Reducing Employee Churn: A data-Science Approach

****To understand how to stop a wave of departures from happening,
we first need to understand why waves happen in the first place****

Motivation

What is employee turnover.

It's a very competitive world out there where each organisation is trying to hire the best talent. Because they know that employees are the most essential contributors toward profits. Corporate leaders often proclaim that their employees are their most valuable asset. They are the most essential contributors toward profits and shareholder value. While every team and company is unique in some ways, each face the problem of employee churn.

Employee churn is basically voluntary or involuntary departures of workforce.This is also called as employee turnover. Company has to hire a new employee to fill up the position. They have to allocate a whole budget for marketing the position and then going through the whole hiring cycle and exhausting the resources for finally getting a new talent onboard. When you calculate the annual price for the whole process it cost more than 2 times the employee's annual salary who just left Ex: A 100-person organization that with an average salary of $50,000 could have turnover and replacement costs of approximately $660,000 to $2.6 million per year.

why it is such an huge issue and why we need a data science approach to address that
It barely matters if a company's employee turnover is above or below average, its workforce becomes temporarily less efficient every time someone leaves. In most companies When an employee leaves, there's an exit interview to know why employees are leaving and hoping to make changes that might minimize churn in the future. In some other companies surveys are done each month and variety of questions are asked about how they are doing, things going on in the company, and so on. A team will read the anonymous answers, and try to resolve the common issues stated by majority. However This method is not personalized. We need a proactive measure to find out who is having intent to leave and then figure out a way to make them stay. Retaining Talent Is More Cost Effective Than Hiring.It's not enough to just recruit the top talent. Current methods includes conducting exit interviews and feedback surveys to reduce employee churn in future which is highly ineffective as people barely answers their true reasons for leaving in hopes of leaving at good terms. Hence the cycle continues since there is no current way to determine who is going to leave next and why.

Secondly we need an effective way so that we can reduce the annual turnover cost. Apart from affecting the company expenses employee churn also affects the existing employees. Before the new employee is hired, though, the open position will probably be covered by other employees, diverting them from their regular work or requiring overtime or activity is simply not done. If it's a sales position, that's a direct hit to the top line. Longer lead times on products could be costly as well.

Finally It's the satisfaction level of your employees that matters the most. So, if an employee isn't happy, she might spread a negative word about the organization, even after leaving it. All this leads to bad company Image.

What's more, is that an unhappy employee will lack motivation and will not perform well, leading to unsatisfactory performance. This results in unachievable performance targets, low profits, and employee churn.

So we mounted a study to investigate the motivations of employees to stay or leave and reasons behind It. This data was then fed into a Random Forest Classifier for training.
" Our main aim is to develop an automatic system that takes basic/publicly available info of employees and figure out if employee is dissatisfied and has intent to leave and the reason that is making him/her do so, then organization can act to reinforce the right reasons to make them stay and stop reinforcing the wrong reasons." In other words, they can take a positive approach to managing retention in present , which will be more effective over the long run than the ordinary approach of simply conducting the exit interview in order to hope for reducing future turnover so that others don't leave.

****1) The Data****

IBM Attrition dataset is contains the data of 1470 IBM employees records and 35 feature variables.the Data is hosted as part of a competition by Kaggle. The target variable is "Attrition".

****2) Data Cleaning****

Once we have all the Data. We performed Data cleaning on it. Data Cleaning is necessary because the more clean the data is better is the model predictions. Garbage data results in inaccurate models that cannot make good predictions. So First we checked for missing values(nan, 0 or absence of any other value.). We also checked outliers using boxplots and capped the outlier data to a higher and lower limit. Next we checked if the datatypes of columns are consistent to the value stored in them. We also checked for duplicate data and removed them.

columns such as Employee count, Over18, StandardHours that has same value for all employees with single value for all column so we dropped those columns as they dont contribute to model predictions. Next we visualised correlation between variables. Correlation usually affects the performance of model. So if there is any feature value that can be derived from any other feature simply by applying some mathematical functions than we can remove that. Finally we converted categorical variables to numerical variables using one hot encoding as Model can only take data in numerical form.

