A LOOK AT ATTRITION

BY ROBERT CARSTENS

Introduction

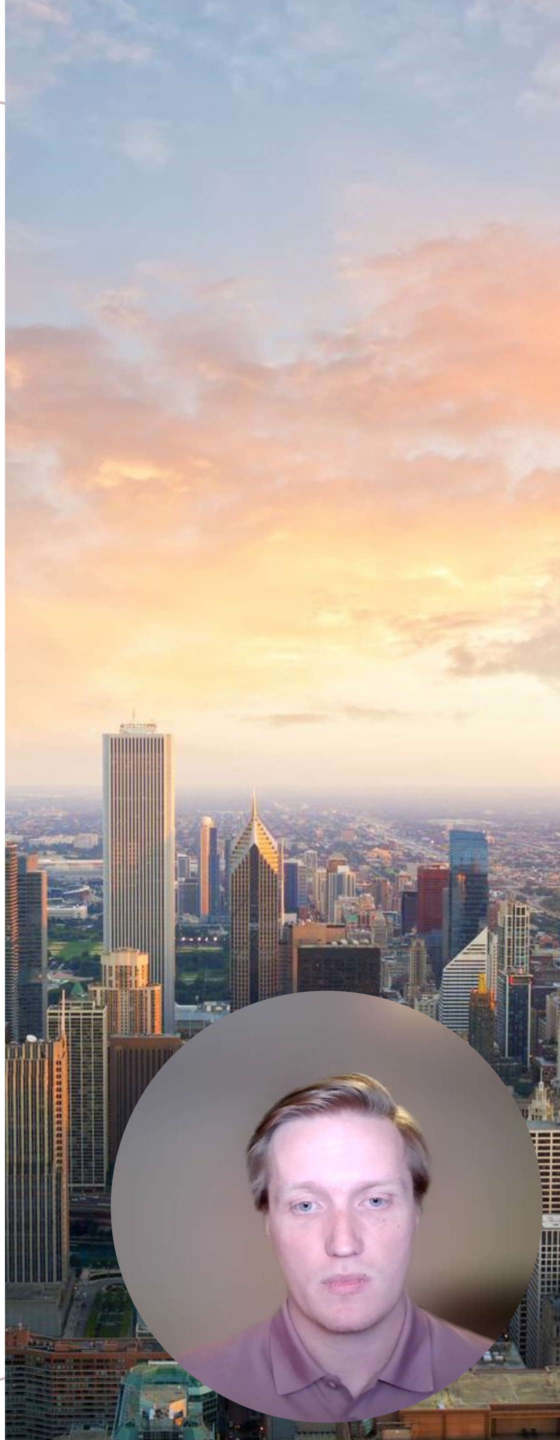
Exploratory Data Analysis

Most Important Factors

Feature Selection + Feature Engineering

Final Model

Conclusion



WHO AM I

- Robert Carstens
- Data Scientist at DDSAnalytics
- Expert on all things Machine Learning
- Over 23 years of experience

OUR OBJECTIVE

- We aim to understand attrition and the factors that affect it, in order to help Frito Lay predict employee turnover.

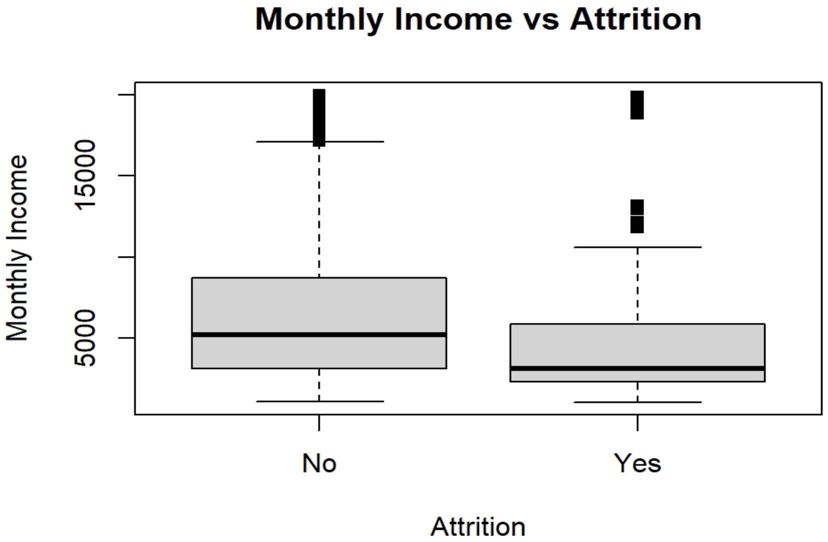
- To do so we will be analyzing some employee data given to us by DDSAnalytics, performing an exploratory data analysis with this data, and then building a predictive model using Naïve Bayes to predict attrition.

