Black_Raven (James Ng) Follow
Nov 27, 2019 · 6 min read ★ · Listen

Save

Feature Engineering, Over-Sampling, Predictive Machine Learning, Supervised Learning, Attrition Prediction

The key to success in an organisation is the ability to attract and retain top talents. It is vital for the Human Resource (HR) Department to identify the factors that keep employees and those which prompt them to leave. Organisations could do more to prevent the loss of good people.

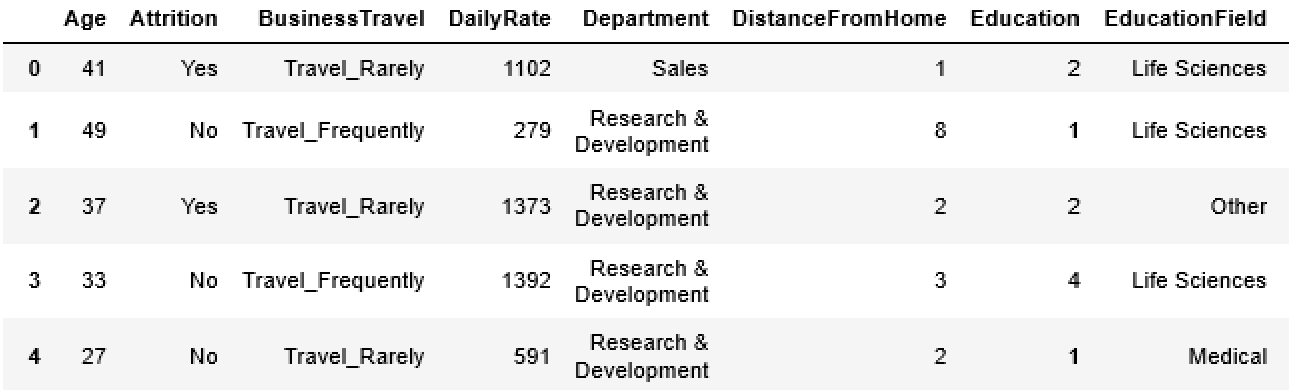
Which key factors result in employee attrition?

This project is based on a hypothetical dataset downloaded from IBM HR Analytics Employee Attrition & Performance. It has 1,470 data points (rows) and 35 features (columns) describing each employee's background and characteristics; and labelled (supervised learning) with whether they are still in the company or whether they have gone to work somewhere else. Machine Learning models can help to understand and determine how these factors relate to workforce attrition.

Exploratory Data Analysis (EDA)

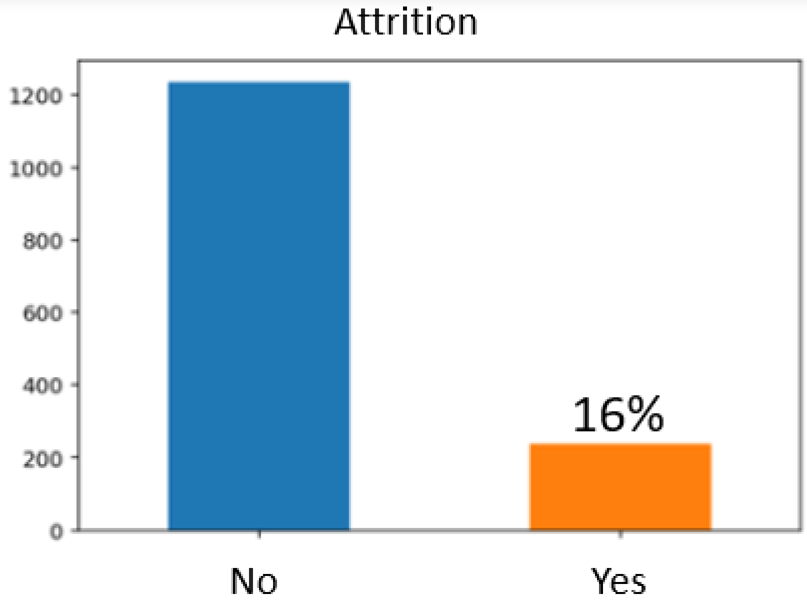
Let's import the relevant Python libraries, and read in the data file. Jupyter notebook with Python codes here.

```
df = pd.read_csv('WA_Fn-UseC_-HR-Employee-Attrition.csv')
df.head()
```

