Attrition Analysis and Prediction

Project Team:
Zhipeng Luo(brianluo@umich.edu) & Chihshen Hsu (cshsu@umich.edu) & Qi Zhang(qizhan@umich.edu)

Motivation:

Talent consistency has long been an important factor plaguing companies, especially in the rapidly evolving Internet era where markets become more volatile with increased uncertainty risks in the macro environment. Attrition rate control and talent retention are always important parts of a company's development strategy and one of the biggest challenges for the Human Resources (HR) department, especially under the "new normal" condition with COVID-19 impacting world widely. From the newest research from McKinsey published on September 8, 2021, the "Great Attrition" is happening and will probably continue (Smet A.D.).

In this project, we target to explore that employees voluntarily choose to leave which could be caused by poor work environment, unsatisfied pay, lack of advancement opportunities, long overtime, or even just the employee wanting to make some changes.

In addition to exploring the correlations, we would like to design a model with the function to predict whether an employee would leave based on diverse independent variables and identify the key features.

Key Questions:

- **Does over-time work trigger attrition?**
- **How about the salary of IBM employees compared to the market ?**
- **How about the effect of position promotion on attrition?**
- **Which model is more appropriate for prediction and how to evaluate?**

Data Sources: IBM Attrition Data

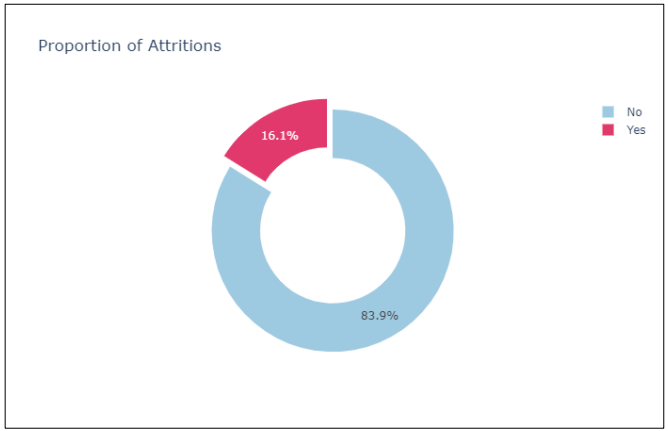
The Attrition dataset has 1470 observations with **35 variables**. Out of the 35 variables, the target variable Attrition has binary outcomes, Yes and No.

In this dataset, all the data provided has **no missing values**. The dataset contains several numerical and categorical columns providing various information on employee's personal and employment details. We will use this dataset to explore the correlations of different variables with Attrition and predict when employees are going to quit by understanding the main drivers of employee churn.

As stated on the IBM website: "*This is a fictional data set created by IBM data scientists. Its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition.*" The main weakness of the dataset is that **it has no temporal data, but rather the static information**. We can't observe the tend of the attrition along the time triggered by the changes of other independent variables.

In addition, our target data, Attrition, is suffering **skewness** and can be regarded as **Moderate-level imbalance** (Google Developers), referring to the minority proportion as **16.1%** shown in the Donut chart. This imbalanced property is vital to be considered for hyper-parameter fine-tuning, model and evaluation metrics selection for any machine learning algorithm application.

Name	**attrition.csv**
Size	**222 K**
Format	**csv**
Records	**1,740**
Time Period	**Q4, 2017**
Source	https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset



Data Sources: Stack Overflow Survey Data

The data is a questionnaire survey conducted by stack overflow for software engineers worldwide in 2018. According to stack overflow's backend data, each questionnaire takes an average of 30 minutes to answer, and the total number of records received is close to 100,000.

The questionnaire dataset has **129 variables**, includes not only some features existing in IBM attrition data, such as age, gender, salary and job satisfaction, but also other important variables deserves exploration, such as employment status, size of the company, job searching status, Salary Type, Country, etc.