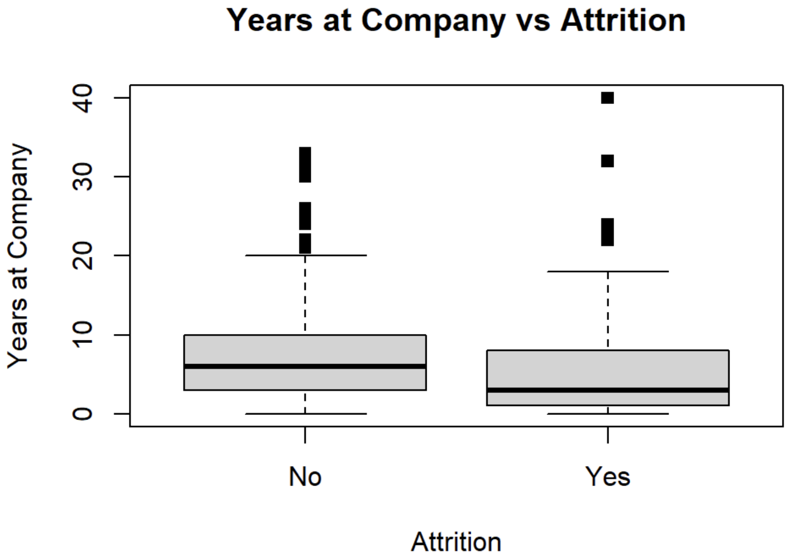
EXPLORATORY DATA ANALYSIS

WHAT DOES THE DATA TELL US

EXPLORATORY DATA ANALYSIS

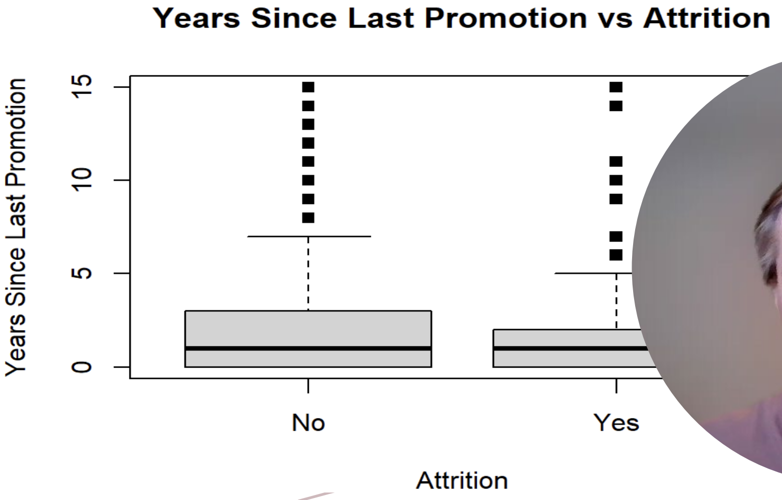
- I explored several factors and their impact on attrition through my exploratory data analysis.

