HR Analytics: Predicting Promotions

Predictive Analytics & Machine Learning
MSBA 315

Prepared by:
Tarek Oueidat
Makram Noureddine

HR Analytics background

"Employees are the **biggest cost** as well as the **biggest asset** of any organization"

HR Analytics is an innovative field that makes use of Machine Learning tools to enhance the overall employee's experience by improving the rewards and career development program.

The Key **advantage** here is the ability to apply statistical modeling in objective **data-driven** promotion and recruiting recommendations.

Researches show that applying analytics at the HR level can boost profit margins by 4%, while achieving a 23% talents' ROI.

Introduction and Objectives

- **Problem tackled:** The overwhelming impediments facing HR professionals in today's firms when trying to factor in objective criterions as part of the scoring process to elect employees for promotions.

- **Project Objective**: To explore a real-time employees' dataset & develop a Predictive ML Model capable of predicting the likelihood of Promotion based on a set of relevant characteristics.

Related Works

Spoiler Alert: Despite the fact that this experiment addresses the issue of promotion prediction, it is worth mentioning that such topic is not yet popular in previous studies,
However, some of the Previouses we listed did hover around the spectrum of the HR analytics as contingent to promotion.

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

- The trained data is a real dataset provided by IBM analytics, including 35 features and 1500 samples.
- Many classifiers were tested: Bernoulli-NB, Logistic Regression, KNN, Random Forest, SVM.
- Interested in minimizing False Negatives.
- Adopted Gaussian Naïve Baye.
- The highest Recall 54% and overall False Negative of 4.5%.

Paper 2: "Early Prediction of Employee Attrition", by: B. Sri Harsha, A. Jithendra Varaprasad, L.V N Pavan Sai Sujith (2020)

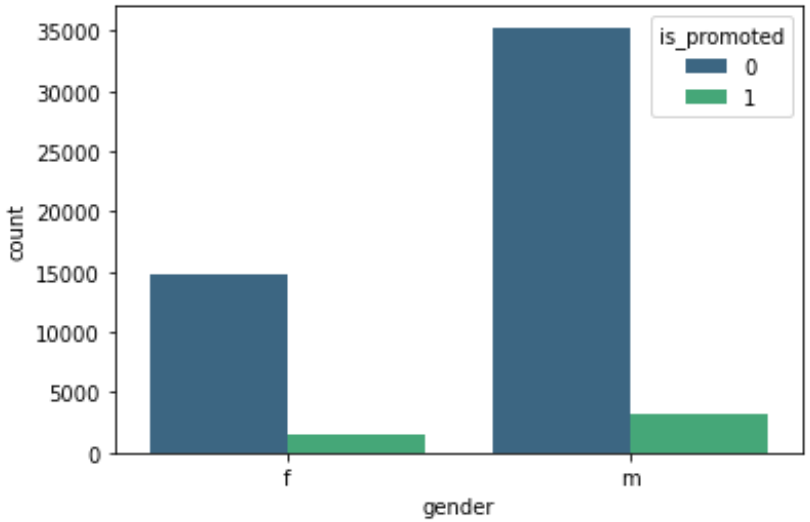
- The dataset used is the openly available engineered IBM Watson Analytics1.
- It contains 1470 workers observations with 32 HR features: 1233 were labeled NO-Attrition, while 237 were labeled YES-Attrition.
- The models trained were: Naïve Bayes, KNN, Random Forest, SVM.
- Tested using: accuracy, precision, recall, F1 and AUC.
- The winner was the SVM model having: accuracy 88.44%, precision 45%, recall 72.72%, F1 55.65%, AUC 70.91.