Since the data was released in January 2018, the survey data is reasonable to be used to make cross analysis with the IBM attrition data, especially for R&D department. Of all the 98,855 responses, 67,441 completed the entire survey. So, some of the nulls need to be treated before analysis. As the questionnaire survey is world widely implemented, the value of some features may not be as tidy as the one in IBM data, which required data manipulation before comparison. For example, both datasets has the variable "**Salary"**. In the IBM attrition dataset, it is named as **Monthly Income** with specific and standardized meaning. However, it includes both monthly income, yearly income, weekly income and null value in survey data.

Name	**survey_results_public.csv**
Size	**186 MB**
Format	**csv**
Records	**98,855**
Time Period	**January 2018**
Source	https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey

Data Manipulation & Cleaning

The IBM dataset is clean and tidy, without any requirement for empty value treatment. Besides of Split-Apply-Combine strategy, one of the main manipulation for machine learning application is to convert the object-type columns by different methods based on different conditions. For the features with binary values, “Yes” and “No”, we can directly use **map** or **LabelEncoder** to transform the value to be numeric value as 1 (Yes) and 0 (No), such as our target data [Attrition] and the variable [OverTime]. If the value of the feature has ordinal property, we choose to convert them using **map** as its intuitive and readable characteristics, such as [BusinessTravel] with three levels of value, Non-Travel, Travel_Rarely , and Travel-Frequently, converted to 0, 1 and 2, respectively . The rest of the object-type features are converted by **get_dummies**, such as [Department], [Gender] and so on. The variables that only have the serial number or constant value, such as [Over18, EmployeeCount, StandardHours] , will be removed by **drop**.

The survey data from stack overflow was collected globally. Compared to IBM dataset, the survey data has larger volume with approximately 100,000 records. To make cross analysis between the two data set, our first manipulation step for the survey data is to filter the dataset by three variables, including [Country] as “United States”, [CompanySize] over 5000 employees and [Employment] as “Employed full-time”.

The two features of the survey data, [JobSatisfaction] and [Salary], are used to connect with IBM dataset to make cross analysis after manipulation. The seven levels of [JobSatisfaction] are **grouped** and **mapped** to be the same 4 levels as that in IBM dataset. To compare the salary condition, we take the **subset** of survey data by [SalaryType] as “Monthly”, ignoring “Yearly”, “Weekly” types and removing the missing value by **dropna** in row.

Split-Apply-Combine strategy are heavily used in the whole exploratory data analysis. The API **groupby** is commonly used to **aggregate** required statistics. By **merge** or **concat** function, the new dataframe is made to support our data visualization.

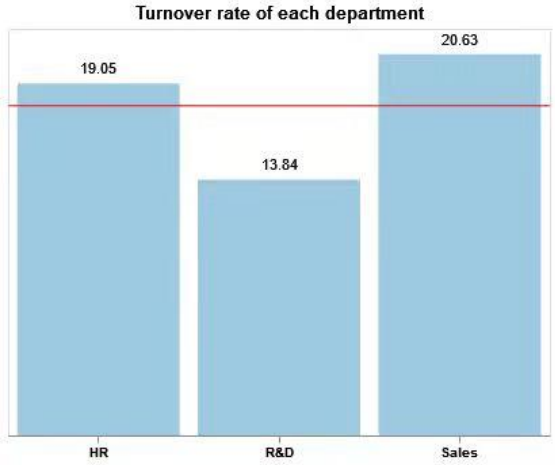
Analysis: OverTime Triggers Attrition ?

