

# Predicting Early Attritions in the Technology industry using Machine Learning

Clinton Ngan

*"In general terms, attrition is defined as a process of reducing the strength of a particular thing to reduce the effectiveness."*

*- Chourey (2019)*

Resham Sundrani (2012)

\$2.5x

Call et. Al. (2015)

-8.9%

Singh (2018)

~49M

Bean (2019)

#### Sectors with the Highest Turnover Rates

1.	 Technology (Software)	13.2%
2.	 Retail & Consumer Products	13.0%
3.	 Media & Entertainment	11.4%
4.	 Professional Services	11.4%
5.	 Government/Edu/Non-Profit	11.2%

Source: LinkedIn.com

### Hypothesis

**Socializing with teammates outside of work settings and number of promotions would be important factors on an employee's decision to depart a technology company early. Thus, resulting in an "early attrition".**



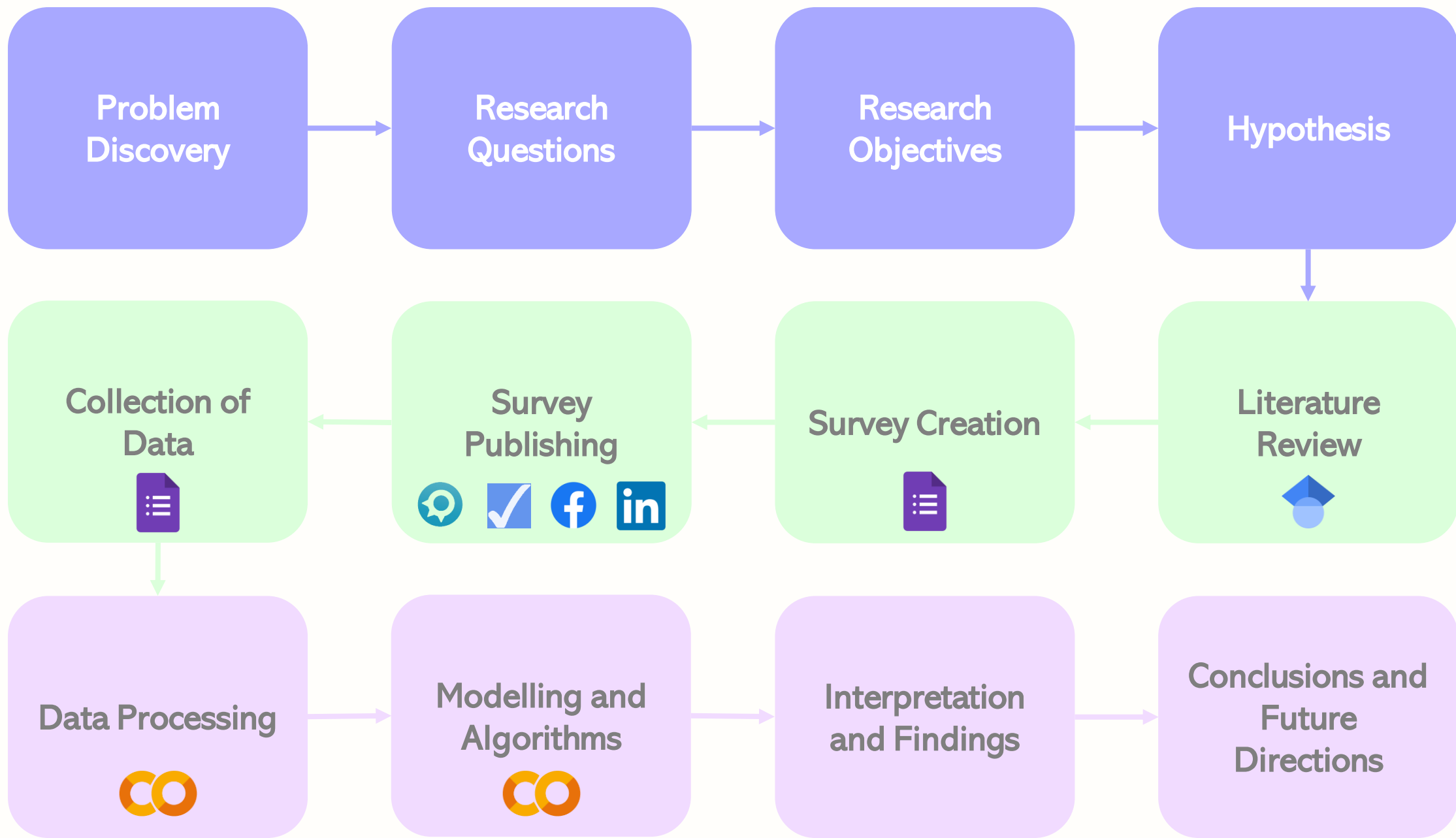
Alrazehi, H. and Noor Aina Amirah (2020)  
 Bean, D. (2019)  
 Babu, G. M., Student, M. C. A. and Assistant, S. (2020)  
 Call, M. L. et al. (2015)  
 Chourey, A. (2019)  
 Chukwu, B. . (2019)  
 Dubey, A. K., Maheshwari, I. and Mishra, A. (2018)  
 Ellis, J. (2020)  
 Foley, A. and Foley, A. E. (2019)  
 Gabrani, G. and Kwatra, A. (2018)  
 Gao, X., Wen, J. and Zhang, C. (2019)  
 Halim, Z. et al. (2020)  
 Keshri, R. (2020)  
 Khara, S. N. and Divya (2019)  
 Kumar, A. and Mathimaran, B. (2017)  
 Lema, G. et al. (2015)  
 Mathur, R. and Jain, A. (2015)  
 Mckibbin, W. J. and Member, C. (2020)  
 Mendonsa, K., Stolberg, M. and Crum, S. (2020)  
 Punnoose, R. and Ajit, P. (2016)  
 Resham Sundrani (2012)  
 Ribes, E., Touahri, K. and Perthame, B. (2017)  
 Seger, C. (2018)  
 Shore et., A. (2014)  
 Singh, A. (2018)  
 Srinivas, K. et al. (2020)  
 Tanmay Prakash Salunkhe (2017)  
 Technische Universtität München, L.-M.-U. M. (2018)  
 Younge, K. A., & Marx, M. (2016)  
 Zhao Y., Hryniewicki M.K., Cheng F., Fu B., Zhu X. (2019)



