EMPLOYEE ATTRITION ANALYSIS

Advised by Professor Foster Provost and Joshua Attenberg

Group #9

(Data Mining for Business Analytics)

Jeffrey Younghoon Kim

Saumya Goyal

Shimeng Cao

Yajash Pandey

Abstract

Employee retention is an important problem to businesses. Retaining talented employees is critical to the success of the firm. Also, replacing employees is costly as companies must incur expenses related to hiring new employees. In short, employee turnover impacts organizational performance. The data is a fictional set created by IBM, but we will treat it as real life data. Our project aims to predict employees that are likely to leave. However, in order to produce more useful insight from our investigation, we attempt to do causal inference and figure out the drivers for this attrition.

Table of Contents

I. Business Understanding

The world of data mining presents us to unique business problems with different companies in different sectors. To streamline the process of phrasing the business problem, we have learned a framework to apply the data mining process which comprises of many steps.

The first step of this process is Business Understanding which helps to understand our business problem, the use phase of the model and also how predictive modeling is helping us resolve the problem at hand. This is a key phase of the data mining process which helps us identify what we are trying to achieve through the data mining process at any stage in the entire process and helps us align with the business objective.

For an organization, Employee retention is an important problem. Retaining talented employees is critical to the success of the firm. Companies lose their employees due to various reasons. This results in additional cost of hiring new employees, training the employees and aligning the employees to the company culture. Resigning also results in increase of workload impacting other members of the team and overall work environment. The cost of an employee leaving is variable as different employees at different levels with different experience have different salaries and different impact when they leave the company. These values are not present in the current data set but can be evaluated using historical data. Employee attrition also impacts the performance of the organization. There is another aspect of analysis which is understanding the correlation between attrition (our target variable) and the features being used in the process of predictive modeling.

For our analysis, target variable is the probabilistic estimate of employees leaving the company within next one year. We have labelled training data for our modeling process which makes it a supervised learning problem. Our predictive model will identify employees that are likely to leave which will help the company to analyze and reduce attrition. Based on employee attributes, they will be classified as either likely to stay or leave the firm, with the probabilistic estimate. Our model also aims to figure the causal drivers of the attrition.

Furthermore, our model later can help the organization to design policy and procedure to retain the top talents critical to the success of the business. Acquired training data can also help the companies to analyze the replacement cost which is due to attrition and reduce that cost.

The second step in this process is to dig into the data, get a understanding of the data which leads us to data understanding.

II. Data Understanding

Data Link: https://docs.google.com/spreadsheets/d/1gvtGFeWLUrOM1_m0yJJVB_c0RKXnzQU91-l-dKLys_4/edit?usp=sharing

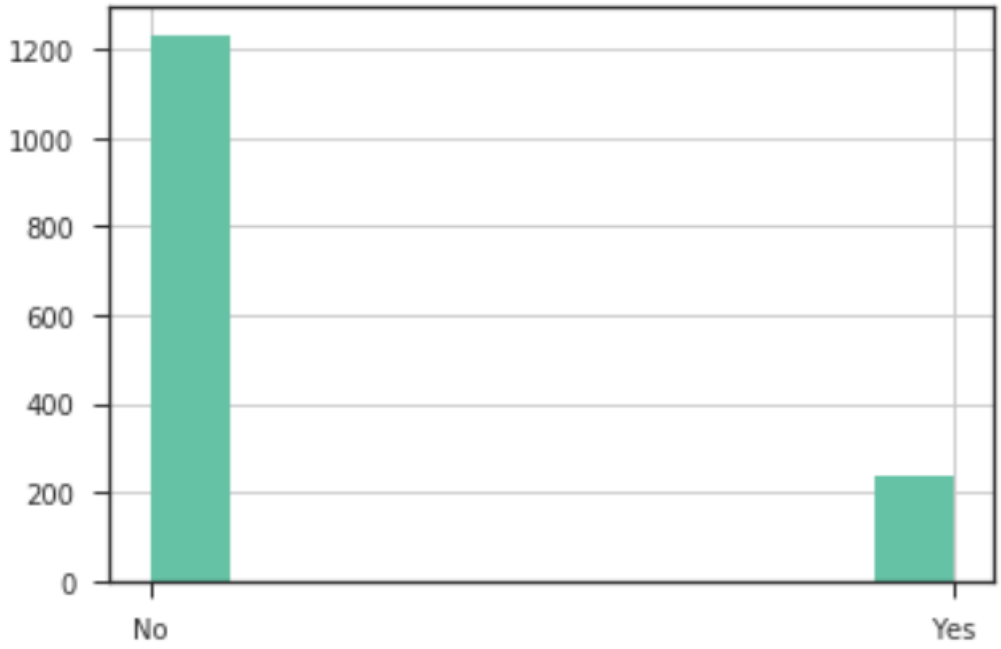
This phase is the second step in the data mining process and helps us analyze the other relevant data which can be used and analyze current data set to improve our predictive model. Data has to be considered as an asset which will improve our business solutions. If there are scenarios, where we don't have labelled data we have an option to buy the relevant data or may be pilot projects which can help us acquire data to perform better predictions or follow an unsupervised learning approach. As we have labelled training data, supervised learning can be done using this dataset.