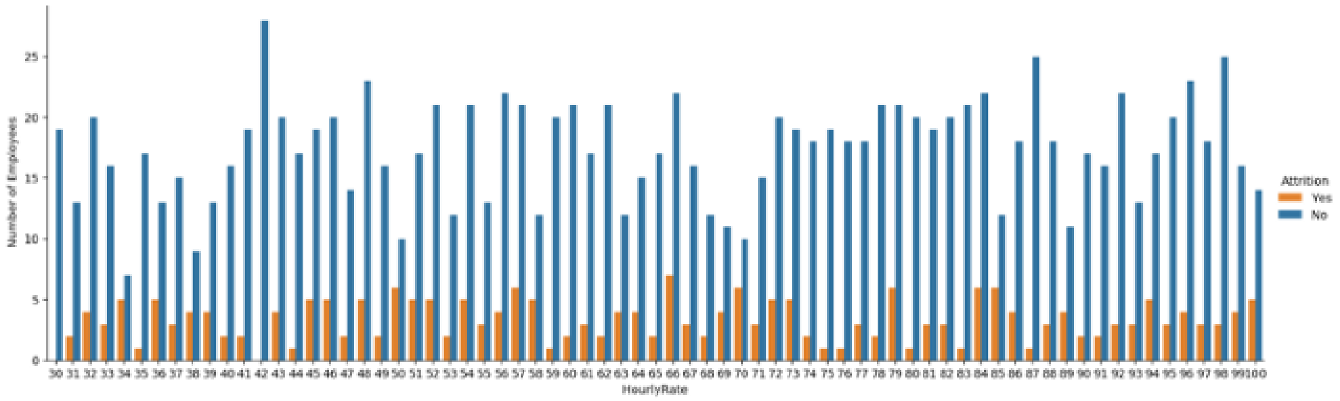


	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical

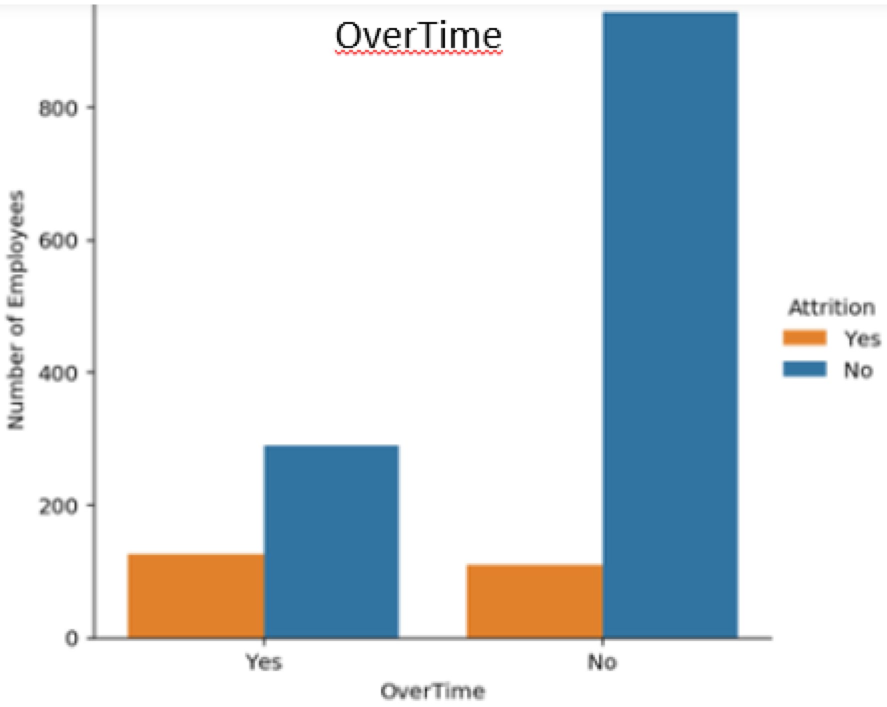
The dataset is well organised with no missing values. Target class is imbalance, with attrition rate of 16%.



Are employees leaving because they are poorly paid? Employees are paid an hourly rate of $30 to $100, and attrition seems to happen at every level regardless of employee hourly rate. This can be confirmed later at feature importance.

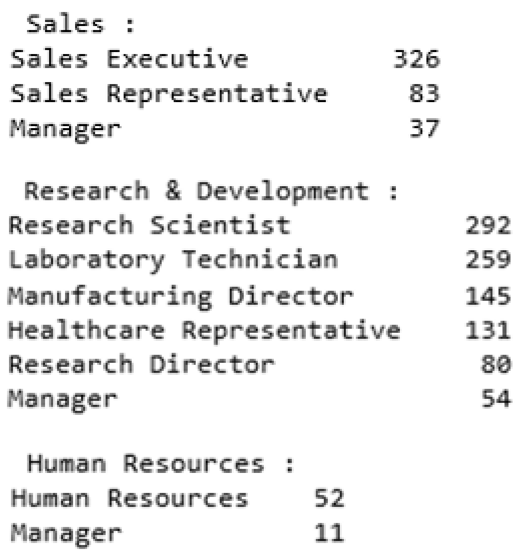


Overtime seems to be one of the key factors to attrition, as a larger proportion of overtime employees has departed.



There are only 3 departments included for this analysis.

```
for dpmt in df['Department'].unique():
    print('\n', dpmt, ':')
    print(df[df['Department']==dpmt]['JobRole'].value_counts())
```



```
 Sales :
Sales Executive          326
Sales Representative      83
Manager                   37

 Research & Development :
Research Scientist           292
Laboratory Technician        259
Manufacturing Director       145
Healthcare Representative    131
Research Director             80
Manager                       54

 Human Resources :
Human Resources    52
Manager            11
```

Data Preprocessing

Data has to be preprocessed as machine learning models are better at reading numbers

```
for col in df.select_dtypes(['object']).columns:
    print(col, ':', sorted(df[col].unique()))
```

Baseline Model Performance

Next I split the data into 80:20 ratio. Do specify parameter "stratify=y" in the code so that the proportion of classes in the output dataset will have the same proportion as the stratify column provided.