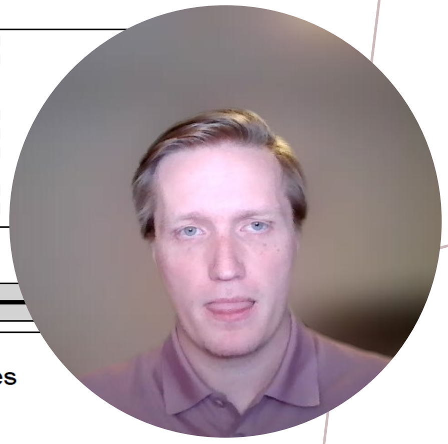




A look at stock options and its relationship with attrition

	0	1	2	3
No	281	328	78	43
Yes	98	27	3	12





FINDING THE MOST IMPACTFUL FACTORS

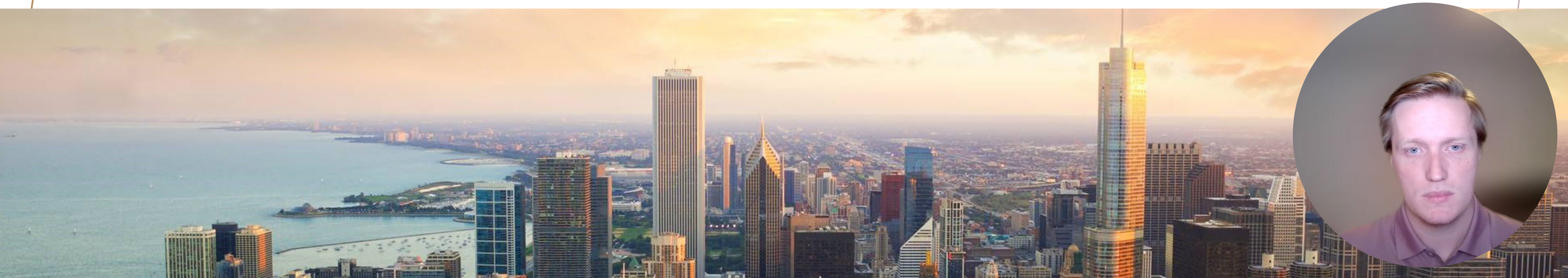
IDENTIFYING IMPORTANT FACTORS

- To find the significant factors, I ran a regression using all the factors in the dataset. To do this successfully, I had to remove all columns that contained only 1 value.
- From here, I evaluated the p-value associated with each variable and select those with the lowest p-values
- From here I was able to find that Overtime, Number of Companies Worked, and Job Involvement were the 3 most statistically significant factors, as they had the lowest p-values.

```
Call:
glm(formula = Attrition ~ ., family = binomial, data = Attrition)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -9.107e+00  7.403e+02  -0.012 0.990185
ID                               2.073e-04  5.174e-04   0.401 0.688726
Age                             -2.983e-02  1.944e-02  -1.534 0.124997
BusinessTravelTravel_Frequently  1.629e+00  5.233e-01   3.114 0.001846 **
BusinessTravelTravel_Rarely      7.213e-01  4.623e-01   1.560 0.118726
DailyRate                       -3.578e-04  3.253e-04  -1.100 0.271415
DepartmentResearch & Development 1.363e+01  7.403e+02   0.018 0.985307
DepartmentSales                  1.391e+01  7.403e+02   0.019 0.985010
DistanceFromHome                 5.240e-02  1.545e-02   3.392 0.000693 ***
Education                       -1.112e-02  1.238e-01  -0.090 0.928423
EducationFieldLife Sciences     -1.448e+00  1.224e+00  -1.184 0.236550
EducationFieldMarketing         -1.588e+00  1.295e+00  -1.227 0.219995
EducationFieldMedical           -1.569e+00  1.221e+00  -1.285 0.198845
EducationFieldOther             -1.273e+00  1.290e+00  -0.987 0.323868
EducationFieldTechnical Degree  -7.791e-01  1.256e+00  -0.620 0.535222
EmployeeNumber                  -2.622e-04  2.193e-04  -1.195 0.231964
EnvironmentSatisfaction         -3.069e-01  1.205e-01  -2.548 0.010850 *
```