The motivation driving the project is to create a pipeline with similar features to predict employee attrition. The data is a fictional set created by IBM scientist, but here we will treat it as real life data and our purpose of this project is to create a prediction that can apply for other Human Resource Departments with dataset similar to our original data set.

The data used here has contains 1470 instances of employee data who either left the firm or are still with the firm. The data consists of employee information including age, job role, travel time, hourly wage, gender, education, performance-rating, job satisfaction, years at company, years with current manager etc. Each row is for an employee with different data attributes. All this data about employees help us to implement different models to predict the target variable i.e attrition. Along with that we are using cross validations to use data effectively for training and testing purposes.

The different columns give us a feature vector which is the input for the model. There are different types of data attributes which are categorical and numerical data. For example, it can be inferred just by looking at the data that if an employee has low job satisfaction rate there might be attrition. The target variable in the data set, Attrition, is binary with labels which has been transformed to the values zero 0 and 1.

Histogram of the target variable (attrition) is shown below:



The details about how we handled different attribute types leads us to the third step in the data mining process which is data preparation.

III. Data Preparation

The major goal of the data preparation phase is featurization. The data consisted of both numerical and categorical data. A sample of the original dataset is shown below:

Attrition	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField
Yes	41	Travel_Rarely	1102	Sales	1	2	Life Sciences
No	49	Travel_Frequently	279	Research & Development	8	1	Life Sciences
Yes	37	Travel_Rarely	1373	Research & Development	2	2	Other
No	33	Travel_Frequently	1392	Research & Development	3	4	Life Sciences
No	27	Travel_Rarely	591	Research & Development	2	1	Medical

While some categorical data were strings and therefore easily identifiable (e.g. Department), others were categories represented as numbers (e.g. Education). In addition, numerical data themselves had different ranges of values with different order of magnitude (e.g. Age and Daily Rate).

Therefore, we used the pipeline utility to convert the categorical data to numerical data. We then transformed the numerical data via standardization and scaling to a range. After transformation, the number of columns increased from 35 to 80 (including target variable), and column names were explicitly assigned to new columns that were generated through the conversion of categorical to numerical data for ease. A sample of the transformed dataset is shown below:

Attrition	Age	BusinessTravel_ None	BusinessTravel_ Frequently	BusinessTravel_ Rarely	DailyRate
1.0	0.547619	0.0	0.0	1.0	0.742527
0.0	0.738095	0.0	1.0	0.0	-1.297775
1.0	0.452381	0.0	0.0	1.0	1.414363
0.0	0.357143	0.0	1.0	0.0	1.461466
0.0	0.214286	0.0	0.0	1.0	-0.524295

With enhanced featurization, the maximally informative data were fed into the machine learning models.

