HR Analytics: Predicting Promotions

Predictive Analytics & Machine Learning
MSBA 315

Prepared by:
Tarek Oueidat
Makram Noureddine

Presented to:
Dr. Wael Khreich

Table of Contents

Abstract

Employees are significant resources contributing to the success and failure of any firm. Ensuring a fair appraisal and rewards system is key for a proper employees' retention and career development. Furthermore, with the rapid evolution of machine learning ML techniques, it is now crucial to employ ML tools in HR analytics as assistive data-driven methods capable of minimizing biases in organizational decisions. This experiment aimed at producing a reliable classifier predicting the likelihood of promotion for an employee with a specific set of relevant features. For the sake of credibility, the data was running anonymously on a presumably real employees' dataset. The final classification model issued a relatively high performance achieving a precision score of 92%, overcoming previous research limitations, and providing a plausible framework for further studies.

Introduction

Human Resources' role has been developing drastically in the latest decade, going beyond adopting a classical recruitment, training and promotion style towards driving efficiency into the firm and impacting the bottom line. Since organizations are moving rapidly towards cutting costs and driving innovation, taking into account the ESG compliance mandate, a well-established HR system became a necessity: the ability to recruit the best-fit for the job and RETAIN high performing employees through a fair allocation of career development and promotions. To do so at a large scale and ensuring minimum biasness, the HR departments are seeking quantitative assistance in decision-making. The answer is HR Analytics making use of Machine Learning tools to enhance the overall employee's experience and rewards program based on statistical modeling. Research show that applying analytics at the HR level can boost profit margins by 4%, while achieving a 23% talents' ROI. Additionally, analytics uncover biases and generate actionable insights that usually don't come handy in conventional ways, even to the most seasoned HR professionals.

Literature Review

'Employees are both the biggest cost as well as the biggest asset of any organization.'[1]

While no study or research concentrated solely in the promotion side of HR roles and the analytics tailored to it, a significant and increasing number of studies are

1 Dr. Naman Sharma, HR Analytics: Opportunities, Issues and Challenges, Pragyaan: Journal of Management, vol 17: Issue 2, Jul-Dec 2019

addressing the role of analytics in human resources management and an imperative benchmark for HR data-driven decision making at any organization. Traditionally, many organizations relied on manual input of data in order to track their people's processes, but this has proven over time that it's a very inefficient.

Historical overview - The evolution of HR processes

Historical approaches of HR analytics and HR Information systems (HRIS) show the importance of employees' data in assisting and streamlining the business's performance. Building up a data infrastructure is a necessary means for Data analytics and predictive modeling for human resources over the past decades. Businesses began automating HR processes as far as the1980s, consisting of handling payroll and administrative tasks. The accumulation of data led to a wider adoption of HRIS by organizations (DeSanctis, 1986; Davis, 1989) [3][2]. The 1990s witnessed the real wave of HRIS pushing through studies emphasizing on the benefits of integrating such system to reduce HR administrative processes (Kossek et al., 1994; Hannon et al., 1996) [5][4], let alone the effect of the Internet on the need for automation. After the year 2000, HR professionals became aware of the crucial need for data and information technology in their decision-making process. Technology was increasingly used on data availability to make decisions about performance management and compensation (CedarCrestone, 2006) [1]. If we expand on that statement, promotions fall under this category. In Deloitte's 2020 Global Human Capital Trends, surveys showed that companies were least likely to collect workforce information. This suggests that many organizations may not be focusing their workforce data collection efforts effectively (2020 Deloitte GCAT) [7]. After conquering reluctance to provide personal and professional data by individuals across societies, the last decades uncovered massive data provision. We emphasize on data provision since it is the primary driver of analytics and therefore HR analytics. Digitization gave firms means for strategizing in HR. Based on the outcomes of Delloitte (2016) [8] study, 24% of the companies expressed their needs for analytics in HR role as well as in other departments. Google implemented a people focused strategy, which granted the Tech-Giant a competitive advantage in that field (Sullivan J, 2013) [6]. Furthermore, a study by "Kaur and Fink" in 2017 stipulated that, in terms of technical tools, the most used technologies for HR Analytics include R, Tableau, Python, SPSS and Excel. It is worth mentioning that the analysis and visualization conducted throughout this project uses Python as a statistical and analytical language.

Related Studies

Paper 1: "Predicting employee attrition using machine learning techniques", by: Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano and Ernesto William De Luca (2020)

The study is seeking data-driven decision in HR to uncover the likelihood of employee's attrition, which in turn reflects on the organizational employee related

decisions for retention, career development, promotions and so on. The researchers used ML methods to develop a model explaining the extent to which objective factors influence attrition. The trained data is a real dataset provided by IBM analytics, including 35 features and 1500 samples. Many classifiers were test (Bernoulli-NB, Logistic Regression, KNN, Random Forest, SVM), however since they were interested in predicting the greatest number of attritions therefore minimizing False Negatives: Gaussian Naïve Bayes was adopted having the highest Recall 54% and overall False Negative of 4.5%.

