Stats 404 final project

PRESENTOR: Huilin Tang
DATE: 3/6/2019

Business Questions

- Competitive Companies are eager to woo and vie for elite college graduates’ attention; however, companies have not figured out a way to improve recruitment effectiveness.
- What are the most important components to boost recruitment effectiveness?
- What is more important to recruitment effectiveness: work life balance or company culture?

Overview of data set

Over 67k employee reviews for Google, Amazon, Facebook, Apple, and Microsoft

We have a total of **16** variables, **7** of which are NUMERICAL and **9** that are STRING

OUTCOME VARIABLE:

OVERALL RATING

NUMERICAL 1- 5

Details

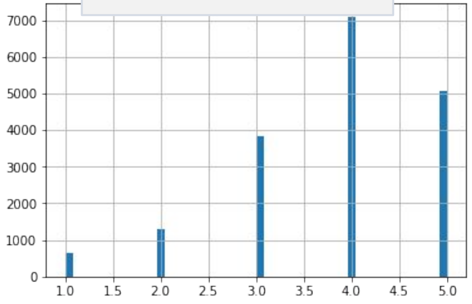
VARIABLE	DESCRIPTION	TYPE
Index	index	numerical
company	Company name	string
Date Posted	in the following format MM DD, YYYY	string
Job-Title	This string will also include whether the reviewer is a 'Current' or 'Former' Employee at the time of the review	string
Summary	Short summary of employee review	string
Pros	pros	string
Cons	cons	string

Details (con't)

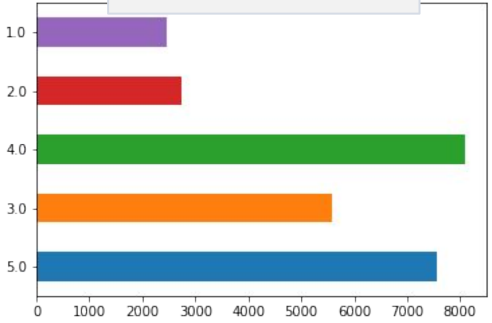
VARIABLE	DESCRIPTION	TYPE
Overall Rating	**1-5**	**numerical**
Work/Life Balance Rating	***1-5***	***numerical***
Culture and Values Rating	***1-5***	***numerical***
Career Opportunities Rating	1-5	numerical
Comp & Benefits Rating	1-5	numerical
Senior Management Rating	1-5	numerical
Helpful Review Count	A count of how many people found the review to be helpful	numerical