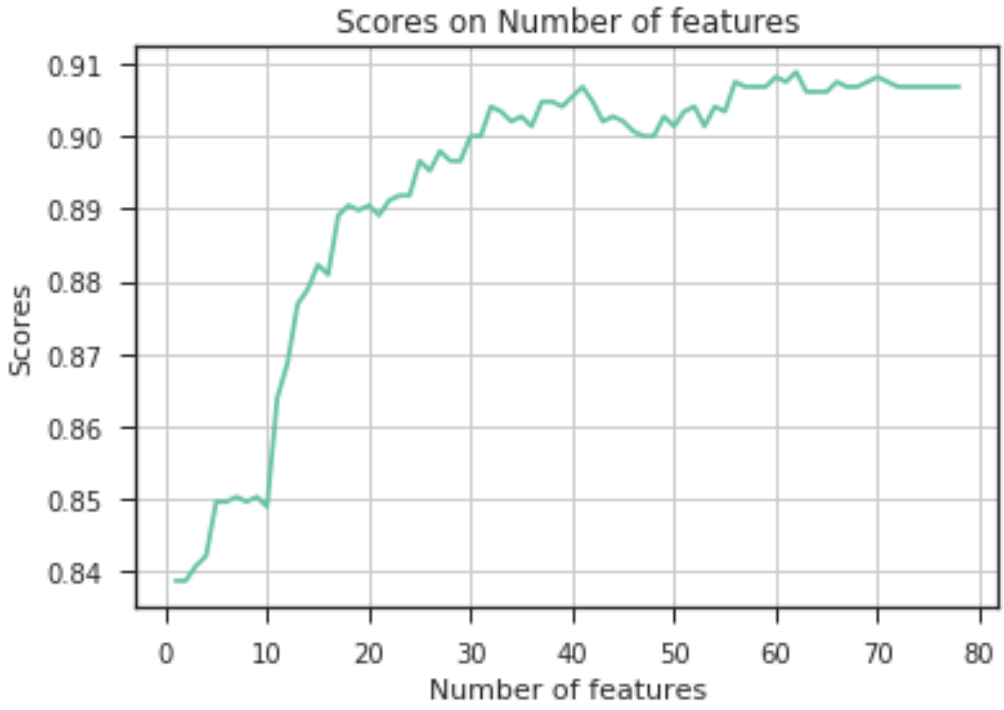
Another check performed for the data preparation process is to check for leakage. A leak is a situation where an informative variable in the historical data is not actually available when the decision has to be made. Such variables have to be checked for during the data preparation phase, and it appears that there isn't any obvious feature in our dataset that would create leakage.

Features which were almost the same for all employees such as Over 18, Employee Count and Standard hours were removed directly from the set as these were not predictive of the target variable whatsoever. The values had no impact on the Attrition. Similar results for these features are also shown as a part of Recursive Feature Elimination (RFE) where the rank of these features confirm our analysis.