There are other steps we could perform like grouping of data or binning data into appropriate range and values. Collapsing too many categories into few categories by keeping categories with high value counts and Replacing categories with very low value counts as 'other'.

****3) EDA****

1470 employees described by 35 features comprise the dataset. Eachemployee is represented exactly once in the dataset and thus the shape of the training data is (1470, 35). Figure 1 below illustrates the

large class imbalance with the following breakdown: 1233 - no attrition,237 - attrition. This stratification had to be taken into account when creating train/test splits.

****Is there a high corelation between certain factors?****

We could see that There is a high correlation between the columns "joblevel" and "monthly income". which is plausible as higher your job level, higher income you get. Another good correlation is between the column job level and and total working years so more experience people tends to go to higher job level. We could plot a heat map to see if there is any other corelation that exists in dataset.

****What factors have higher effect on atrrition?****

One of the factors that lead to atrrition is frequent travels. We could see from the image below that People with more frequent travel are willing to quit the company more often as compared to non travelling employees.

Another factor that seems to contribute to attrition is their current year in company. Attrition is high in employees in their first year of joining, then it again increases around 5th year. and then goes down after 10 or 11 years of work experience

Stock options also affects an employees willingness to leave. Most of the companies gives stocks to the employees progressively by the end of the year or after they finish certain time period of employment. Hence the employees has motivation to stay. Employees that dont get any stock options do not have any motivations to stay and hence leave early.

****4) Results:****

To evaluate the performance of my models, I chose the precision, recall, and f1-score metrics. These are defined as follows:

Precision and recall are favored over the true positive and true negative rates when there is a significant class imbalance, as in this case. The F1-score is the harmonic mean of precision and recall and gives us the ability to compare a single metric from one model to another.

We tried to establish a baseline model in this work which can then be compared to other models in future work, a random forest model. A Randomised CV algorithm is used to tune the hyperparameters of a model. Random Forest models are popular for many classification tasks due to their accuracy and ability to give feature importances. Precision, recall, and f1-score are calculated for each class in a binary fashion. The results are drastically different between classes. The precision was highest for the Attrition class (0.83) . In addition to the large class imbalance, the attrition class may simply be intrinsically harder to predict given the same set of predictor variables.

The recall score is very low (0.14). The results imply that the model varies widely in the number of false positives between classes.

We also plotted the RUC_auc score for this model and it is 87.24%. This translates to an average of 87.24% accuracy in predicting each class correctly.

Here accuracy score of the model was 87%. However our main evaluation matrix is confusion matrix. Here we chose precision/recall as our main matrix because we want to know that if the model is predicting 10 employees and out of those 10 if only 4 are actually intending to leave even then doing something good would be a nice gesture for other 6 employee and will increase their productivity. However if the resources are low and We want the model to only predict the employees who are leaving and not include the false positives as it hits the company resources. Then we can go for precision as our final matrix. Here we choose precision but Still the final metrics depends on what's the end goal of business.

Feature Importances:

Random forest also lets us to predict what are the top features that are contributing to attrition. So if most of the employees are leaving because of a certain factor the company can implement a new policy or try to chance the existing work conditions to make it easy for employees.

****5) Conclusion:****

Retaining existing talent is more economical and good in building a company image. In other words, through this we can figure out what factors are causing employee attrition and then we can take a positive approach to managing retention in present , which will be more effective over the long run than the ordinary approach of simply conducting the exit interview in order to hope for reducing future turnover so that others don't leave. This was a single class classification task. since attrition is something that cannot be verified instantly. It seems that our attrition is depended on multitude of factors but the top 3 were:

High in employees who are working a lot of overtime, so the company could address this by making sure people working overtime are supported with more resources so that it doesn't happen indefinitely into the future

High in employees who have to travel frequently, so the company could address this by making sure people who are willing to travel should be sent and non-important travels should be cut off.

High in employees with No stock options, company can either give other incentives/perks every year to these employees to compensate for lack of stock options.