Paper 2: "Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm", by: Ananya Sarker, S.M. Shamim, Dr. Md. Shahiduz Zama & Md. Mustafizur Rahman. (2018)

The study aims at predicting the employees' performance for the next year based on personal and professional performance factors (i.e., punctuality, personality…) assessed as metrics. The result of the developed model will result in an assistive data-based recommendation of promotion for efficient workers and likelihood of discharge for others. The data trained consisted of 100 worker samples over 4 years of data keeping. Attributed to each sample (record) were 19 performance features (as metrics) to be tested for prediction. The model adopted was hybrid using K-Means Clustering and Decision Tree. The output was five performance classifications ranging from 1 to 5, 1 being poor and 5 being very good. K-Means Clustering (Unsupervised Learning) split the data into 5 classes for each year and labeled them to be fed later on into the Decision Tree (Supervised Learning) to develop a predictive model based on the Entropy technique based on previous years partition. The Decision Leaves will predict the employee's performance and measure it whether it will be Excellent, Good, Medium or Poor. The result was compared against the real data and was fairly performing.

Paper 3: "Early Prediction of Employee Attrition", by: B. Sri Harsha, A. Jithendra Varaprasad, L.V N Pavan Sai Sujith (2020)

The study tries to develop an analytical tool using Machine Learning to predict the likelihood of attrition among employees to have more data insight in HR the decision-making process since new hires are expensive in terms of time and money consumption. The model will predict employee attrition rate to anticipate the problem and solve it. The dataset used is the openly available engineered IBM Watson Analytics1. It contains 1470 workers observations with 32 HR features. 1233 were labeled NO-Attrition, while 237 were labeled YES-Attrition. The models trained were: Naïve Bayes, KNN, Random Forest, SVM. The final performance was tested using accuracy, precision, recall, F1 and AUC. After feature selection and preprocessing, the models were running on the dataset and the winner was the SVM model having accuracy 88.44%, precision 45%, recall 72.72%, F1 55.65%, AUC 70.91.

Paper 4: "Predictive analytics of HR, A machine learning approach", by: V.

Kakulapati, Kalluri Krishna Chaitanya, Kolli Vamsi, Guru Chaitanya, and Ponugoti Akshay (2020)

The goal is to categorize employees as the furthermost likely to grow promoted. The dataset is from UCI machinery containing above 50,000 records with 7 attributed features, 1000 were trained after splitting 70/30 and preprocessing. The model used was the optimized K-Means Clustering to define how many clusters to have based on which the HR will elect to employ new candidates or promote existing ones depending on the characteristics needed. It is an unsupervised learning approach, and the patterns are subject for change as well as the accuracy.

Data Exploration and Description - EDA:

As we began examining the dataset in hand, which represents, allegedly, 54808 records of a real company's employees posted recently on Kaggle for the purpose of predicting the likelihood of "promotion" based on certain characteristics, the first step to think of was to look into the type of data and its initial distribution.
The examination is to be conducted based on initial assumptions that should be validated in order to generalize the findings and results:

- Data is not presumed normally distributed therefore in scaling we use MinMaxScaler instead of standard scaling.
- The firm under investigation abides by the industry rules of hiring and promoting by taking into accounts equal opportunities amongst genders and social segments.
- The experiment targets employees falling under the same promotion criteria.

First-off, the records specify 11 features for each employee, beside his/her ID number, stating the academic, demographic, and professional status of each and whether he or she had been promoted or not, here we point at the dependent variable 'y' to be predicted. Therefore, after dropping the id, the features consist of 6 categorical variables and 5 numerical variables. One of the **constraints** to acknowledge during the project is the memory and computational limitations; therefore, it is crucial to select the most meaningful yet computationally efficient features to include in the modeling section. The feature that showed very low relevance to the purpose of the experiment and had 34 subcategories, which causes computational delays after encoding, was the "region" so we dropped it ahead of testing limiting the 10 remaining independent features. These 10 features were asserted its correspondent data types to avoid confusion. **Next** is confirming the previously stated assumptions. After plotting the distribution of the numerical features, we figured that it was mostly positively skewed and requires scaling and normalization post-splitting to avoid data leakage. In terms of industry equality standards, the figures show gender imbalance (38496 males in the firm versus 16312 female employees), however, the equal opportunity standard is upheld in terms of having both genders equal 'Average Training Score' averaging ~63, and the

percentage of promotion within each gender category apart is balanced as evident in (figure 1).