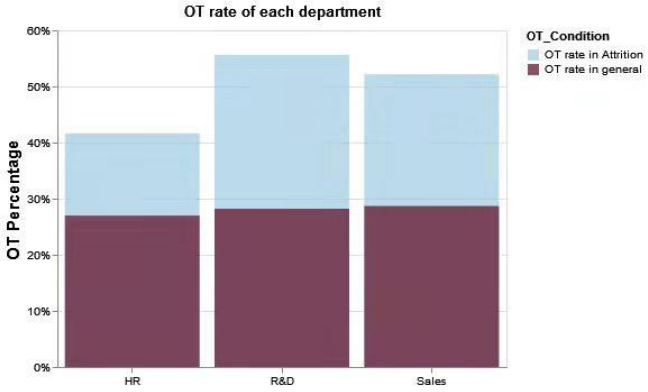
The debates and concerns about over-time work is unstoppable. One blog on CIRCADIAN discussed the 5 negative effects of high overtime levels, including increased turnover rates as one of its direct results. In August 2021, one of the hottest news is that China ruled the “996 work style”, a practice of work from 9 am to 9 pm per day and six days per work, is illegal.

In IBM data, **OverTime** is a binary variable, including two nominal value, Yes and No. The general turnover rate is 16.1%, which is higher than the overall turnover rate of Technology Industry (13.2%) based on the data of LinkedIn in 2017. After drilling down to see the turnover rate by department, we can find that R&D (Research & Development) has the lowest turnover rate, while, Sales suffers the most serious turnover rate, achieving over 20%.

To evaluate the over-time work conditions, we plot the layered non-stacked bar chart to illustrate the general OT (overtime) rate in purple and OT rate of Attrition group in light blue by department. Generally, the OT rate of the three departments almost equal to each other at around 28%. However, taking all the attritions from R&D department into consideration, around 58% of them experienced overtime, which is the highest proportion of all the three departments.

After exploring the attrition and overtime conditions by departments, we can’t simply conclude that overtime impacts the decisions of employees to leave or not.



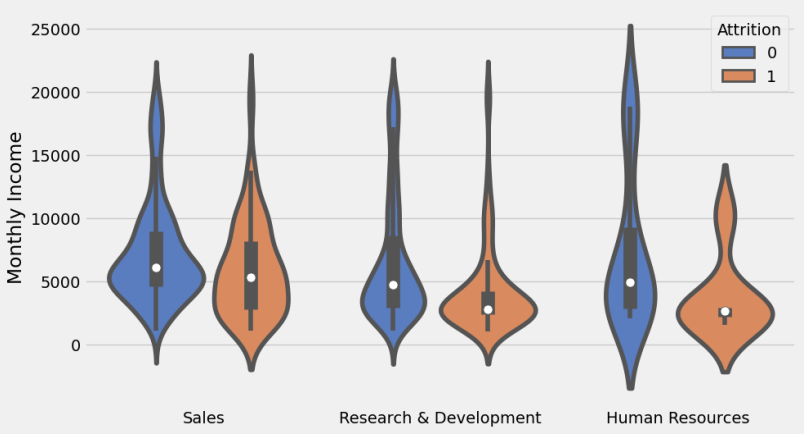


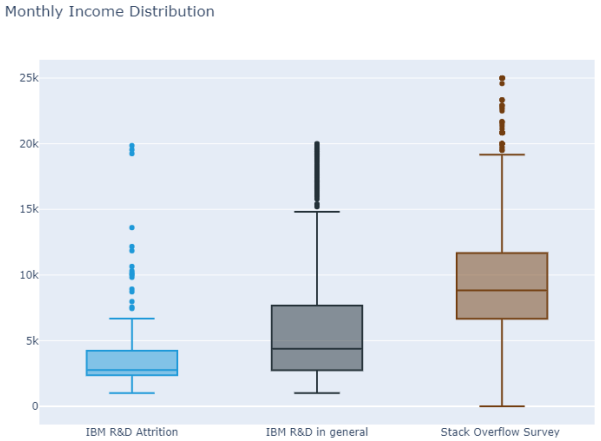
Analysis: How about Salary ?

Salary is always an important parameter to be considered for the design of talent retention strategy. Randstad illustrated *"if your employees' compensation levels are not competitive, don't expect them to stick around for long."*

By demonstrating the salary distribution of each department for the two groups of employees by attrition status through violin plot, we can find the median values of the salary of all the three departments who stay in the company are higher than those who choose to leave.