**30**  
 Literatures Reviewed

**2012 - 20**  
 Published

**~ 3**  
 Tech Industry specific



## Survey Creation

- 24 Questions total
- Previous and currently in Tech industry
- Consent form attached
- Participants can skip anytime



## Survey Publishing

- Joined >10 FB groups to promote
- Published on Survey Tandem
- Published on Survey Circle
- Promoted on LinkedIn



## Data Collection/Processing

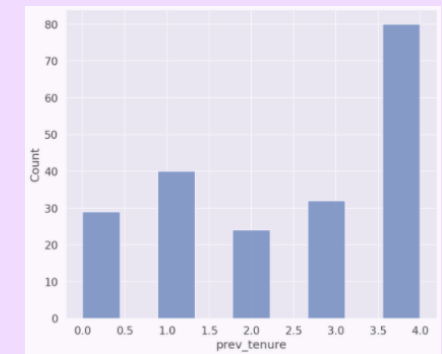
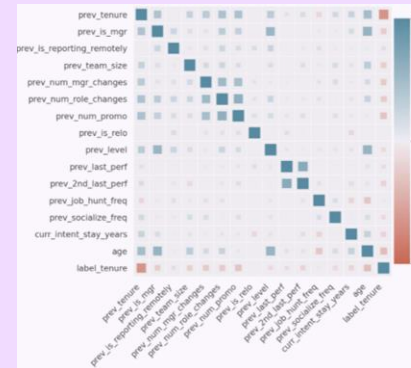
- 222 total responses
- 205 total unique responses
- Performed EDA to catch outliers
- Explored missing values for imputation
- Renaming columns for housekeeping

```
#rename columns
for col in data.columns:
    print(col)

data.rename(columns={"Username":"id"
                    "How many years
                    "From the list
                    "Were you a man
                    "Were you repor
```

```
# Count NaNs% in each column
na_analysis = round(data.isna().sum()
                    print(na_analysis)

prev_tenure           0.00
prev_leave_reason     0.00
prev_is_mgr           0.00
prev_is_reporting_remotely 0.49
prev_team_size       0.00
```





## Algorithms

Extra Trees

Decision Tree

Gradient Boost

Logistic  
Regression

Random Forest

## Imputations

Mean

Median

Mode

Constant

## Processes

1

Utilizing all algorithms and imputation strategies without balancing the data set's class distribution

2

Utilizing all algorithms and imputation strategies by manually balancing the data set's class distribution

3

Utilizing all algorithms and imputation strategies by automating balancing process using under sampling

4

Utilizing all algorithms and imputation strategies by automating balancing process using SMOTE