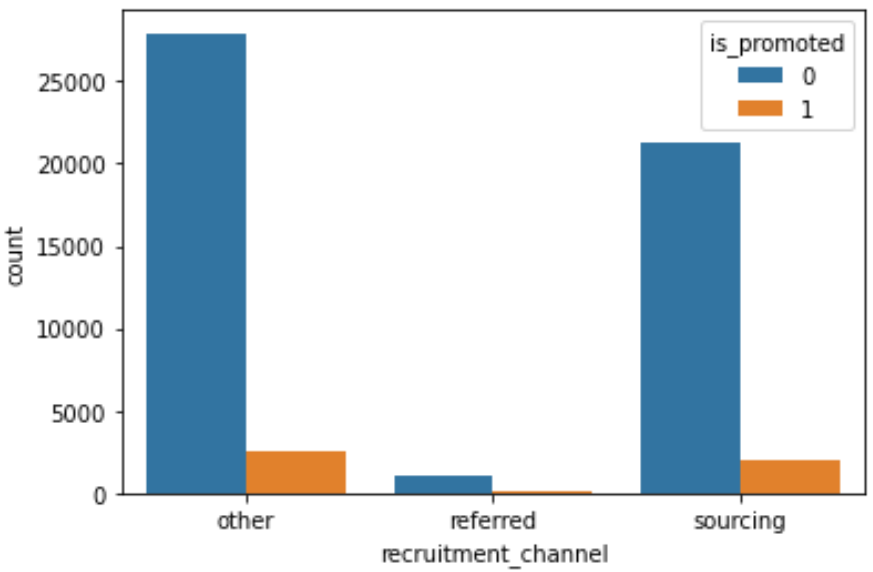
Data Description

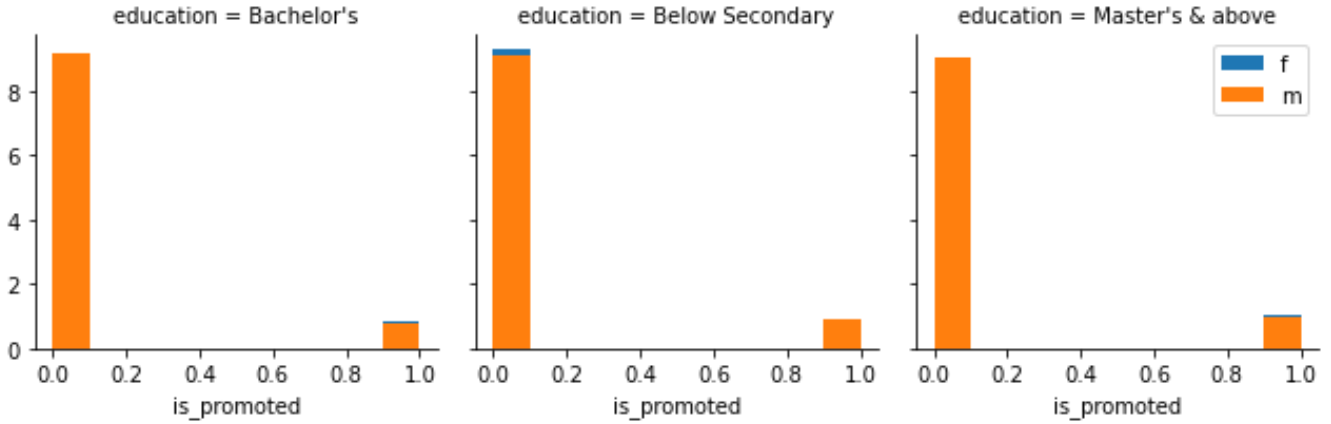
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54808 entries, 0 to 54807
Data columns (total 13 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   employee_id           54808 non-null  int64
 1   department            54808 non-null  object
 2   region                54808 non-null  object
 3   education             52399 non-null  object
 4   gender                54808 non-null  object
 5   recruitment_channel   54808 non-null  object