To make a strong comparison between the salary of IBM employees and the market condition, we use box plot to show the results. Within IBM R&D department, the monthly income of the attrition groups are lower than the department average. The brown box is generated by the survey data from Stack Overflow, filtered the country as the US and company size with over 5000 employees, to make apple-to-apple comparison. The gap between IBM attrition group from R&D department to the market condition is larger!





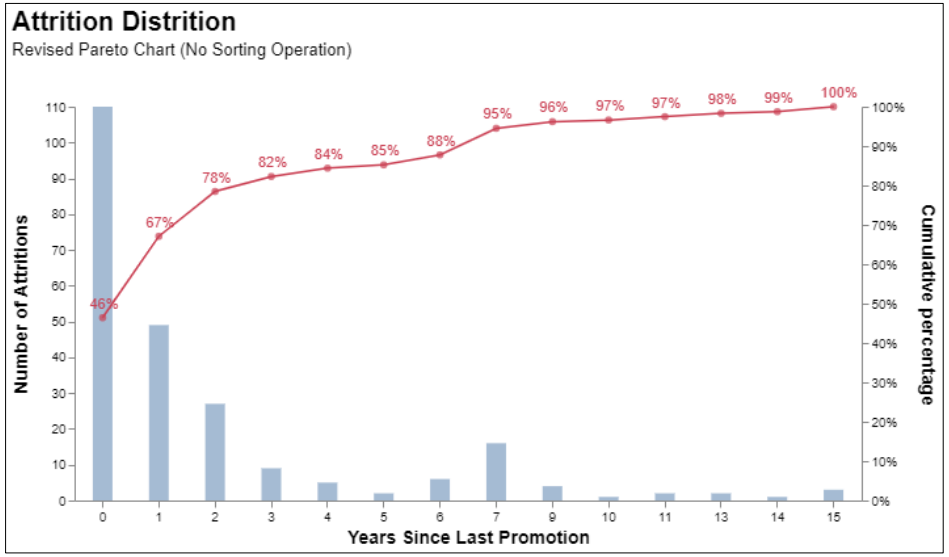
The exploration of salary can reasonably support our inference that people tend to leave and choose a job with better income packages.

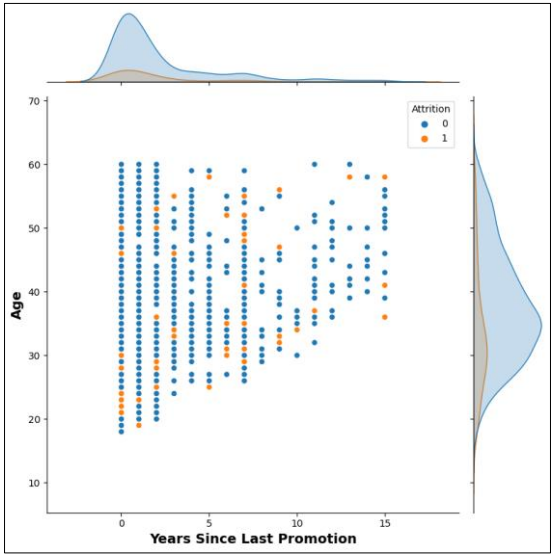
Analysis: Does promotion prevent attrition ?

"Employees want to feel recognized, valued and engaged", Joel Garfinkle believes that one way to insure you keep valuable employees is to promote them. However, the IBM data tells us another story. The revised Pareto Chart illustrates that 82% of the attritions achieved promotion within the past 3 years.

The joint plot between Age and Years-Since-Last-Promotion demonstrates that IBM has healthy age composition with a significant portion of youngsters, most of which has been promoted within the past 3 years.

Even though being promoted, some of the employees may still choose to leave because of other factors or more desired career or academic opportunities outside.