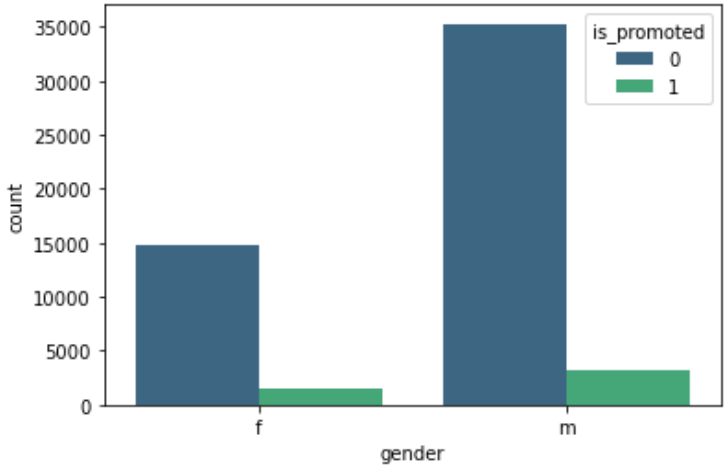

Figure 1

In terms of age allocation of employees, the majority of employees fall within the range of 23 to 46 years old (figure 2), while still having the distribution skewed towards ages above 50, which might affect our final model's judgment. The majority of this age category (above 50) does not follow the same promotion criteria regarding higher seniority and years of service. The best practice here was to cap both the age at 50 years and the length of service at 25 years to avoid misrepresentation.

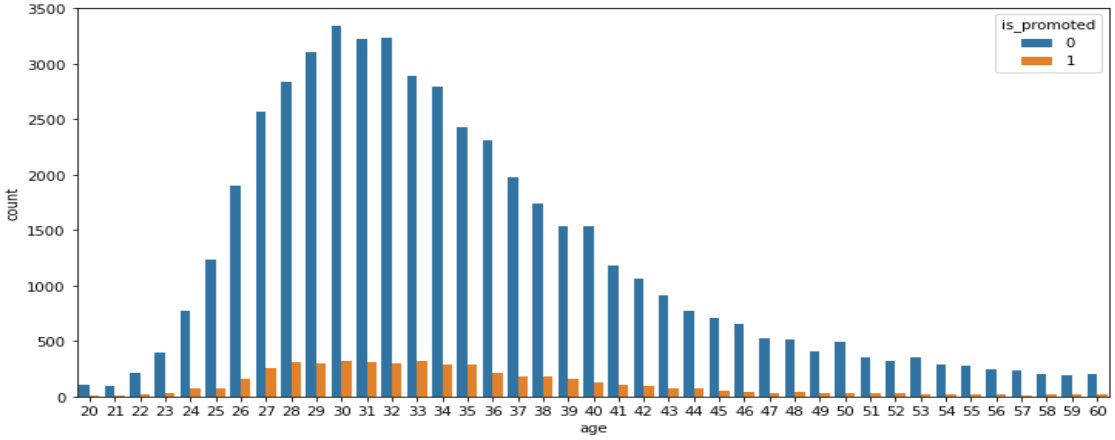

Figure 2

Additionally, the data showed three sources of recruitment: (a) sourcing, (b) referrals, and (c) other. The lowest number of recruits was by referral, but looking at it from a percentage standpoint, we see clearly that referrals had a higher percentage of offers, which makes sense (figure 3).

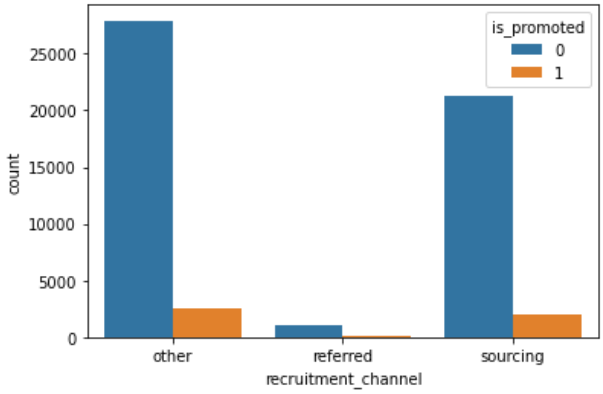

Figure 3

The promotion rate by education, taking into account the gender, showed advantage for male employees in jobs with below secondary academic level, whereas it showed a female advantage when it comes to jobs requiring a bachelor's and master's degree or above (figure 4). Promotion at this firm also factors in the awards won by the employees subject to promotion as shown in the number of employees promoted with and without awards.