 6   no_of_trainings       54808 non-null  int64
 7   age                   54808 non-null  int64
 8   previous_year_rating  50684 non-null  float64
 9   length_of_service     54808 non-null  int64
 10  awards_won?           54808 non-null  int64
 11  avg_training_score    54808 non-null  int64
 12  is_promoted           54808 non-null  int64
dtypes: float64(1), int64(7), object(5)
memory usage: 5.4+ MB
```

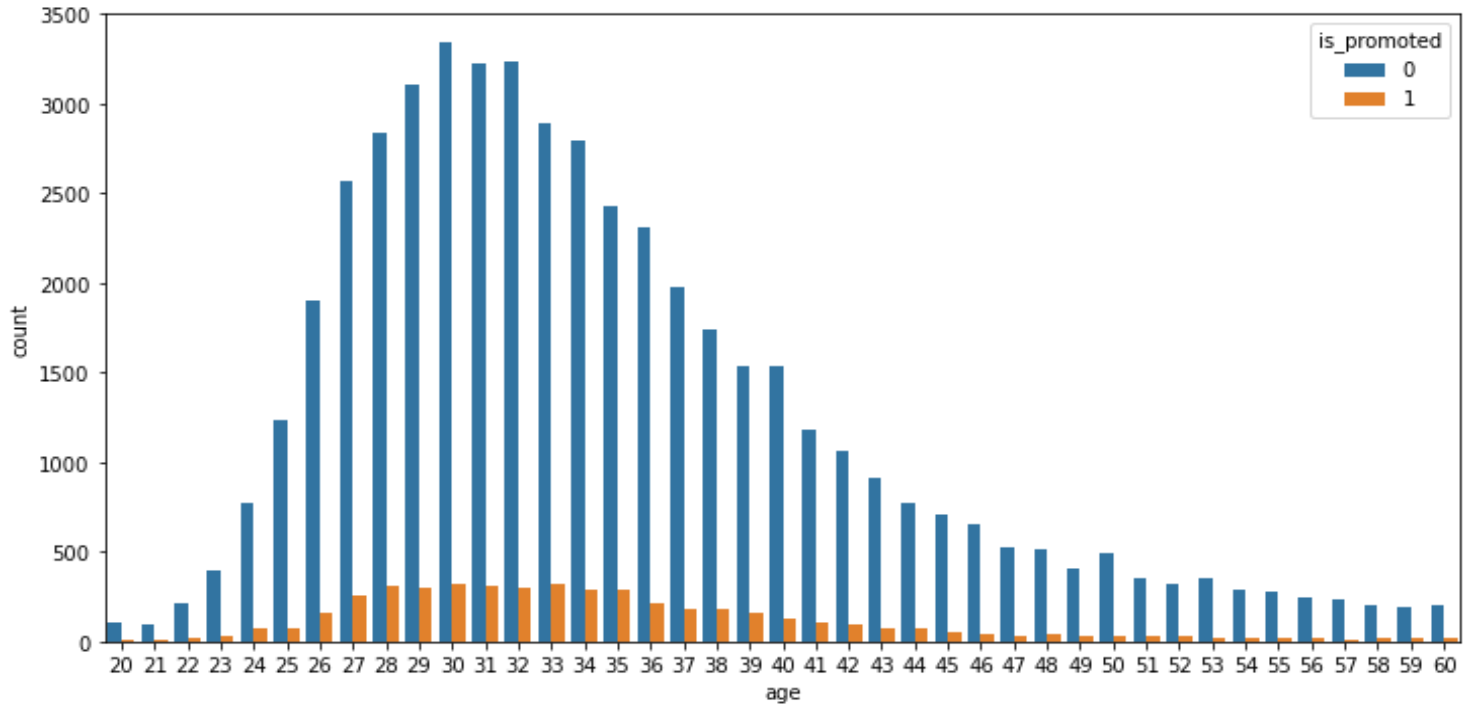
kaggle.com

EDA

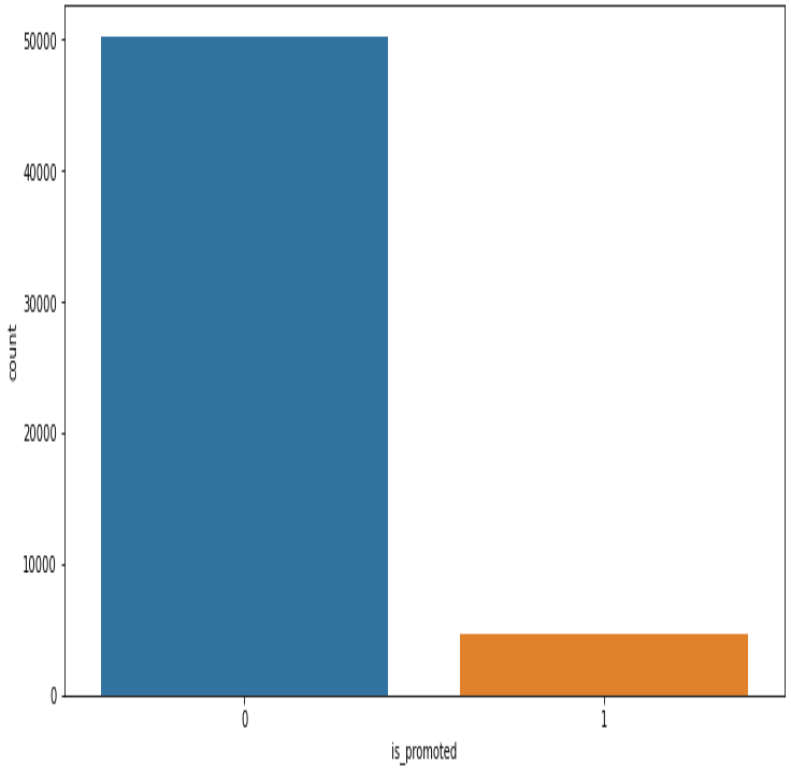

is_promoted
0
1
count
gender
f
m