EXPLORATORY DATA ANALYSIS - OVERALL RATING

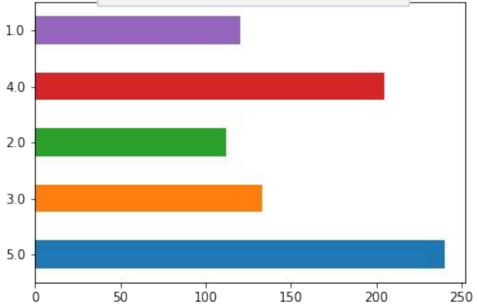
MICROSOFT



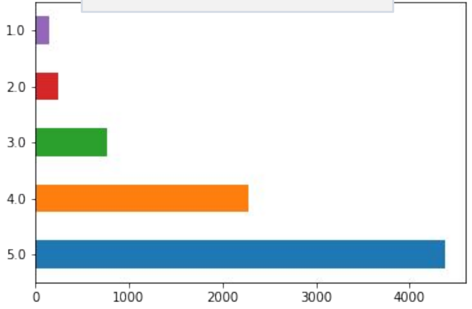
AMAZON



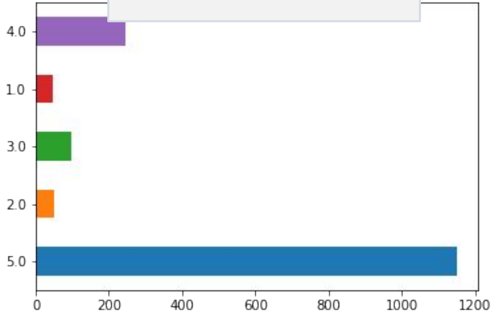
NETFLIX



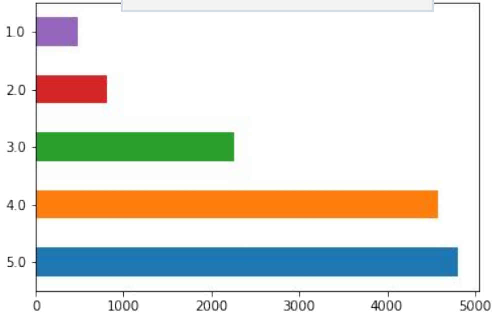
GOOGLE



FACEBOOK

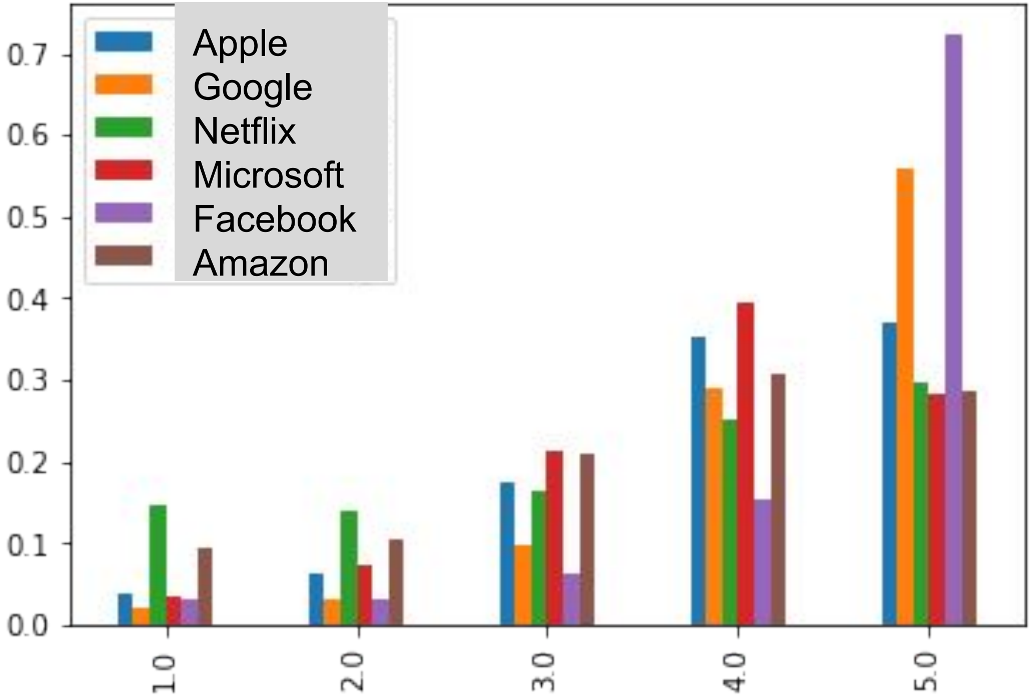


APPLE

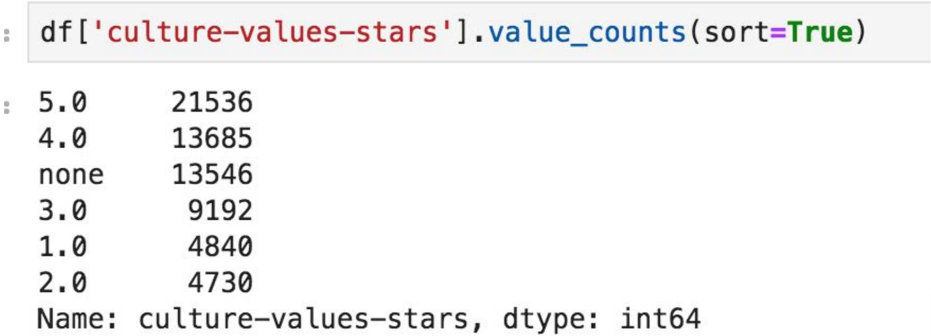


EXPLORATORY DATA ANALYSIS - OVERALL RATING

Histogram Of
All companies'
"Overall-rating"
as comparison



EDA (CON'T) - PREDICTORS

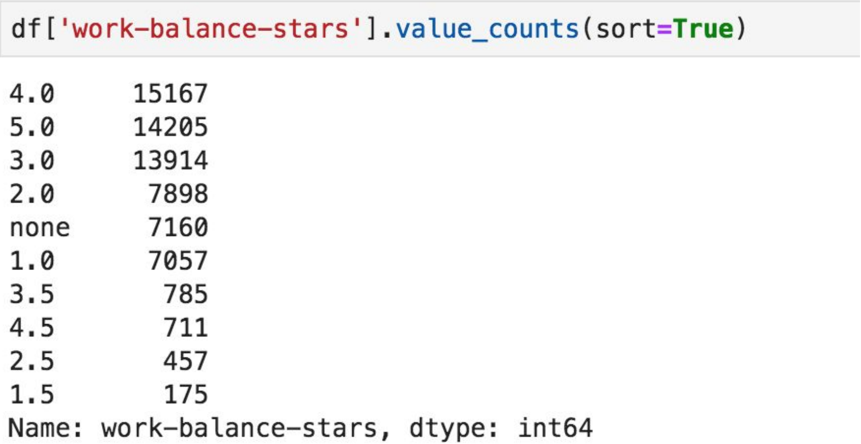


```