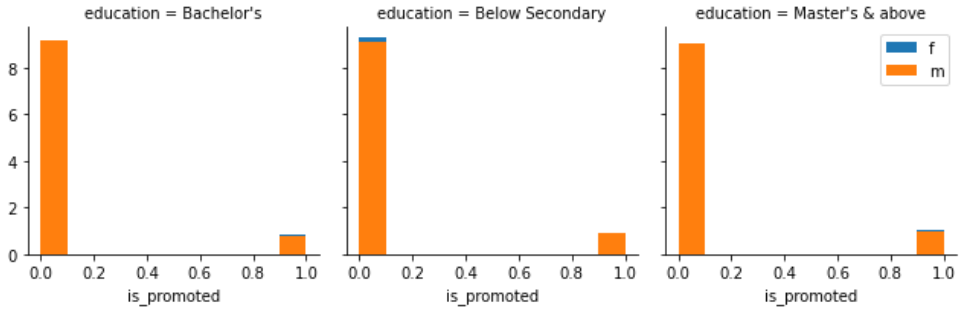

Figure 4

Null values existed in two columns: 'education' and 'previous_year_rating.' This will be tackled in the training and testing phased by adopting an Imputer integral to a Pipeline as a preprocessing step. Moreover, the outliers detected by the boxplots of distribution by features (figure 5) will also be tackled as part of the preprocessing transformation in the Pipeline.

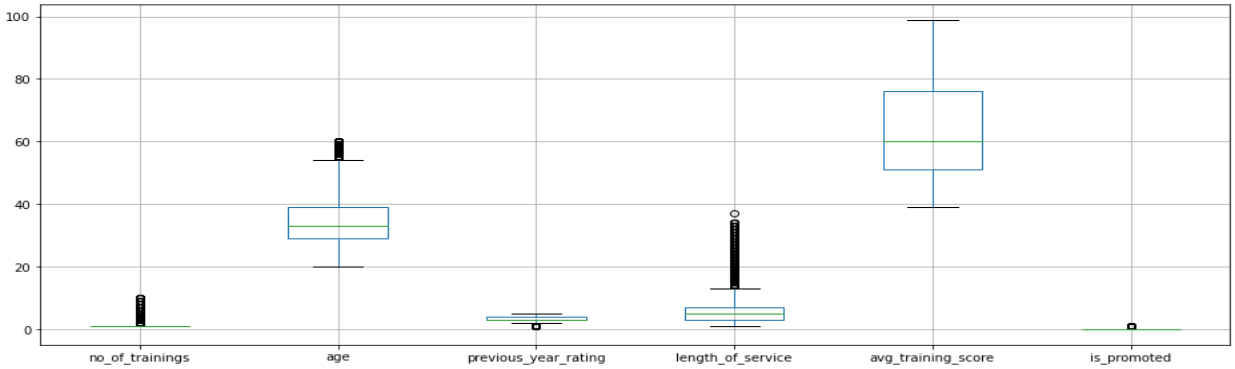

Figure 5 Boxplots

At this level, we recognized the purpose of the experiment, which is to generate an optimized predictive classifier to detect the likelihood of granting a promotion or not since the process of promotion and recruitment on an HR level as well as at Operational level is costly to firms in such a dynamic business environment.

Methodology:

Imbalance Classification

The target variable in the data set in a binary output (0 or 1, 1 being the likelihood of promotion). Just as it is the case in most binary classifiers, we need to examine first the proportion of positive versus negative outputs to avoid falling in the trap of disproportionate sampling rendering the results void in explanatory power. The dataset showed 91.48% of the result to be 0, meaning that without recurrence to any predictive model, if we guess 'No Promotion' every time, we will be accurate 91.48% of the time. It is therefore an Imbalanced Dataset. What is Imbalance? An imbalanced classification problem is an example of a classification problem where the distribution of examples across the known classes is biased or skewed. The imbalanced classifications are a serious challenge in machine learning, especially that most of the real-world data collected are usually imbalanced and must be tackled properly as most of the algorithms were built around the assumption of equal class distribution.

Data Split

We divided the original dataset into train and test subsets based on an 80:20 ratio and using a "stratified sampling" method as a parameter to ensure the optimal post-split balance regarding the target label, which is "is_promoted."

Data Preprocessing

Data preprocessing was performed on both the numerical and categorical columns in the training subset to avoid leakage. We tackled missing, as well as out of bound (outlier) records, by applying imputation and normalization (MinMaxScaler for numerical features and OneHotEncoder for categorical ones) within a Pipeline. Further explanation is as follows:

1- **Outlier Removal**:
Numerical Outliers lying above the upper bound and lower bound, which we have set to our distribution, were converted via a function into NAs. The function we coded and applied to the outliers have them now ready for the imputation phase using the Numerical Imputer.

2- **Numerical and Categorical Imputation**:
Two imputers were created accordingly. One imputer works on replacing all the NA values in the numerical columns by the median value of the same column,

simultaneously, a categorical imputer works on replacing all the NA values in the categorical columns with a token: "missing_value." The purpose of providing a new missing_value category is to avoid having category bias through overloading an existing category to the expense of the others.

3- **Categorical Data Conversion**:

Categorical variables were encoded, always through the same Pipeline steps and on each subset aside, using OneHotEncoder () to be able to feed it to the Machine Learning algorithms at a later stage.

4- **Data Normalization**:

As mentioned earlier, the data was not presumed to be normally distributed therefore in scaling we used MinMaxScaler () instead of standard scaling for our numerical attributes.