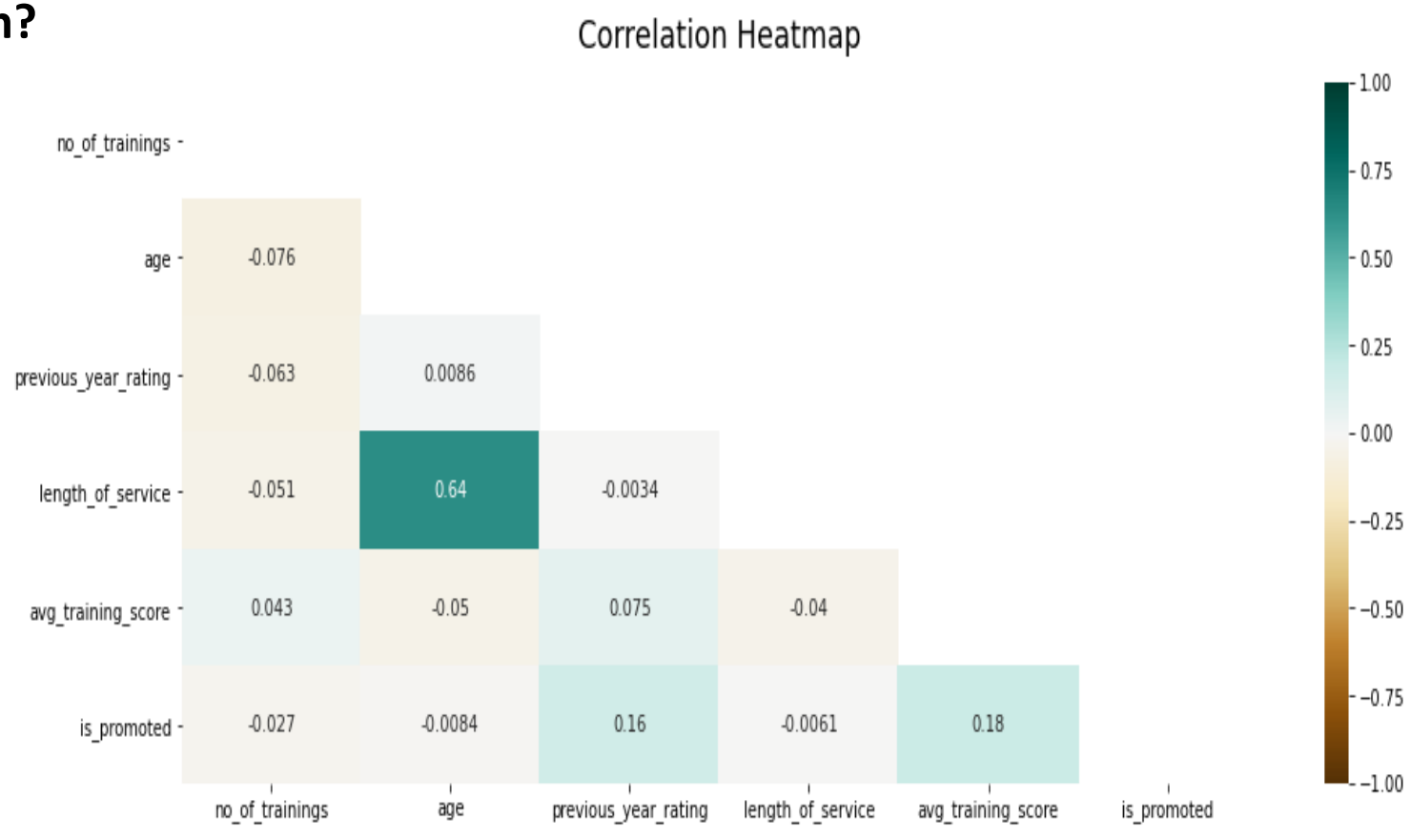

is_promoted
0
1
count
recruitment_channel
other
referred
sourcing


education = Bachelor's
education = Below Secondary
education = Master's & above
f
m
is_promoted


is_promoted
0
1
count
age

OOPS !! Does that mean we have a problem?
It seems we have an imbalanced data

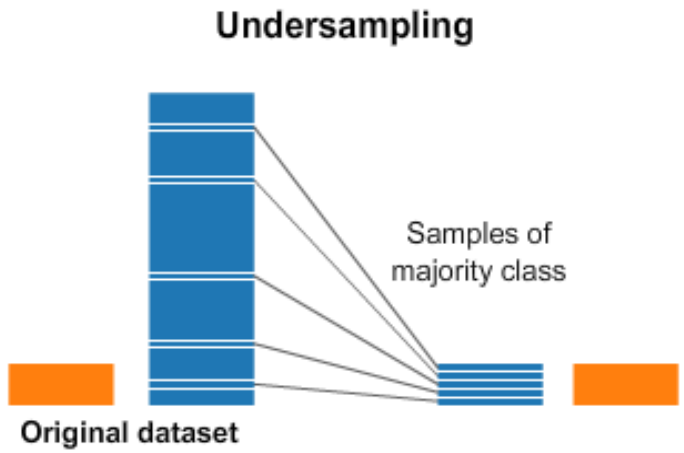


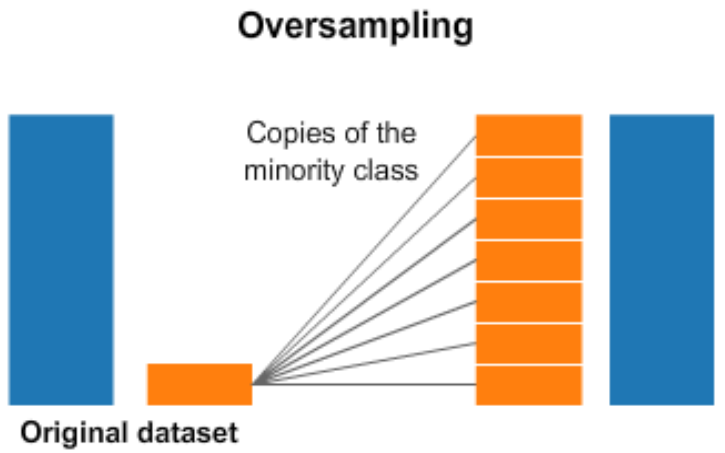


Imbalanced Dataset Issue

- Many machine learning algorithms don't perform well with an imbalance dataset.
- A prominent sign of imbalance is a high Accuracy while suffering a low Precision-Recall combination.