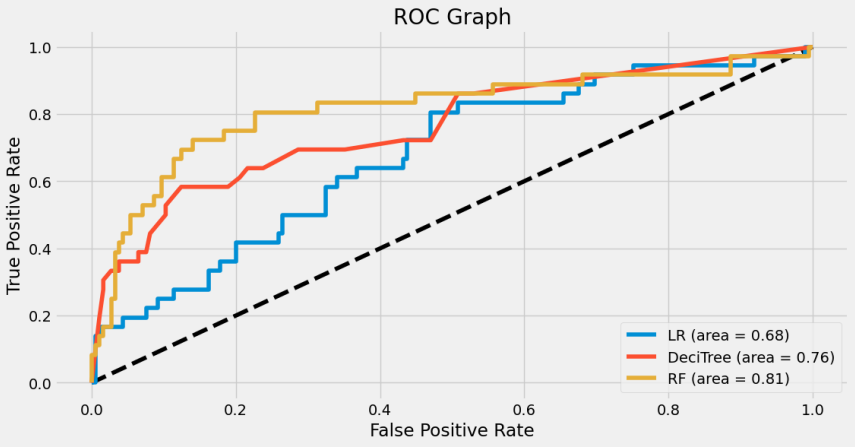


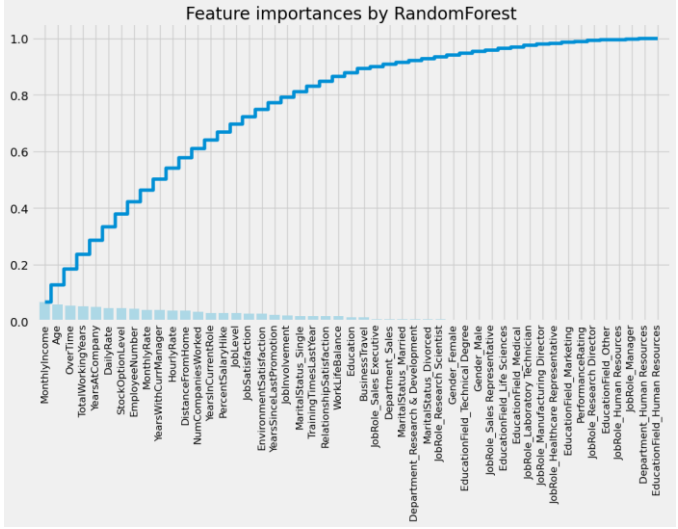


Can we predict the attrition of an employee?

Prior to applying any algorithms, train-test-split technique is utilized to divide the dataset into two subsets. One is for algorithm training; another one is for evaluation. Three common algorithms, **Logistic Regression**, **Decision Tree** and **Random Forest**, are trained to solve this binary classification problem. Because of the imbalance property of the IBM dataset, **Area under ROC Curve** (or AUC for short) is selected, instead of Classification Accuracy, as the metric to evaluate the model performance. Area of 1.0 represents the perfect prediction of the model, while an area of 0.5 indicates the model as good as random.

Based on the results, Random Forest performs the best with 81% probability to make right prediction. By plotting bar chart of feature importance with the cumulative line, the most important features can be revealed which supports the highest weight for the model to make classification, including Monthly Income, Age, Over Time, Total Working Years, etc. Comparing to the prediction accuracy from [Dr. Hamza Bendemra], the performance of Random Forest and Logistic Regression can be further improved after fine tuning of hyper-parameters, especially the latter one achieving 86%.





Conclusions & Next Steps

Conclusions:

As can be seen from the above illustration, **monthly income** and **overtime** can be regarded as the factors with top-level importance that relate to attrition. However, even being **promoted** in recently, someone may still choose to leave for better opportunities outside or other factors. This is important from a learning perspective as it could be used by HR department to guide future talent retention strategy and policy.

- Monthly Income: people on higher wages are less likely to leave the company. Hence, efforts should be made to gather information on industry benchmarks in the current local market to determine if the company is providing competitive wages.
- Over Time: We don't find clear evidence that over-time work triggers people to leave. From my point of view, there are two possibilities. If the employee takes serious about the free time and work-life balance, overtime is high likely resulting in attrition. In reverse, some employees may be possibly very interested in his job or has strong career ambitions, then overtime is not a problem.
- Promotion: Without doubt, promotion opportunity and policy are vital for talent retention in general. But the whole talent retention strategy should be designed from multiple perspectives and dimensions, including talents' career path, salary, working environment, management style, company culture, etc.