5- **Combining the preprocessing steps** into on wholistic Pipeline using the ColumnTransformer () function.

Metrics selection

In terms of metrics, reporting classification accuracy for any imbalanced classification problem could be seriously misleading. Choosing an appropriate metric is challenging as it should take into consideration what result we seek from the experiment. In this case, we are not interested in high accuracy. In our case, we are interested in predicting False Positives and the Precision metric. As contingent to the imbalance treatment, we used the Precision-Recall curve to compare the models' performances rather than ROC curve, which may mislead the observer in cases of imbalanced classification.

Resampling Techniques

Many machine learning algorithms don't perform well with an imbalance dataset. For that, many techniques can be applied in order overcome this impediment, one of which is adopting a Stratified K-Fold cross validation while sampling no matter the method used.

The imbalance library, 'imb-learn,' offers us several sampling techniques to use such as Random over or under sampling, oversampling using SMOTE **(Synthetic Minority Oversampling Technique**), combining SMOTE and Tomek Links Undersampling or SMOTE and Edited Nearest Neighbors Undersampling. Also, some machine learning algorithms supports the cost-sensitive strategy where a class weight parameter is assigned to some algorithms such as Logistic Regression and Decision Trees in order to overcome the imbalance.

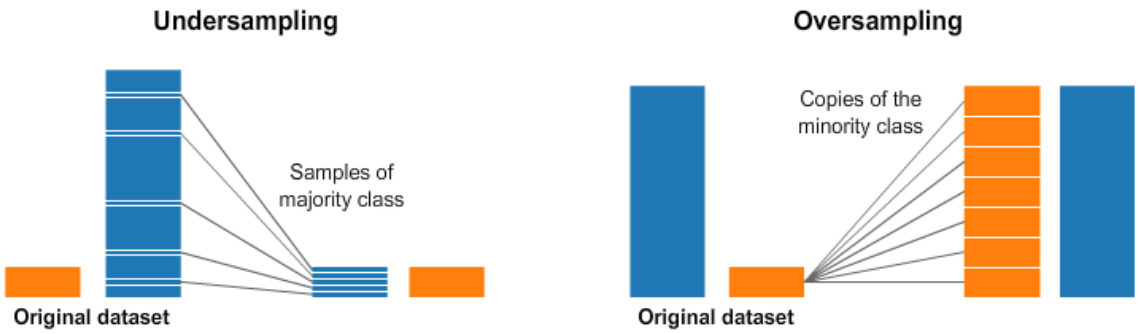

Figure 6 source: KDnuggets website

Many trials regarding the sampling strategy were made in order to choose the most efficient strategy in terms of metric results on the training subset. We need to also factor in the memory capacity and computational speed as a constraint in the research. The final sampling strategy adopted was oversampling the minority class using SMOTE with a sampling strategy parameter of 0.2, and afterwards conducting random under-sample of the majority class also using a sampling strategy of 0.2. This led to an equally distributed dataset, ready for the machine learning algorithmic processing and evaluation. It is important to emphasize that in order to avoid any data leakage, the resampling technique's steps were fed into a specific pipeline provided by the imbalanced learn library.

Predictive data modeling – Machine Learning Algorithms

The modeling process in this project consists of choosing several predictive machine learning techniques, applying them, and finally validate the best model to be generalized based on metric evaluations. The trained models in this experiment are as follows:

- Logistic Regression classifier
- ComplimentNB
- K-nearest neighbours (K-NN)
- Linear Support Vector Machines (LSVM) classification.
- Decision tree classifier,
- Random forest classifier
- Gradient Boost Classifier
- Balanced Bagging Classifier (from the imbalanced library)

Note that: the algorithms that do not require any resampling prior to training were the ComplimentNB, the Gradient Boost, and the Bagging classifier. The reason for that is that it embedded resampling within its internal selection techniques. For instance, the Gradient Boost does not randomly bootstrap the records, it does learn from its previous selection going forward, which means we end up overfitting the data if we apply resampling here.

Models Optimization

For the purpose of optimizing the previously stated models and extracting the best parameters and issue a well-rounded model recommendation, we applied the GridSearchCV () function on all of the models. In each case, we specified the most prominent parameters with different values in order to save the computational power for later, for instance, in the case of Linear SVC, we placed emphasis on the 'C' value in order to minimize any potential under- and over- fitting the data. The best model performance as well as the best parameters will be cited in the next section as part of the result interpretation.

Results and Discussions

To begin with, we initialized each model separately and applied the GridSearchCV optimization model **within the main preprocessing pipeline**. The results sought for model evaluation and optimal selection were primarily the Precision score and also the Recall, Accuracy, and F1 score. Furthermore, to be able to compare and contrast the model's performance visually, and since it was initially an imbalanced experiment, we substitute the area under the curve AUC or the ROC curve by the Precision-Recall curve after researching previous studies for that matter.
The concluded results after having cross-validation conducted within each model on the training subset are displayed in the table below.