df['culture-values-stars'].value_counts(sort=True)
```

```
5.0      21536
4.0      13685
none     13546
3.0       9192
1.0       4840
2.0       4730
Name: culture-values-stars, dtype: int64
```

Counts of “culture_values_stars”

Counts of “work-balance-stars”

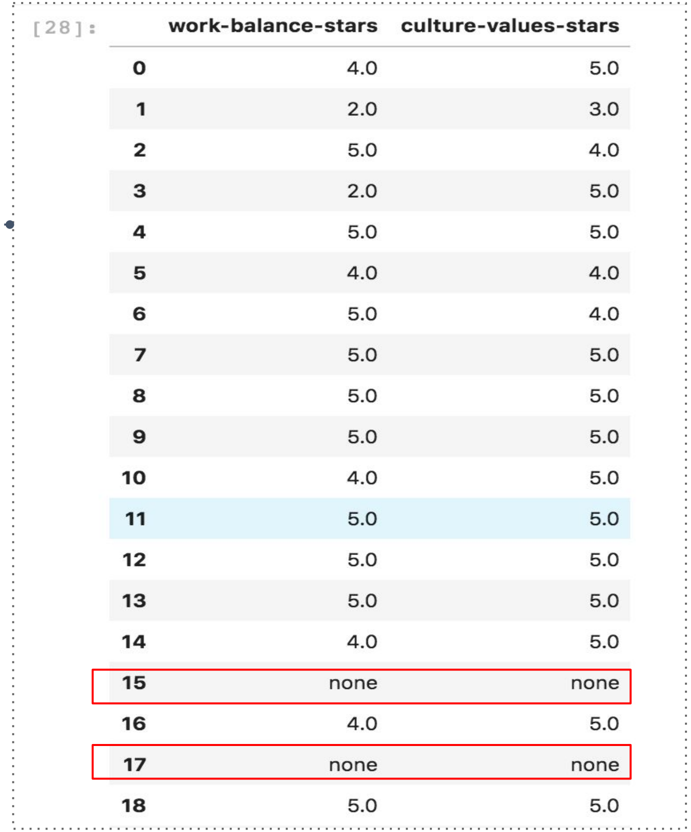


```
df['work-balance-stars'].value_counts(sort=True)
```

```
4.0      15167
5.0      14205
3.0      13914
2.0       7898
none      7160
1.0       7057
3.5        785
4.5        711
2.5        457
1.5        175
Name: work-balance-stars, dtype: int64
```