IV. Modelling

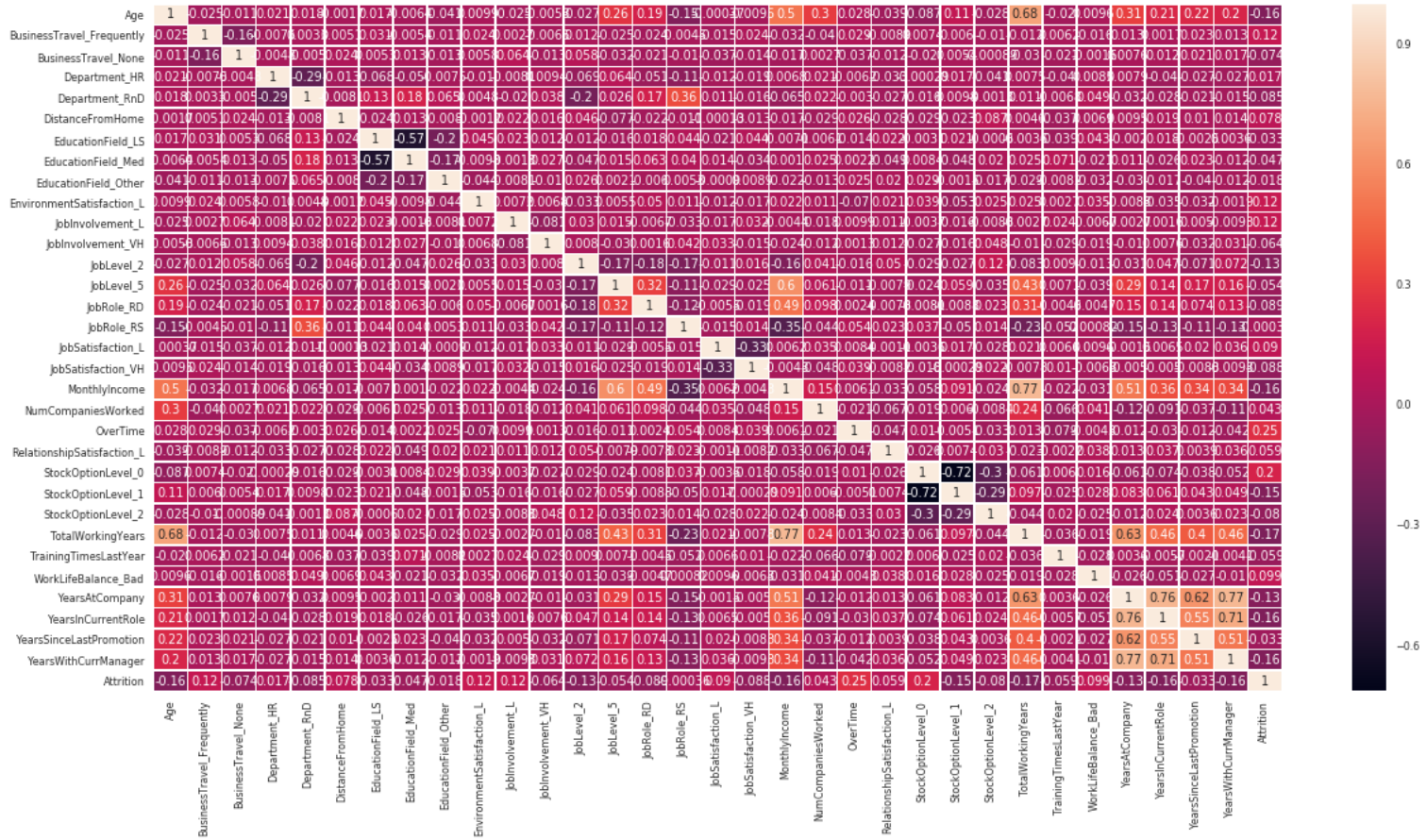
- Feature Selection

As a result of featurization, the data set ended up with a total of 79 features. As fitting all 79 features into the model could result in overfitting, we decided to utilize feature selection method to eliminate the unnecessary features from the group and select the most valuable ones. We utilized Best Subset Feature Selection method to determine the best subgroup of features we would use in our model within total 79 features. With Best Subset Feature Selection, we started with a null model which contains no predictors, and fit all possible combinations of k features (k ranging from 1 to 79) from 79 features into models that contain exactly k predictors. Then we used logistic regression model because the target variable is binary, and leveraged RFE from sklearn.feature_selection for selecting subset of features. For each selection and model fitted, a score was generated based on fitting. After plotting all the scores against number of features as illustrated below, we picked the number of subset based on the accuracy score, and in this case, 32 features were returned as the best number of features for its high score and least complexity involved as to avoid any future overfitting issue. Later, we used cross-validation and selected the top 32 ranked feature variables as the result of subset selection.



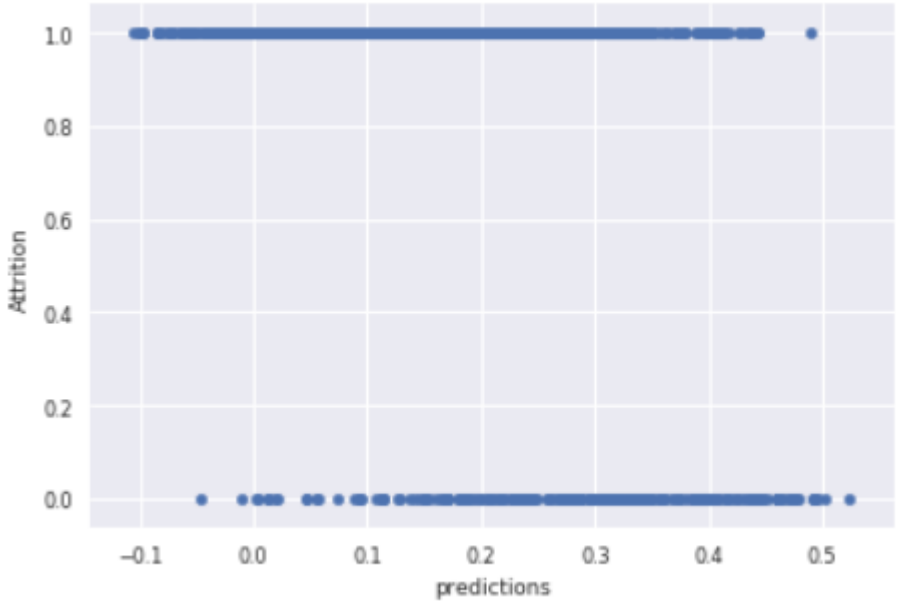
- Heatmap

After feature selection, to understand the data better we created a heatmap of the selected 32 features to find correlations with each other and with the target variable. This is a very simple way for us to understand how each variable correlates with each other, and potentially try to avoid confounders in future causal inference analysis. For example, based on the heatmap Joblevel_5 and Monthly_income is highly correlated and this can be explained that a senior person in the organization tends to earn more. Also it is observed that Overtime has the most correlation with Attrition and it can be interpreted that if an employee works for tremendous amount of over time he/she is more likely to leave the company than someone who doesn't work overtime. Although the coefficients help us understand the data better, it does not imply any causal inference between these features and Attrition. Spurious patterns in data can always be analyzed and based on correlation, we should not generalize causal relationships. To perform the causal analysis, the heatmap provides us a baseline to make hypothesis and test them using propensity score in later section of our analysis.