	Model	Precision	ROC_AUC	RECALL	F1	ACCURACY
0	Logistic Regression	0.864987	0.773898	0.275806	0.417793	0.933862
1	KNN	0.439367	0.678275	0.234741	0.292908	0.903271
2	Compliment NB	0.135550	0.653761	0.530525	0.215850	0.667296
3	Linear SVC	0.896802	0.761311	0.320401	0.471938	0.938244
4	Decision Tree Classifier	0.767141	0.736402	0.229154	0.351217	0.927532
5	Random Forest Classifier	0.801077	0.766461	0.164695	0.272457	0.924451
6	Gradient Boost Classifier	0.915659	0.784182	0.349653	0.505734	0.941127
7	Balanced Bagging Classifier	0.677938	0.739079	0.366001	0.474962	0.930305

Table 1 Training data modeling metrics

Based on the selected results in Table 1, we elected to choose the best model to be the model having the highest Precision while still to the upside in terms of F1 and Recall scores. Here we have the **Gradient Boosting Classifier** displaying the best performance with:

- Precision: ~92%
- Recall: ~35%
- F1: 50.1%
- Accuracy: ~94%

The next best model is Linear Support vector Machine (Linear-SVC) when we consider both the Precision and the F1 scores. That being said, we can apply the best algorithm to the testing subset and issue a final recommendation. However, we elected to reapply all the fitted models to the test subset and further validate the findings observed in the first model processing phase. Before moving forward to the testing level, we extracted the feature importance for the chosen model (i.e., Gradient Boosting Classifier). This step will provide us with a solid insight on which characteristics contributed more to the promotion process.

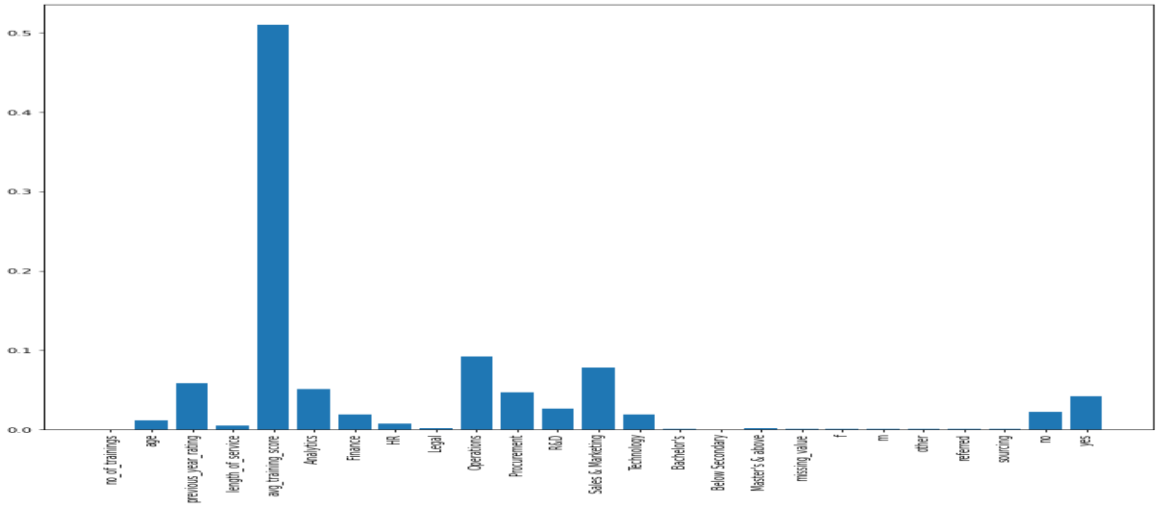

Figure 7 Feature importance - Gradient Boosting Classifier

The results as displayed in the bar plot show significance of the following features:

- Average training score
- Previous year rating
- People in Operations
- People in Procurement
- People in Sales and Marketing

After applying the previous models with its best estimators to the test subset, the results were as follows:

	Model	Precision	ROC_AUC	RECALL	F1	ACCURACY
0	Logistic Regression	0.850000	0.630749	0.265922	0.405106	0.932730
1	KNN	0.625000	0.581695	0.173184	0.271216	0.919834
2	Compliment NB	0.140265	0.614896	0.544134	0.223036	0.673467
3	Linear SVC	0.882736	0.649501	0.302793	0.450915	0.936483
4	Decision Tree Classifier	0.787879	0.627409	0.261453	0.392617	0.930324
5	Random Forest Classifier	0.824324	0.566787	0.136313	0.233941	0.923107
6	Gradient Boost Classifier	0.944625	0.661116	0.324022	0.482529	0.940141
7	Balanced Bagging Classifier	0.662420	0.665930	0.348603	0.456808	0.928592

Table 2 Testing data modeling metrics.