Ethical Statement:

The primary dataset for this analysis is fictional from IBM. Before referring to the findings to design any talent retention policy or strategy, we highly recommend to verify the results with real datasets. In addition, taking care of the applicable range and area of the conclusion to avoid biases, such as the limitation of country, industry, etc. Finally, the conclusion or prediction may relate to sensitive features, such as age and gender, which could possibly lead to discrimination. Before any algorithm applied for this task, the risks and possible damages should be systematically evaluated.

Next Steps:

- **Data collection:** Find more comparable data set with more features.
- **Modeling:** Optimize model selection, and further optimize parameters, improve prediction accuracy

Statement of Work:

The whole project is really a teamwork. Even with only 3 of us in this team, we experienced the four stages of team development including **Forming**, **Storming**, **Norming** and **Performing** (Tuckman). Zhipeng is good at team building and project management, driving the discussion and direction of the project. Chihshen owns enormous knowledge about correlation and causation, and equipped with high-level skills of data manipulation and visualization. Qi Zhang is a quick thinker and always performing in high efficiency.

Zhipeng Luo:

- Team Building and management.
- Data source research and selection.
- Working platform setup (GitHub)
- Data manipulation.
- Data visualization.
- Machine Learning model discussion and evaluation.
- Project report writing.

Chihshen Hsu :

- Data source research and selection.
- Working platform setup (DeepNote)
- Data manipulation.
- Data visualization.
- Machine Learning model discussion and evaluation.
- Project report writing.

Qi Zhang:

- Data source research and selection.
- Data manipulation.
- Data visualization.
- Machine Learning model design, discussion and evaluation.
- Project report framework design.
- Project report writing.

Reference:

1. Bendemra, H. (2019, March 11). *Building an Employee Churn Model in Python to Develop a Strategic Retention Plan.* Towards Data Science. https://towardsdatascience.com/building-an-employee-churn-model-in-python-to-develop-a-strategic-retention-plan-57d5bd882c2d

2. Booz, M. (2018, March 15). *These 3 Industries Have the Highest Talent Turnover Rates*. LinkedIn. https://www.linkedin.com/business/talent/blog/talent-strategy/industries-with-the-highest-turnover-rates#:~:text=The%20sectors%20seeing%20the%20most,2017%20with%20a%2013.2%25%20rate.

3. Circadian. (n.d.). *5 Negative Effects of High Overtime Levels*. https://www.circadian.com/blog/item/22-5-negative-effects-of-high-overtime-levels.html

4. Garfinkle, J. (n.d.). *Employee Retention & Promotion.* Garfinkle Executive Coaching. https://garfinkleexecutivecoaching.com/articles/how-to-retain-employees/employee-retention-promotion

5. Smet, A.D., et al. (2021, September 8). *'Great Attrition' or 'Great Attraction'? The choice is yours.* MckinSey. https://genesishrsolutions.com/peo-blog/types-of-employee-turnover/

6. Lim, S. (2021, September 2). *China rules '996 work culture' as illegal to prevent people from overworking*. The Drum. https://www.thedrum.com/news/2021/09/02/china-rules-996-work-culture-illegal-prevent-people-overworking#:~:text=China%20has%20ruled%20that%20the,being%20forced%20to%20work%20overtime

7. Tuckman's stages of group development. (2021, September 10). In Wikipedia. https://en.wikipedia.org/wiki/Tuckman%27s_stages_of_group_development

8. Imbalance Data. (2021, September 10). Google Developers. https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

9. The impact of salary on employee retention: what should you be paying? (2021, September 10). Randstad. https://rlc.randstadusa.com/for-business/learning-center/employee-retention/the-impact-of-salary-on-employee-retention-what-should-you-be-paying