- Model Selection

Using a Linear Regression for our analysis is not a viable approach as the the target variable is not numerical. Also with linear model, classifying the instances which are very close to the decision boundary is very difficult which leads us to use other models for our business problem.



- Grid Search

In order to select the best algorithm, three models were built and compared: Logistic Regression, Random Forest, and Decision Tree. Grid search was used to tune hyper-parameters for these models and the area under the ROC curve (AUC) was used as a scoring measurement.

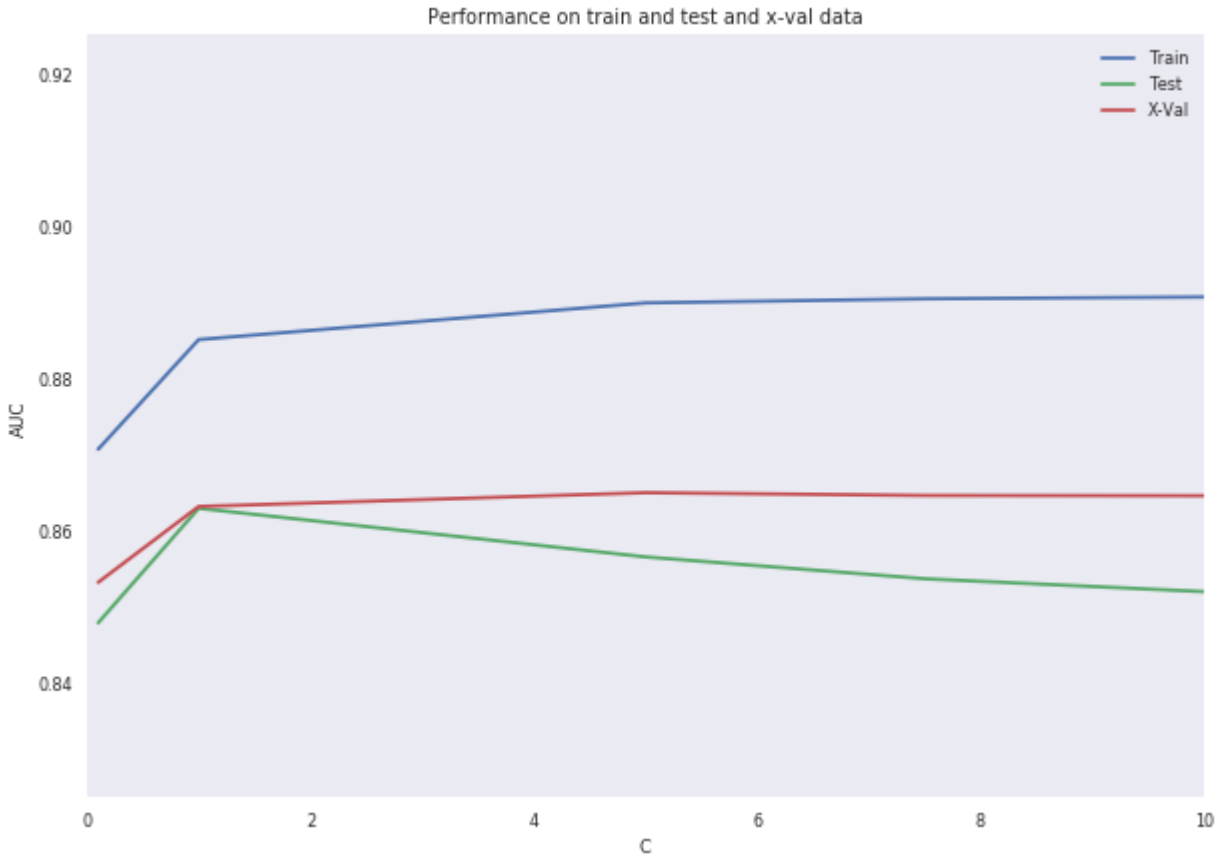
The grid search is a hyper-parameter optimization technique that explores all possible combinations of hyper-parameter values specified in a grid with a brute-force exhaustive search. It evaluates the model performance to select the optimal combination of hyper-parameter values that produces the best cross-validated score.