- Many techniques can be applied in order overcome this problem, one is adopting a Stratified K-Fold cross validation while sampling.
- More advanced methods will be: - Undersampling the majority
  - Oversampling the minority
  - SMOTE **(Synthetic Minority Oversampling Technique**)

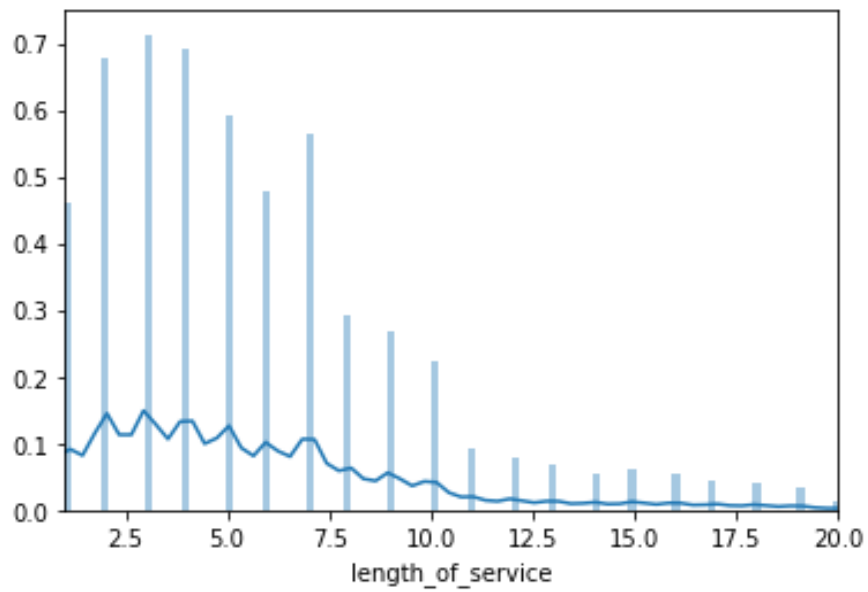
- **OR** a **combination of SMOTE followed by undersampling**, which is what this experiment adopted after multiple trials.

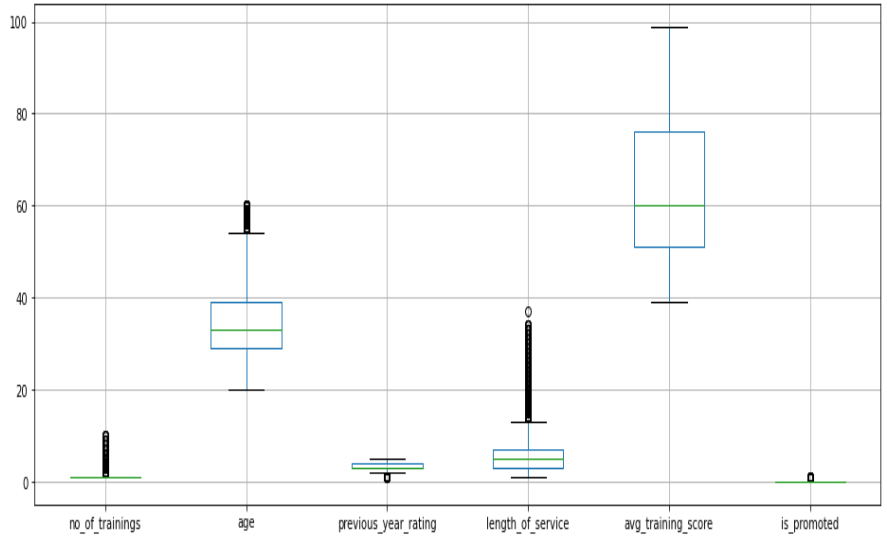




Data Cleaning