Based on the metrics above we confirmed the selection of the final fit in the Gradient Boosting Classifier model. Therefore, the model can be stored and generalized over other out of sample data that qualifies based on the initially stated assumptions.
We can further establish a visual confirmation of the models' relative performance through the Precision-Recall curve of each plotted against the rest for comparison.

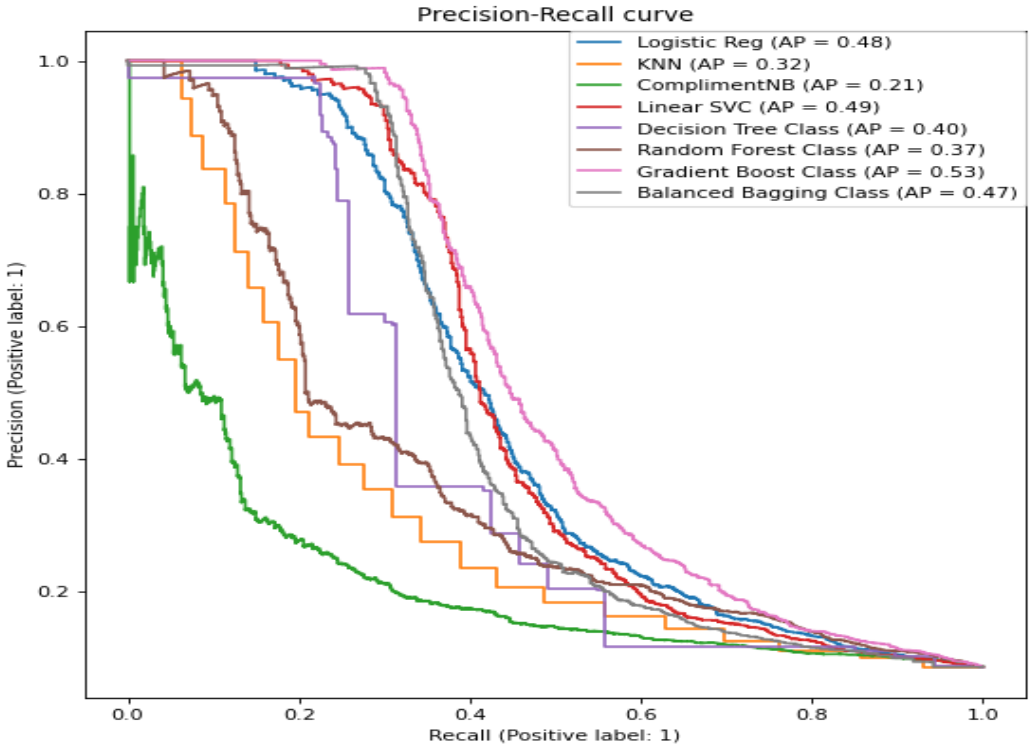

Figure 8 Precision-Recall Curves

Usually, the precision is more concerned by direct costs, the greater the number of false positives, the greater the cost per true positive. In this case, not being capable of correctly classifying the true positives puts the company in a bad financial position. It then offers more employees a chance for promotion resulting in higher salaries on the payroll. It also requires more senior positions/roles regardless of the business model and the firm's hierarchy. Such repercussions caused by issuing poor promotion judgments adds more burdens to the bottom line of the firm, pushing it outside of its competitive advantages towards bankruptcy in worst cases. Precision-Recall curves give a visual representation of the models achieving high recall while proportionally preserving a high precision. Technically speaking there is a tradeoff between the precision and the recall score; therefore, by looking at figure-8, we deduce that the most convenient model, having the greatest area under the precision-recall curve relative to its peers, is also the Gradient Boosting Classifier.

With respect to the obstacles and limitations confronted throughout the experiment, we need to shed light on the limited domestic computational power and memory constraints. Such limitations, when overcame, may result in further optimization of the algorithms and may possibly skew the results in favor of other models. However, for the purpose of this test, the results obtained were satisfactory.

We also plotted the Learning Curve using the Precision as a reference to identify whether additional optimization steps such as reducing overfitting or adding more complexity and/or more observations to enhance the model's performance.