For this grid search, our transformed data with 32 attributes chosen from the feature selection were split into train and test sets with ratio of 0.75 to 0.25. The grid search then ran on the training data with three-fold cross validation and produced the following results:

	Best AUC	**Best Parameters**
Logistic Regression	0.865	penalty: L2, C: 5.0
Random Forest	0.813	n_estimators: 66
Decision Tree	0.753	min_samples_leaf : 20

Out of the three models that were analyzed, logistic regression produced the highest AUC. Therefore, selected as the model of choice going forward.

- Fitting Graph



A further study was conducted to analyze the impact of the model complexity. For this purpose, a fitting graph was constructed to show the accuracy of the model as a function of complexity. The fitting graph shows that the range of AUC values for all three data sets (training data, test data, and cross validation) are tight, ranging from 0.850 to

0.890. As the model becomes more complex, the accuracy of the training data continues to improve while the accuracy of the test data starts to decline. This is the symptom of the model overfitting to the training data. The accuracy of the cross-validation held steady despite increasing complexity.

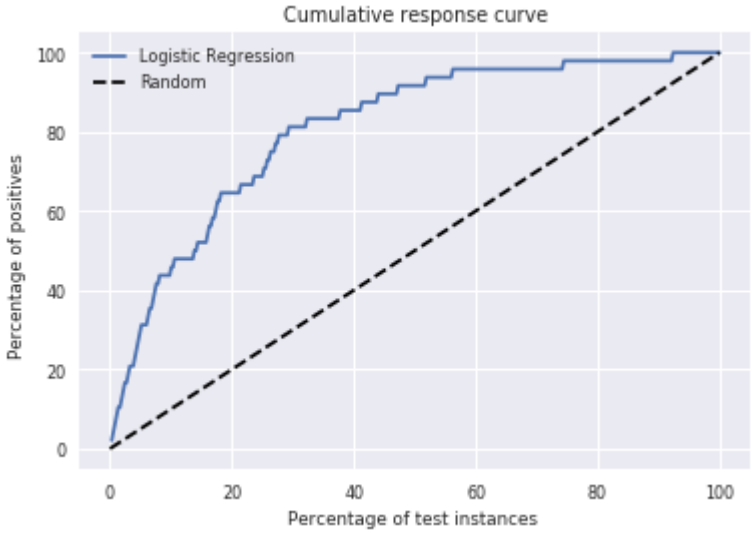
- Learning curves:

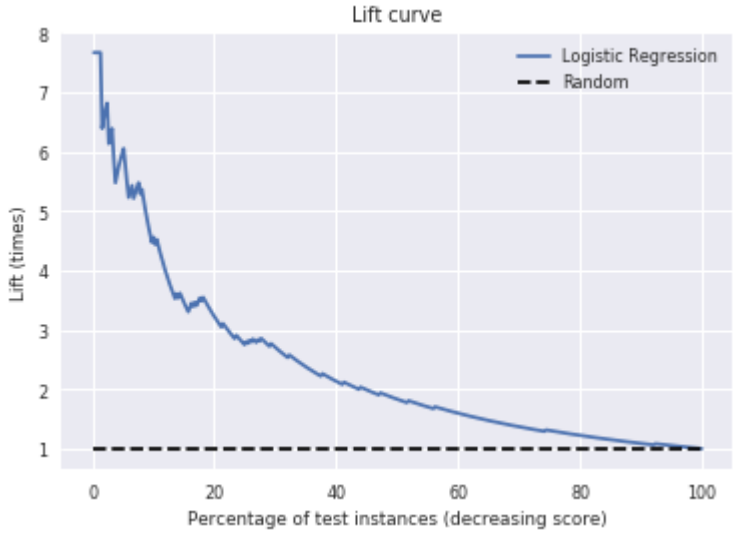
A plot of generalization performance against the amount of training data is a learning curve which helps us analyze the increase in accuracy as the amount of training data increases. It also helps to evaluate if we need to invest more in training data if the performance of the model has leveled off. In the below diagram, we can observe that logistic regression is able to generalize more with the same amount of training data as compared to decision tree and random forest.

- Cumulative Response Curve and Lift Curve:

The performance of the model is more intuitively visualized in the graphs presented below. A cumulative response curve is a plot of the true positive rate(tp) as a function of the percentage of the targeted population. As the

cumulative response curve shows below, the model is able to identify a larger proportion of the actual positives better than simple random guessing. However, since the degree of "bowing" of the cumulative response curve is difficult to judge by eye, the numeric lift is plotted on the lift curve.