Algorithms	Processes	Mean	Median	Mode	Constant
Decision Tree	Process 1 (imbalance)	0.46	0.52	0.49	0.48
	Process 2 (manual balance)	0.79	0.79	0.74	0.79
	Process 3 (undersampling)	0.66	0.62	0.67	0.67
	Process 4 (SMOTE)	0.91	0.92	0.92	0.92
Random Forest	Process 1 (imbalance)	0.47	0.5	0.45	0.4
	Process 2 (manual balance)	0.9	0.89	0.89	0.9
	Process 3 (undersampling)	0.79	0.79	0.78	0.8
	Process 4 (SMOTE)	0.95	0.94	0.94	0.94
Extra Trees	Process 1 (imbalance)	0.41	0.45	0.37	0.34
	Process 2 (manual balance)	0.91	0.92	0.9	0.91
	Process 3 (undersampling)	0.79	0.82	0.82	0.81
	Process 4 (SMOTE)	0.97	0.96	0.96	0.96
Gradient Boost	Process 1 (imbalance)	0.4	0.4	0.39	0.44
	Process 2 (manual balance)	0.88	0.9	0.9	0.87
	Process 3 (undersampling)	0.7	0.73	0.74	0.74
	Process 4 (SMOTE)	0.92	0.92	0.92	0.92
Logistic Regression	Process 1 (imbalance)	0.38	0.37	0.24	0.17
	Process 2 (manual balance)	0.84	0.86	0.85	0.86
	Process 3 (undersampling)	0.81	0.8	0.82	0.81
	Process 4 (SMOTE)	0.87	0.86	0.86	0.86



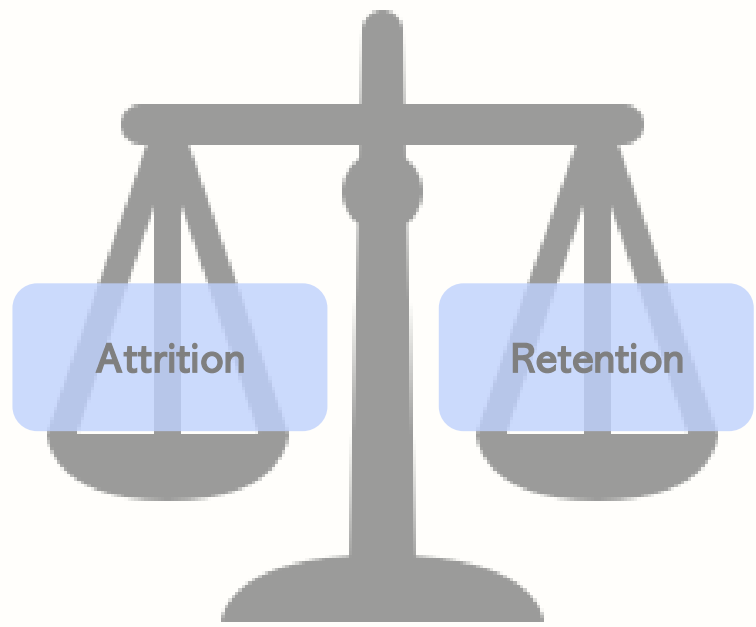
Algorithms	Mendonsa 2020	Gabrani 2018	Foley 2019	Chourey 2019	Gao 2019	Ngan 2021
Random Forest	0.89	0.98	0.48	0.33	0.70	0.94
Decision Tree	0.83	0.95	-	0.31	-	0.92
Logistic Regression	0.80	0.76	-	-	0.26	0.86
Gradient Boost	-	-	-	0.45	-	0.92
Extra Trees	-	-	-	-	-	0.96

- Gabrani yielded best performance for Random Forest and Decision Tree
- This study yielded 2<sup>nd</sup> best performance for Random Forest and Decision Tree
- This study yielded best performance for Logistic Regression and Gradient Boost
- No one else used Extra Trees but it was this study's best performing algorithm
- Overall, this study's performance holds up well against other studies'





Features	Correlation
prev_num_promo	-0.54
age	-0.49
prev_num_role_changes	-0.49
prev_num_mgr_changes	-0.44
prev_team_size	-0.44
prev_is_mgr	-0.35
prev_is_reporting_remotely	0.17
prev_level	-0.16
prev_job_hunt_freq	0.05
prev_last_perf	-0.03
prev_is_relo	0.03
prev_socialize_freq	0.02



### Hypothesis

**Socializing with teammates outside of work settings and number of promotions** would be important factors on an employee's decision to depart a technology company early. Thus, resulting in an "early attrition".

### Future Direction

By understanding both attrition and retention, HR leaders in technology companies can ensure they are hiring and retaining effectively but also creating appropriate and suitable pipelines for candidate recruiting to combat early attritions.

# Q&A

Clinton Ngan