BUILDING OUR MODEL



FEATURE SELECTION

- I wanted to include those factors which proved to be statistically significant based on my earlier analysis. While still keeping our model relatively simple to avoid overfitting.
- In addition to this, I wanted to find a way to improve one of the factors found to be in the top 3 most significant. Number of Jobs

FEATURE ENGINEERING

Job Hop Score is essentially a way to look at the length of a subjects employment, in comparison to the average length of their previous employments.

$$\mathit{Number\ of\ Years\ at\ Company} * \frac{\mathit{Total\ Working\ Years} + 1}{\mathit{Number\ of\ Companies\ Worked} + 1}$$

The idea here is that if somebody leaves their job every 1-2 years but they've only been at the company 1-2 years, they might be more likely to leave soon, versus somebody who tends to stay for 5-10 years.

When evaluating models with this feature vs strictly number of jobs. The models with Job Hob Score proved to be more accurate.

OUR FINAL MODEL

$$Attrition = Job\ Hop\ Score + Age + OverTime + Environment\ Satisfaction\ Score + Gender + Distance\ From\ Home + Job\ Involvement$$

EVALUATING OUR MODEL

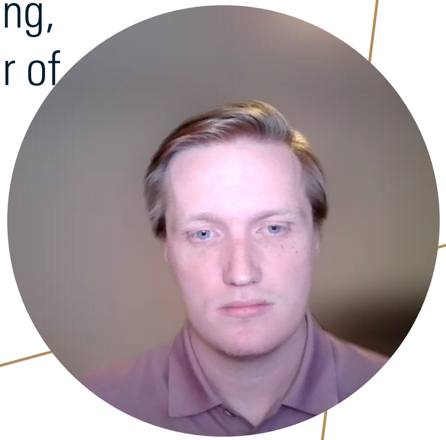
EVALUATING OUR MODEL

Training vs Testing Data

- We the remaining data into Training and Testing data sets using a 70-30 split.
- We felt that giving the training data a higher amount of records to train with, as we were going to lose some in that data set post – balancing it.
- Sample Sizes
  - Training pre-balancing: 609 records
  - Training post-balancing: 189 records
  - Testing 261

Balancing our Data Set

- The existing data set was very unbalanced, meaning that most of the records contained had "No" for their attrition value.
- This can make training models hard, as they will want to then classify everything as "No" to optimize their accuracy
- We balanced our data set by down-sampling, which means that we dropped a good number of those "No's" from the dataset



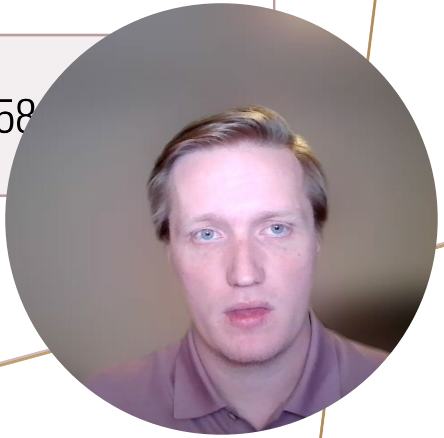
HOW WELL DOES OUR MODEL PERFORM?

Actual

Predicted

	No	Yes
No	154	11
Yes	65	31

Metric	Score	Confidence Interval
Accuracy	70.8%	[65.0% ,76.3%]
Sensitivity	70.3%	[63.8%, 76.2%]
Specificity	73.8%	[58



IN CONCLUSION

1. We have taken a look at and analyzed our dataset in order to understand it.
2. We have identified the 3 most statistically significant factors affecting employee attrition.
    a. Number of Companies Worked
    b. Over time
    c. Job Involvement
3. We have built a model that predicts employee attrition with
    a. 70.8% Accuracy
    b. 70.3% Sensitivity
    c. 73.8% Specificity

THANK YOU

Robert Carstens

469-964-1603

rcarstens@smu.edu