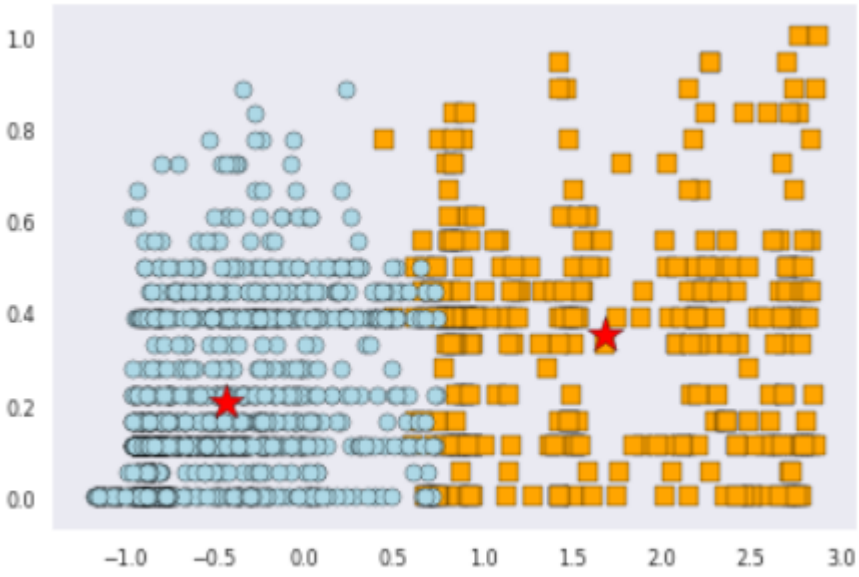




- K Means Clustering:

We then used K-Means clustering to identify two clusters which were passed as a parameter to the K Means Clustering algorithm. This approach helps us classify the employees in two clusters. Input for clustering should be scaled and normalized so that it does not impact the similarity measure calculation. We are trying to identify

employees in a particular cluster who have target variable of attrition as 1 and employees who have target variable of attrition as zero. This approach helps us evaluate two employees with similar features but a different value of the target variable which is Attrition. This analysis also helps us structure our hypothesis for Causal Inference where we can identify a particular attribute which might have impact on our target variable. By using similarity measures, and by computing |x0-x1|/x1 where x0 comes from cluster group with attrition equals 0 and x1 from cluster group with attrition equals 1 has helped us identify the features which were least similar. All features were computed using the "incremental change" methodology and ranked based on their values. The outcome of the analysis shows the highest ranked attributes: WorkLifeBalance_Bad and Overtime. Similarity measures have helped us in calculating these two attributes which have maximum difference.



This graph shows us a well clustered data set with centroids.

- Propensity Score Matching:

Based on the features being most dissimilar for employees in the cluster, we can base our hypothesis for causal modeling that WorkLifeBalance_Bad and Overtime can be potential drivers for attrition. The approach is now to do causal modeling and find employees which have similar propensity scores in the treated group and the non-treated group. Matching of employees is now based on a single number which is propensity score. This approach also reduces our work to analyze matching people based on all the set of features. There are different ways to estimate

the average treatment effect which are using OLS, Weighting and Matching. The average treatment effect can be calculated only after calculating the propensity scores. The feature which we have considered as a driver for the set of employees can be compared to verify our hypothesis.

After receiving the results from clustering and similarity matching, we try to test if the two features- WorkLifeBalance_Bad and Overtime have causal inference to attrition by calculating propensity scores. First of all, a propensity score is an approximate model of how likely a subject is to have been in the treatment group, $P(Di=1|Xi)$. We tried to fit the top ranked features with an intermediate regression $Di=\text{logistic}(\alpha+\beta Xi+\epsilon i)$. Later, we can estimate the propensities for both our control and treatment groups. Our goal is to find out if these two features have low propensity scores in treatment group while the rest have high propensity, and the opposite for the control group. We then find paired matches in the control group for each sample in the treatment group picking matches that minimize the differences in propensity score. Since we're dealing with estimates, we are not expecting perfect hypothesis result from testing the two features.

By leveraging causal inference package from python, we ran two separate set of models for the two features, and as a result, below two statistic tables show for both WorkLifeBalance_Bad and Overtime. Although we see positive treatment effects in all models and most of them are "significant", the Average Treatment Effect for WorkLifeBalance_Bad was too low to be considered. Therefore we are only considering OverTime as our potential driver for attrition.