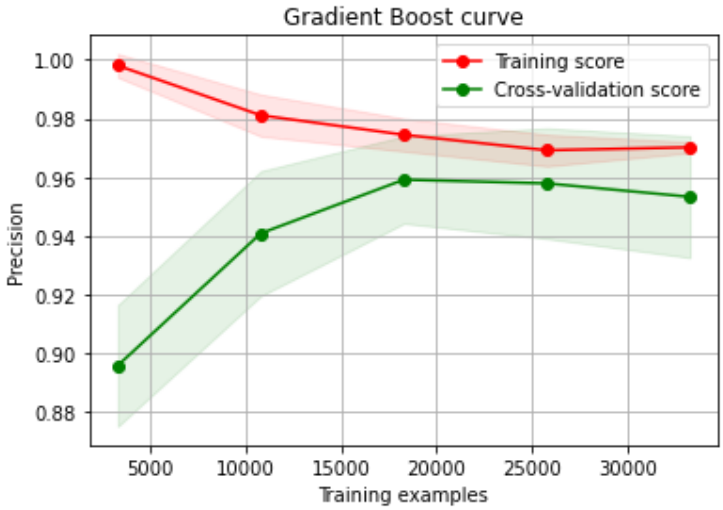

Figure 9 Precision Learning Curve for Gradient Boosting Classifier

The Learning curve shows initial steep convergence between the Precision scores of both training and cross-validation sets, followed by stabilization starting at a training complexity of around 18,000 records going forward. This indicates that any further addition of records is not being useful for optimization. The precision scores are satisfactory ranging between 0.95 and 0.97.

Here some points present the differentiation between this classification experiment and previously selected studies, therefore highlighting a high credibility of outcomes:

1. We specified a set of criterions and assumptions based on which the results of the experiment will be validated instead of simply taking the data as is and applying various modeling and exploration techniques.
2. The experiment reached a significantly higher metric results especially in terms of precision score.
3. It also tackled the problem of imbalance, which is mainly the issue facing most classification projects and consequently causing lower confidence in the results obtained.

Conclusion and Recommendations:

In order to confidently predict employees' promotion likelihood as an assistive recommendation tool, we built predictive machine learning classifiers fitted and tested to a firm's employees' dataset. Prior to modeling, data exploration showed equal opportunity amongst genders, it also confirmed compliance with our prespecified assumptions qualifying for the prediction phase. However, the binary

target, 'promotion likelihood,' showed imbalance. We applied resampling techniques and ran the dataset over eight optimized classification models. The process led to shortlisting the models by performance (mainly by precision) and we ended up selecting the Gradient Boosting Classifier. The chosen model displayed the highest Precision-Recall combination, 92% for precision and 35% for recall. We have chosen to emphasize on the precision mainly because we are mostly interested in accurately predicting True Positives and False Positives to keep the firm on the upside, avoiding financial burdens. The end result showed confidence in adopting the model for further predictions under the condition of having an analogous data following the same promotion criteria.

This experiment recommends further investigation of firms' roles and exploration of other relevant features such as: job level, monthly income, overtime, and business travel. These additional factors might have influence on the promotion decision. We can also seek to get more insights into the business activities and market capitalization (size) of each firm under investigation. We can then cluster similar firms together and then apply separate classification tests for each cluster. This can be helpful in generalizing the models obtained and takes into consideration different criteria and hierarchies in the market. Note that for the sake of clustering, different supervised (KNN) as well as unsupervised machine learning techniques (K-Means-Clustering) can be put together as a pre-classification phase. We can also elect to use larger memory capacity and computation power for further optimization; however, it should consider by case to be as efficient in resource and time management.

References

1. CedarCrestone (2006), CedarCrestone 2006 Workforce Technologies and Service Delivery ApproachesSurvey, 9th Annual ed., CedarCrestone, Alpharette, GA.
2. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3). 319-340.
3. DeSanctis, G. (1986). Human resource information systems: a current assessment. MIS quarterly, 15-27.
4. Hannon, J., Jelf, G., & Brandes, D. (1996). Human resource information systems: Operational issues and strategic considerations in a global environment. International Journal of Human Resource Management, 7(1), 245-269.
5. Kossek, E. E., Young, W., Gash, D. C., & Nichol, V. (1994). Waiting for innovation in the human resources department: Godot implements a human resource information system. Human Resource Management, 33(1), 135-159.
6. Sullivan J How Google is using People Analytics to Completely Reinvent HR Article in HR Management, HR News & Trends, retrieved on 8th November' 2014 from
7. Deloitte Global Human Capital Trends 2016 https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2016/people-analytics-in-hr-analytics-teams.html
8. 2020 Deloitte Global Human Capital Trends https://www2.deloitte.com/content/dam/Deloitte/at/Documents/human-capital/at-hc-trends-2020.pdf
9. "Predicting employee attrition using machine learning techniques", by Francesca Fallucchi, Marco Coladangelo, Romeo Giuliano and Ernesto William De Luca (2020)
10. "Employee's Performance Analysis and Prediction using K-Means Clustering & Decision Tree Algorithm", by: Ananya Sarker, S.M. Shamim, Dr. Md. Shahiduz Zama & Md. Mustafizur Rahman. (2018)
11. "Early Prediction of Employee Attrition", by: B. Sri Harsha, A. Jithendra Varaprasad, L.V N Pavan Sai Sujith (2020)
12. "Predictive analytics of HR, A machine learning approach", by: V. Kakulapati, Kalluri Krishna Chaitanya, Kolli Vamsi, Guru Chaitanya, and Ponugoti Akshay (2020)