How I handle missing data

Snapshot Of missing data

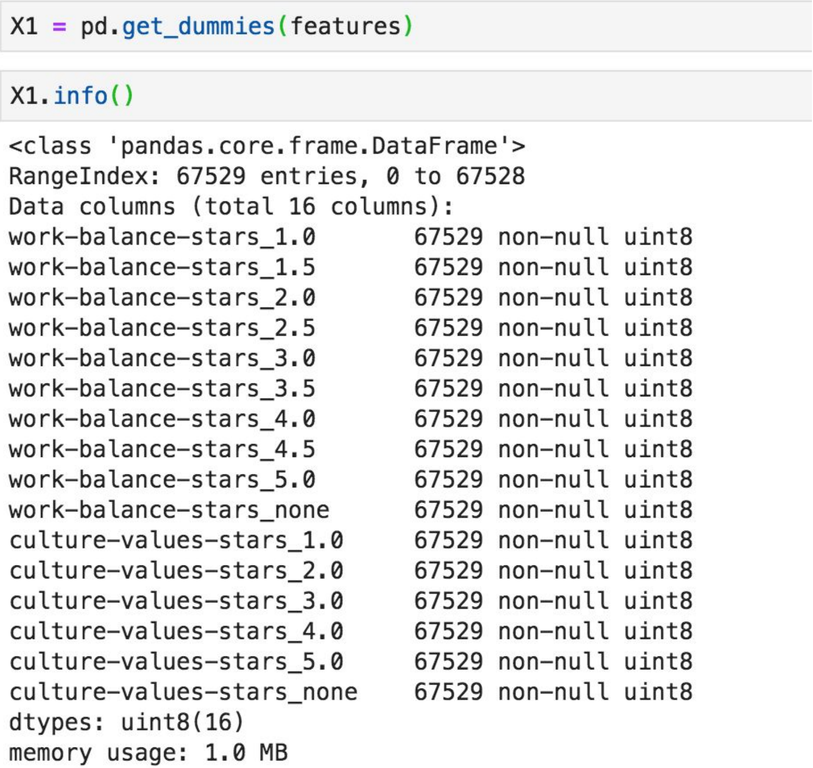


[28]:

	work-balance-stars	culture-values-stars
0	4.0	5.0
1	2.0	3.0
2	5.0	4.0
3	2.0	5.0
4	5.0	5.0
5	4.0	4.0
6	5.0	4.0
7	5.0	5.0
8	5.0	5.0
9	5.0	5.0
10	4.0	5.0
11	5.0	5.0
12	5.0	5.0
13	5.0	5.0
14	4.0	5.0
15	none	none
16	4.0	5.0
17	none	none
18	5.0	5.0

HOW TO HANDLE MISSING DATA (CON'T)

Introducing dummy variables



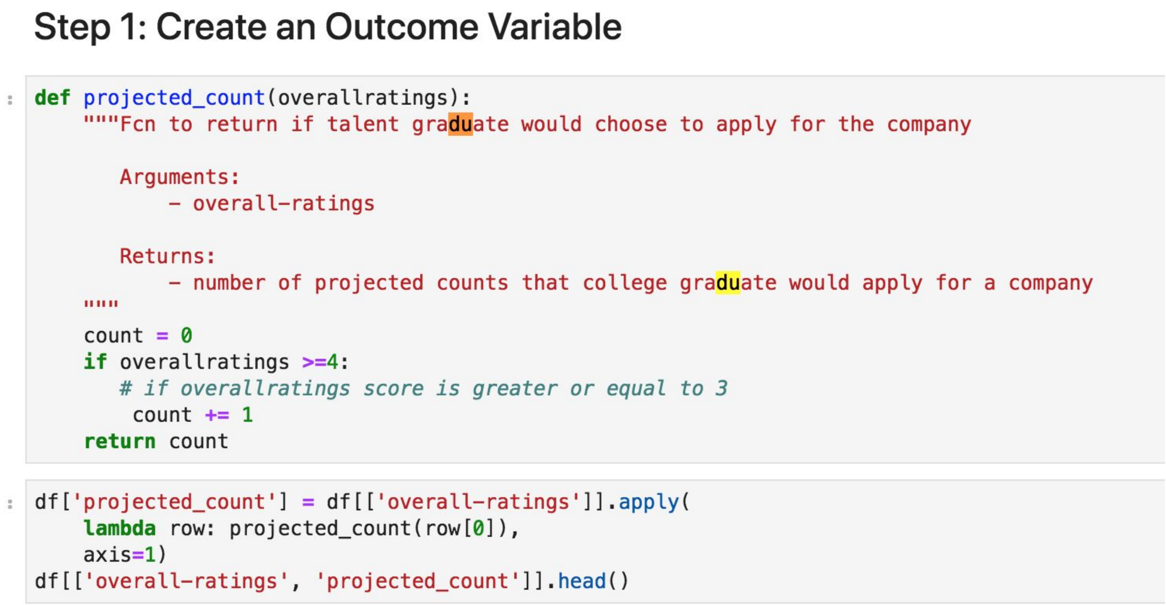
```
X1 = pd.get_dummies(features)
```

```
X1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 67529 entries, 0 to 67528
Data columns (total 16 columns):
work-balance-stars_1.0       67529 non-null uint8
work-balance-stars_1.5       67529 non-null uint8
work-balance-stars_2.0       67529 non-null uint8
work-balance-stars_2.5       67529 non-null uint8
work-balance-stars_3.0       67529 non-null uint8
work-balance-stars_3.5       67529 non-null uint8
work-balance-stars_4.0       67529 non-null uint8
work-balance-stars_4.5       67529 non-null uint8
work-balance-stars_5.0       67529 non-null uint8
work-balance-stars_none      67529 non-null uint8
culture-values-stars_1.0     67529 non-null uint8
culture-values-stars_2.0     67529 non-null uint8
culture-values-stars_3.0     67529 non-null uint8
culture-values-stars_4.0     67529 non-null uint8
culture-values-stars_5.0     67529 non-null uint8
culture-values-stars_none    67529 non-null uint8
dtypes: uint8(16)
memory usage: 1.0 MB
```

