8 AES questions (50%) in our top 13.

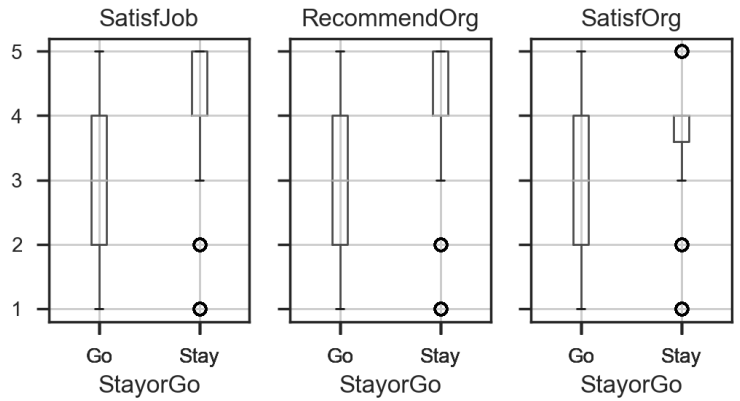


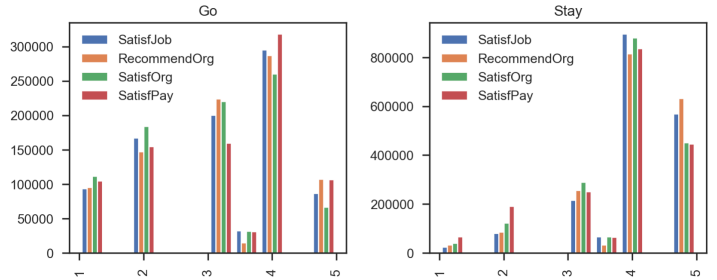


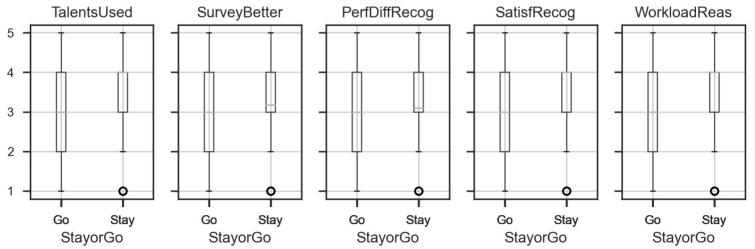
Computation and Analysis

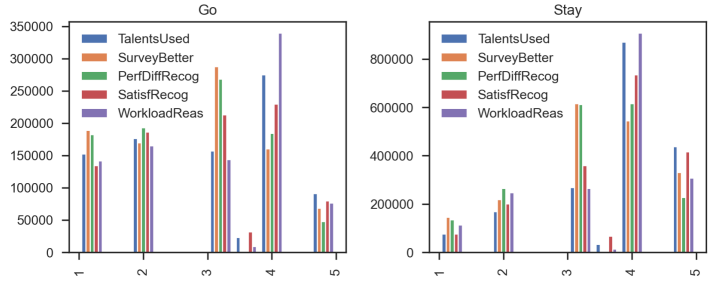
Item	2016	2017	2018	2019	2020
Leadership and Management Practices That Contribute to Agency Performance					
My work unit has the job-relevant knowledge and skills necessary to accomplish organizational goals. (Q. 13)	69	71	80	81	82
Managers communicate the goals of the organization. (Q. 28)	60	62	64	65	68
I believe the results of this survey will be used to make my agency a better place to work. (Q. 18)	41	42	41	41	43
Employee Satisfaction with... Leadership Policies and Practices					
How satisfied are you with your involvement in decisions that affect your work? (Q. 33)	51	53	54	55	58
How satisfied are you with the information you receive from management on what is going on in your organization? (Q. 34)	48	50	51	52	58
Considering everything, how satisfied are you with your organization? (Q. 38)	57	60	60	61	66
The people I work with cooperate to get the job done. (Q. 9)	73	75	76	77	84
My workload is reasonable. (Q. 5)	57	59	59	59	67
Considering everything, how satisfied are you with your job? (Q. 36)	66	68	68	69	72
I can disclose a suspected violation of any law, rule or regulation without fear of reprisal. (Q. 8)	62	64	66	67	68
Employee Satisfaction with... Rewards and Recognition					
In my work unit, differences in performance are recognized in a meaningful way. (Q. 12)	34	36	38	39	51
How satisfied are you with the recognition you receive for doing a good job? (Q. 35)	48	50	52	53	59
Employee Satisfaction with... Opportunities for Professional Development and Growth					
I am given a real opportunity to improve my skills in my organization. (Q. 1)	63	64	66	67	70
My talents are used well in the workplace. (Q. 6)	58	60	60	61	66
Employee Satisfaction with... Opportunity to Contribute to Achieving Organizational Mission					
I know how my work relates to the agency's goals. (Q. 7)	83	84	85	85	87
I recommend my organization as a good place to work. (Q. 17)	64	66	66	67	71

These 8 questions might indicate someone who is ready to say they're going to leave. The other questions might be interesting to ask for other reasons, but they won't necessarily predict exit.









AES = 16 Congress mandated questions (must be asked every year)
Trends from 2020 report

Computation and Analysis

Global Satisfaction Index

The Global Satisfaction Index is an average of the scores of the four items below:

Job Satisfaction	Pay Satisfaction	Organizational Satisfaction	Recommend Organization
Considering everything, how satisfied are you with your job? (Q. 69)	Considering everything, how satisfied are you with your pay? (Q. 70)	Considering everything, how satisfied are you with your organization? (Q. 71)	I recommend my organization as a good place to work. (Q. 40)

GSI results 2018

Global Satisfaction Results and Trends

	2014	2015	2016	2017	2018
Overall Global Satisfaction (An average of the responses for the 4 items below)	59	60	61	64	64
Job Satisfaction	64	65	66	68	68
Pay Satisfaction	56	57	58	61	63
Organizational Satisfaction	55	56	57	60	60
Recommend Organization	62	63	64	66	66

GSI was up again in 2019 and 2020

	A	B	C	D	E	F	G
1	Appendix G: OPM FEVS Indices - Overall Global Satisfaction Index Comparisons						
2	Size Category	Agency	2016	2017	2018	2019	2020
3	Governmentwide	Governmentwide	61	64	64	65	69
4	Very Large	Department of Agriculture	64	67	61	60	64
5	Very Large	Department of Defense, Overall	62	65	65	66	70