```
region_2     12343
region_22     6428
region_7      4843
region_15     2808
region_13     2648
region_26     2260
region_31     1935
region_4      1703
region_27     1659
region_16     1465
region_28     1318
region_11     1315
region_23     1175
region_29      994
region_32      945
region_19      874
region_20      850
region_14      827
region_25      819
region_17      796
region_5       766
region_6       690
region_30      657
region_8       655
region_10      648
region_1       610
region_24      508
region_12      500
region_9       420
region_21      411
region_3       346
region_34      292
region_33      269
region_18       31
Name: region, dtype: int64
```





Data Preprocessing & Feature Engineering

First things First! No DATA LEAKAGE allowed in here!

Split the data into 80/20 training and testing subsets

Outlier Removal

Numerical and Categorical Imputation

Categorical Data Conversion

Data Normalization

Combining the preprocessing steps into on wholistic Pipeline using the ColumnTransformer() function

Reasoning behind evaluation metrics selection

- Choosing an appropriate metric is challenging as it should take into consideration what result we seek from the experiment.
- Reporting classification accuracy for any imbalanced classification problem could be seriously misleading.
- We are interested in predicting **False Positives** and the **Precision** metric.
- We used the Precision-Recall curve to compare the models' performances instead ROC curve.

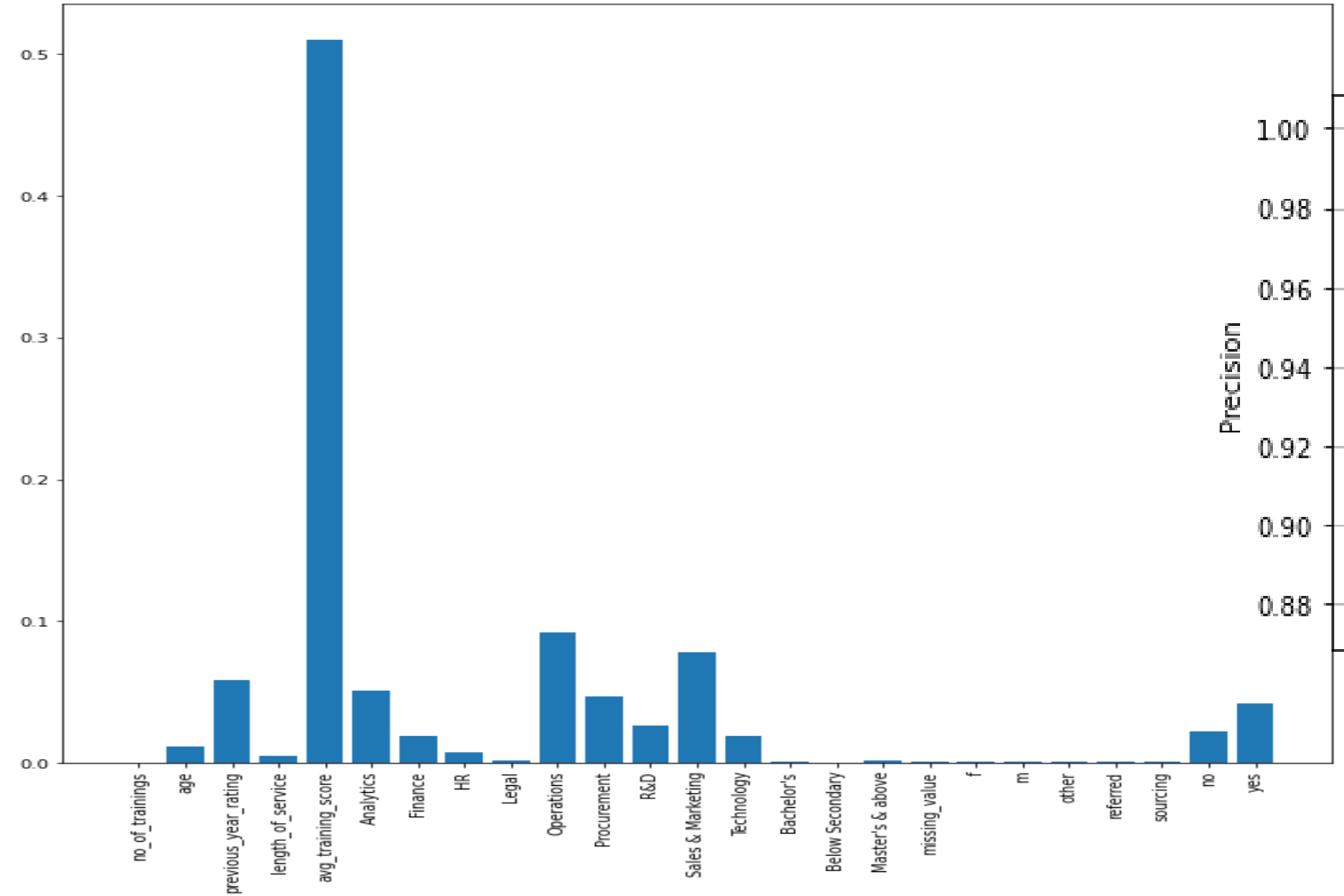
Data Modeling

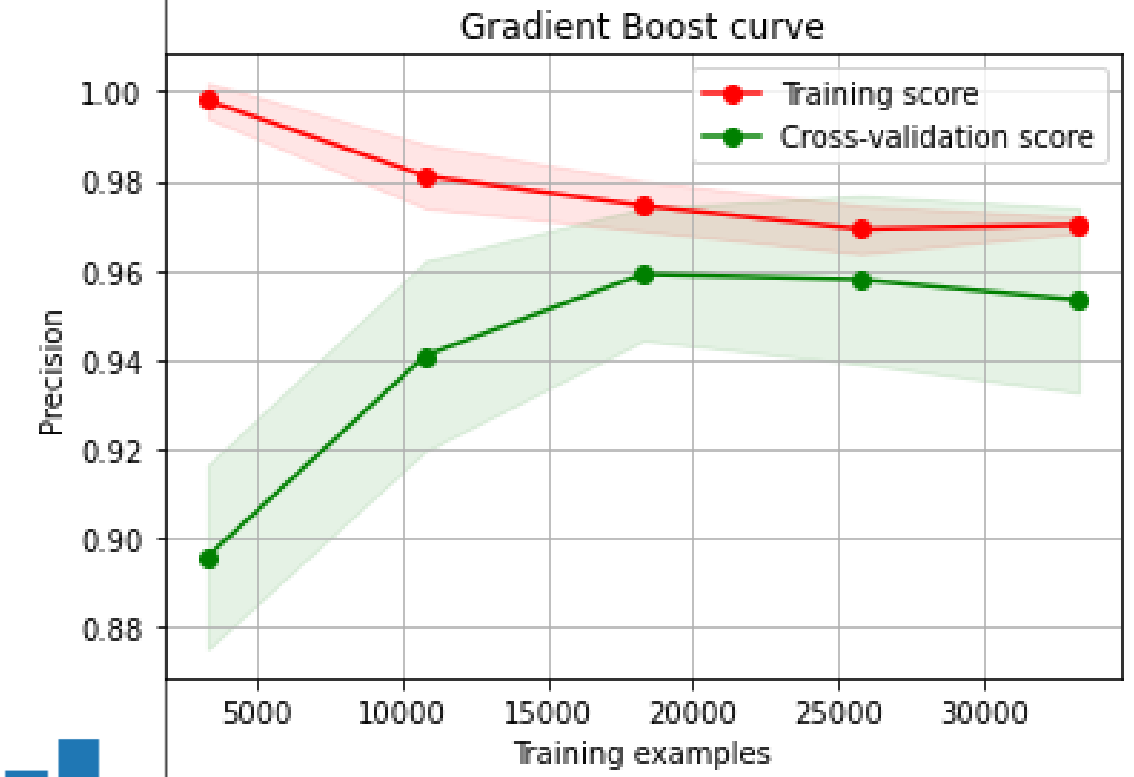
And Don’t Forget to APPLY Stratified Cross-Validation

- 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)

	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

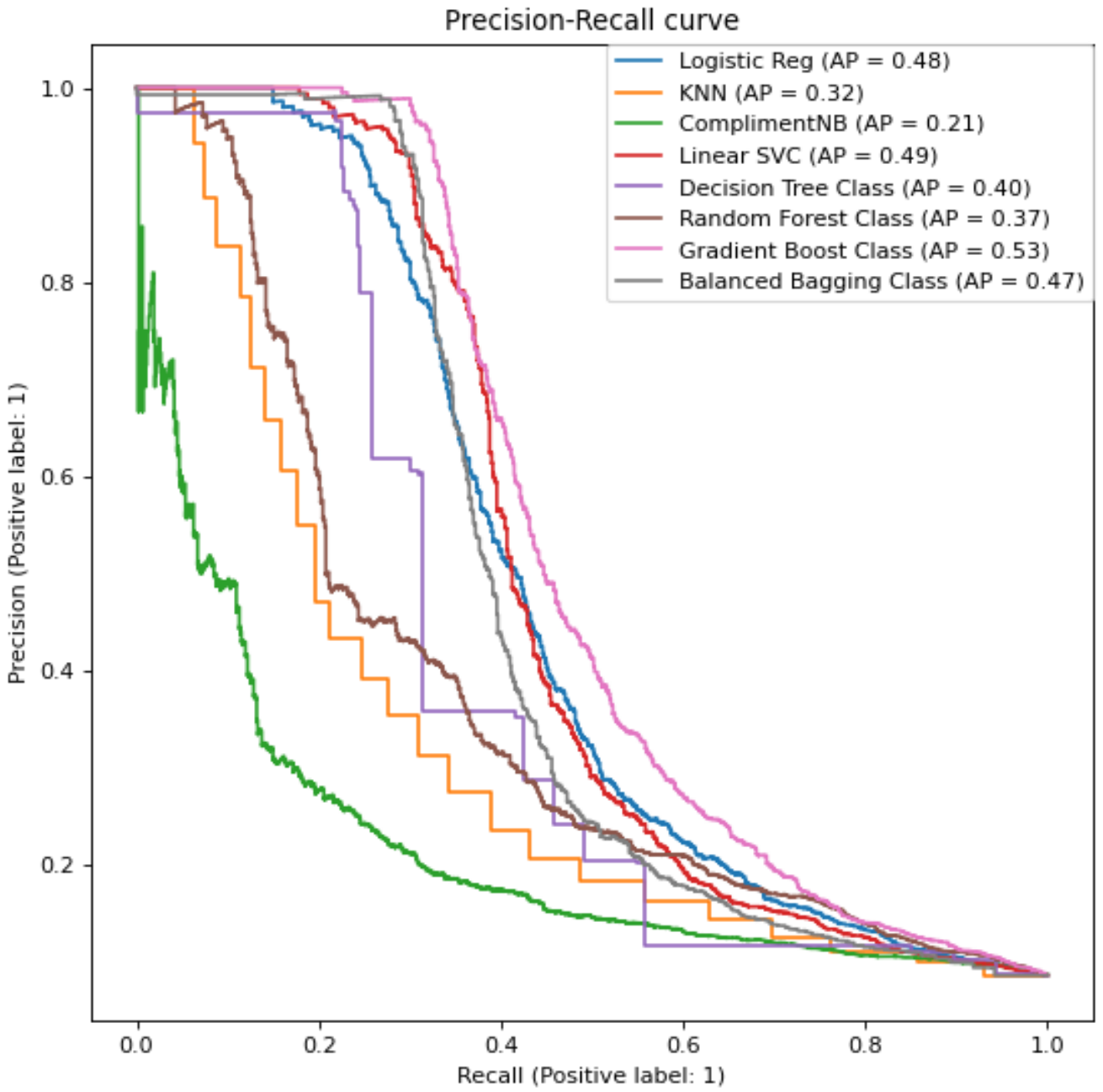
Training and Cross-Validation Results: Best Model





Testing Results

	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

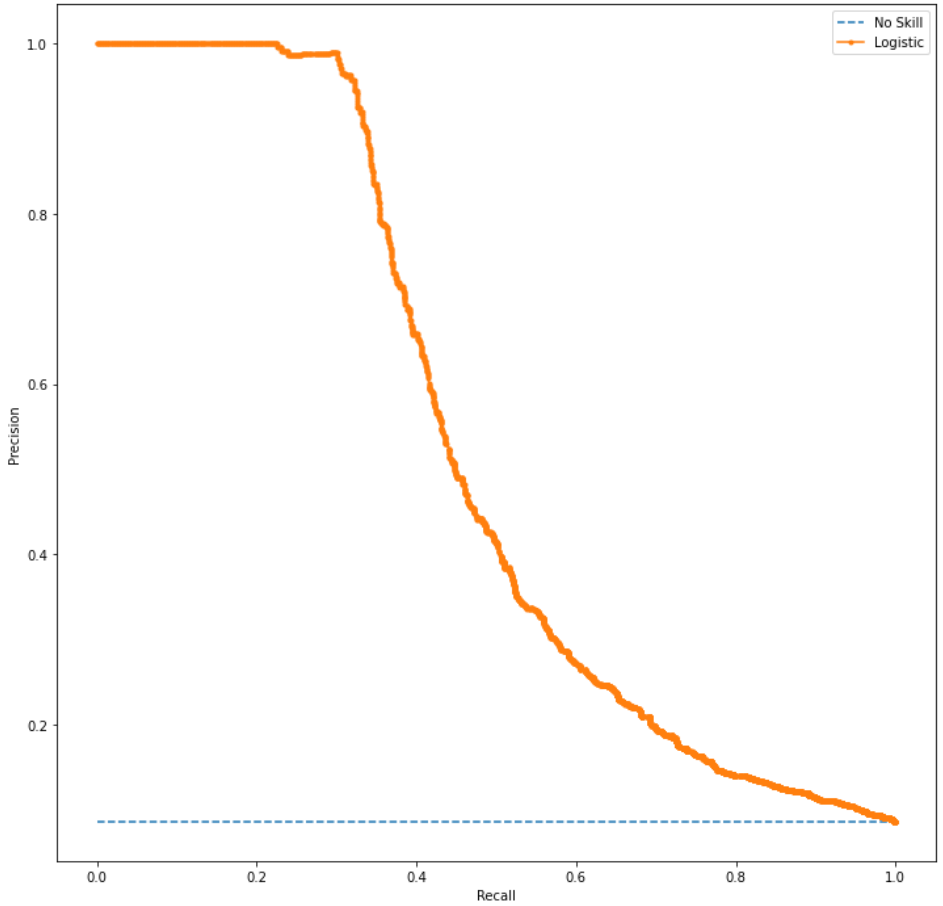


Conclusion and recommendations