OVERVIEW OF MODELING

BASELINE: LOGISTIC REGRESSION



Step 1: Create an Outcome Variable

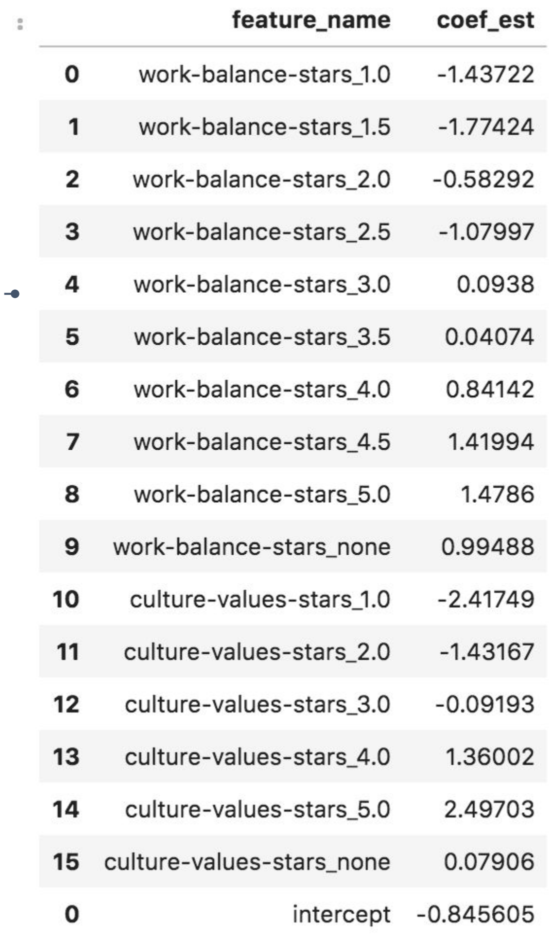
```
def projected_count(overallratings):
    """Fcn to return if talent graduate would choose to apply for the company

       Arguments:
           - overall-ratings

       Returns:
           - number of projected counts that college graduate would apply for a company
    """
    count = 0
    if overallratings >=4:
       # if overallratings score is greater or equal to 3
        count += 1
    return count
```

```
df['projected_count'] = df[['overall-ratings']].apply(
    lambda row: projected_count(row[0]),
    axis=1)
df[['overall-ratings', 'projected_count']].head()
```

OVERVIEW OF MODELING

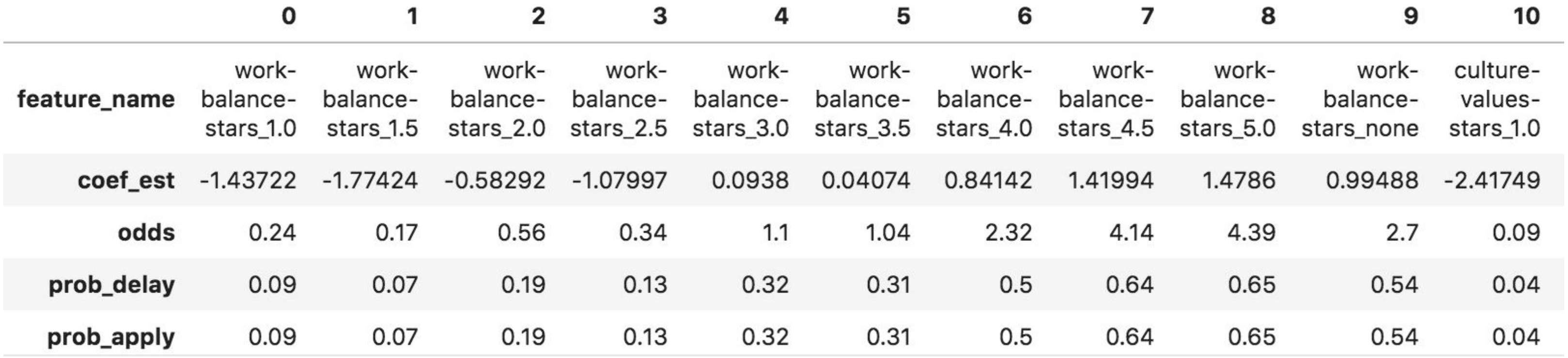
Result

When a binary outcome variable is modeled using logistic regression, it is assumed that the logit transformation of the outcome variable has a linear relationship with the predictor variables. This makes the interpretation of the regression coefficients somewhat tricky.