For WorkLifeBalance_Bad:

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.137	0.053	2.592	0.010	0.033	0.241
ATC	0.138	0.054	2.584	0.010	0.033	0.243
ATT	0.118	0.049	2.408	0.016	0.022	0.214

For OverTime:

Treatment Effect Estimates: Matching

	Est.	S.e.	z	P>|z|	[95% Conf.	int.]
ATE	0.218	0.026	8.437	0.000	0.168	0.269
ATC	0.225	0.027	8.310	0.000	0.172	0.278
ATT	0.201	0.025	8.007	0.000	0.152	0.250

V. Evaluation

Evaluation is a key step in the our framework of predictive modeling.

- Accuracy (Rarely used)

Accuracy is one such way to evaluate model performance but there are issues with using accuracy. When the class distribution is skewed accuracy gives improper results.

- AUC

Therefore, we have used AUC to evaluate the accuracy of the model. Here, Logistic regression has the maximum AUC of 0.865 and by observing the learning curve, we understand how the model is able to generalize.

- Confusion Matrix

Another way to evaluate the model performance is by using the Confusion Matrix. A Confusion Matrix helps us to evaluate performance of the classifier by dealing with errors separately. The confusion matrix for our business problem is:

	p	n
Y	21	11
N	27	309

	p	n
Y	0.057065	0.029891
N	0.073370	0.839674

The value of the Precision from confusion matrix is : 0.656

The value of the Recall from confusion matrix is : 0.438

- Cost Benefit Matrix

In this scenario, using a Cost Benefit matrix for evaluation is difficult because with different employees, there are variable costs involved at different levels. Also, how the attrition impacts the deadline needs to be taken into account which will be different for different projects in the company. Implementing the same with the current data set is not feasible and we will need additional data to determine those costs involved.

A generic equation to structure the problem in the expected value framework will be in the following format:

X is the set of features for an employee.

Probability of an employee that the employee leaves given x is: $P(t|x)$

Probability of an employee that the employee does not leave given x is: $(1-P(t|x))$

Cost the company incurs if person leaves: $-V_t(x)$

Cost the company saves if the person does not leave: $V_{nt}(X)$

Expected loss will be: $P(t|x) *(-V_t(x)) + (1-P(t|x)) *V_{nt}(X)$

The cost benefit matrix with the probabilities help to structure the problem in a framework where more value can be added for business stakeholders to visualize the performance in terms of profit/loss and make better business decisions.

VI. Deployment

After evaluation of different models, we reach the final step of our framework which is deployment of the data mining model. After seeing that the model gives reasonable performance of training data using cross validation, we deploy the model in the production environment. The cycle does not stop after the deployment. It is very important to monitor the impact of the model after being deployed with time. Deploying the model again in production is a time consuming process which involves the effort of people and infrastructure. A model will make better predictions if feedback is taken from the data and is used to train the model again. Our goal is to enhance the model predictions over time which will benefit the business stakeholders.

Ethical Concerns:

Our dataset is pseudo-anonymous. Although an individual's name is omitted, there is sufficient data to enable identification of the individual. The data mining about individuals, in general, lead to ethical and privacy concerns. It has been observed that the effectiveness of business decisions has improved with increased use of personal data in business problems. As future professionals in the data science domain, we have a responsibility to ensure that data is used in the correct way and privacy of individuals is not invaded. The type of data being used for analytics should be known to the the senior management of the company. In different global companies there might be different policies in different geographical location. A company has the responsibility to set up rules and regulations which show the views and thought process of the senior management at the company as well as highlight the culture of the company. HR Analytics team should be taking decisions keeping in mind the policies of the company as well ensuring individual privacy. Individuals should have the right to know about the data which is being stored related to them.

VII. Conclusion

The data mining process has helped us understand us the phases from business understanding to deployment of the model and has added more knowledge for future iterations for predicting employee attrition in the company. Different models have been used in this business problem and we can tune more hyper-parameters using grid search to find the model which will suit our business needs. The causal analysis done shows that Overtime is a potential driver. This analysis can be used to drive company policy and can be used to evaluate the rate of attrition in the company. Also for the next step, we can use Expected Value framework to evaluate the costs of employee attrition to the company and help the company to manage their employee replacement costs. This step involves prediction of such costs using historical data which will be a part of the business understanding phase. We can also invest in acquiring data which helps us analyze the costs to add more value to the model which will can be utilized to form strategic data driven decisions.