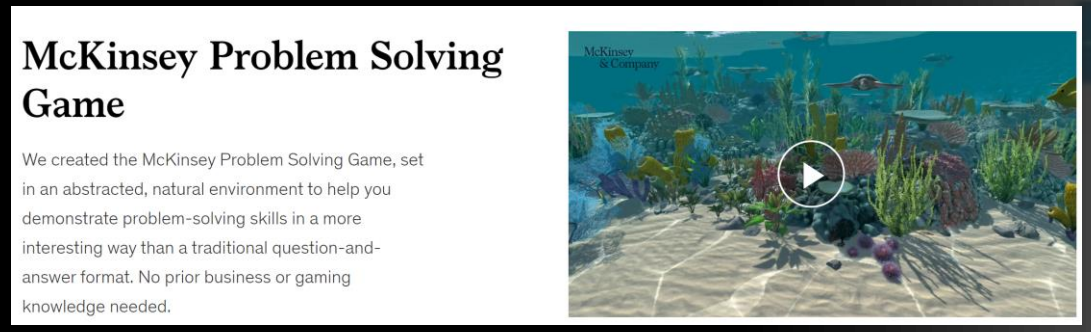
- The chosen model is the **Gradient Boosting Classifier** displayed the highest Precision-Recall combination, **92%** for **precision** and 35% for recall. We have chosen to emphasize on the precision (i.e. True Positives and False Positives), avoiding financial burdens.

- **Recommendations**:
- Exploring other relevant features such as: job level, monthly income, overtime, and business travels etc.
- Getting more insights into the business activities and market capitalization (size) of each firm under investigation.
- Segmenting the market firms beforehand using ML techniques (for instance KNN or K-Means-Clustering), and applying the same framework separately on similar firms from the same cluster for better generalization of outcomes and credibility.

Outperforming previous studies:
- Specifying a set of assumptions and criterions based on which the experiment Is valid
- Tackling the issue of imbalance



Future outlook of HR Analytics



Thank You !!!

Questions?