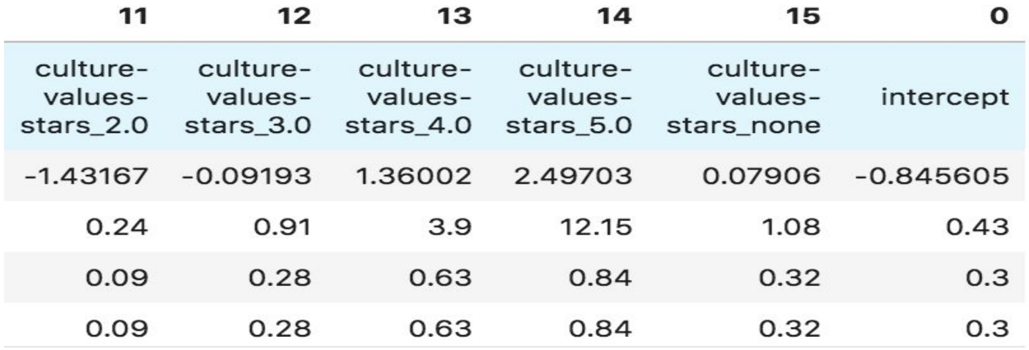
	feature_name	coef_est
0	work-balance-stars_1.0	-1.43722
1	work-balance-stars_1.5	-1.77424
2	work-balance-stars_2.0	-0.58292
3	work-balance-stars_2.5	-1.07997
4	work-balance-stars_3.0	0.0938
5	work-balance-stars_3.5	0.04074
6	work-balance-stars_4.0	0.84142
7	work-balance-stars_4.5	1.41994
8	work-balance-stars_5.0	1.4786
9	work-balance-stars_none	0.99488
10	culture-values-stars_1.0	-2.41749
11	culture-values-stars_2.0	-1.43167
12	culture-values-stars_3.0	-0.09193
13	culture-values-stars_4.0	1.36002
14	culture-values-stars_5.0	2.49703
15	culture-values-stars_none	0.07906
0	intercept	-0.845605

INTERPRETATION



	0	1	2	3	4	5	6	7	8	9	10
feature_name	work-balance-stars_1.0	work-balance-stars_1.5	work-balance-stars_2.0	work-balance-stars_2.5	work-balance-stars_3.0	work-balance-stars_3.5	work-balance-stars_4.0	work-balance-stars_4.5	work-balance-stars_5.0	work-balance-stars_none	culture-values-stars_1.0
coef_est	-1.43722	-1.77424	-0.58292	-1.07997	0.0938	0.04074	0.84142	1.41994	1.4786	0.99488	-2.41749
odds	0.24	0.17	0.56	0.34	1.1	1.04	2.32	4.14	4.39	2.7	0.09
prob_delay	0.09	0.07	0.19	0.13	0.32	0.31	0.5	0.64	0.65	0.54	0.04
prob_apply	0.09	0.07	0.19	0.13	0.32	0.31	0.5	0.64	0.65	0.54	0.04

Result after logit transformation



11	12	13	14	15	0
culture-values-stars_2.0	culture-values-stars_3.0	culture-values-stars_4.0	culture-values-stars_5.0	culture-values-stars_none	intercept
-1.43167	-0.09193	1.36002	2.49703	0.07906	-0.845605
0.24	0.91	3.9	12.15	1.08	0.43
0.09	0.28	0.63	0.84	0.32	0.3
0.09	0.28	0.63	0.84	0.32	0.3

KEY FINDINGS & RECOMMENDATION

- Employees are prone to value company culture more than work-balance feature, when they apply for a company.
-

How WOULD BUSINESS USE THE MODEL?

Regression models can be used to rank the relative importance of quantitative factors impacting any process. This information can be used to target those variables that are key, focusing decision-making.

In other words, regression models could help business to identify the prominent features that impact on the subject they care the most about.

POTENTIAL STEPS & RECOMMENDATIONS

- Comparison between companies
- Adding more features
- Text analysis

QUESTIONS

Thank you!