> *Tip: as target y has binary categorical classes with 84% '0's and 16% '1's, "stratify=y" will make sure that the 80:20 split has 84% of '0's and 16% of '1's in both output datasets*

As the dataset is imbalance, use "StratifiedKFold" in cross validation when training the models, and each baseline model performance can be tabulated.

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=.2, random_state=SEED, stratify=y)

kf = StratifiedKFold(n_splits=5, shuffle=True, random_state=SEED)
```

The baseline model performance results are terrible, with F1-scores ranging from 20% to 40% for most models. After tuning hyperparameters and the threshold, the Logistic Regression has achieved F1-score 50.5% and **Recall 48.9%.**

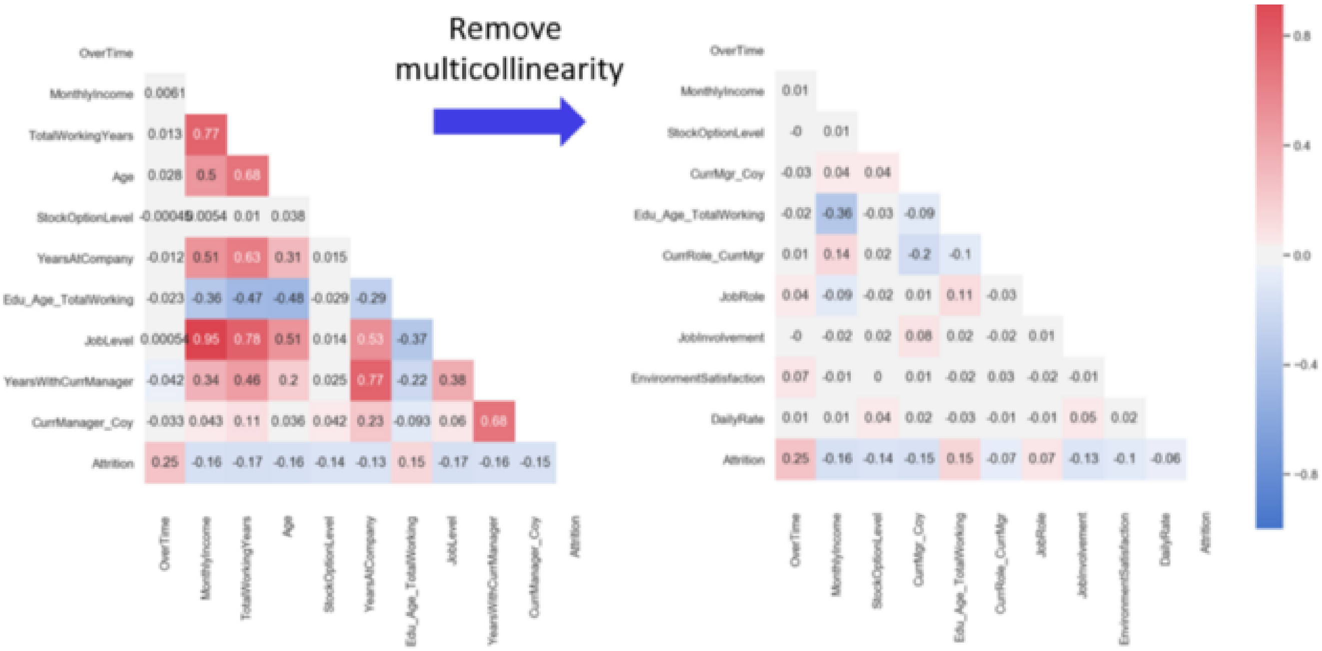
> *Tip: Do NOT use Accuracy as a performance metric for this imbalance class scenario. If a model naively classify all data points to class '0', then this useless model would have scored 84% for Accuracy!*

Feature Selection

Below features are discarded as they do not have any useful information: 'Over18', 'EmployeeCount', 'EmployeeNumber', 'StandardHours'.

Also, if features are closely related to one another (multicollinearity), one of them has to be removed to prevent misleading results to linear models such as Logistic Regression. Although tree-based models are not directed affected, they could also lead

```
df_top10 = df.loc[:,top10_features.index].join(df['Attrition'])
sns.set(style="white")
mask = np.zeros_like(df_top10.corr(), dtype=np.bool)
mask[np.triu_indices_from(mask)] = True
fig, ax = plt.subplots(figsize=(9,8))
cmap = sns.diverging_palette(255, 10, as_cmap=True)
sns.heatmap(df_top10.corr().round(2), mask=mask, annot=True,
            cmap=cmap , vmin=-1, vmax=1, ax=ax)
```



Correlation heatmap before and after removing multicollinearity

Feature Engineering

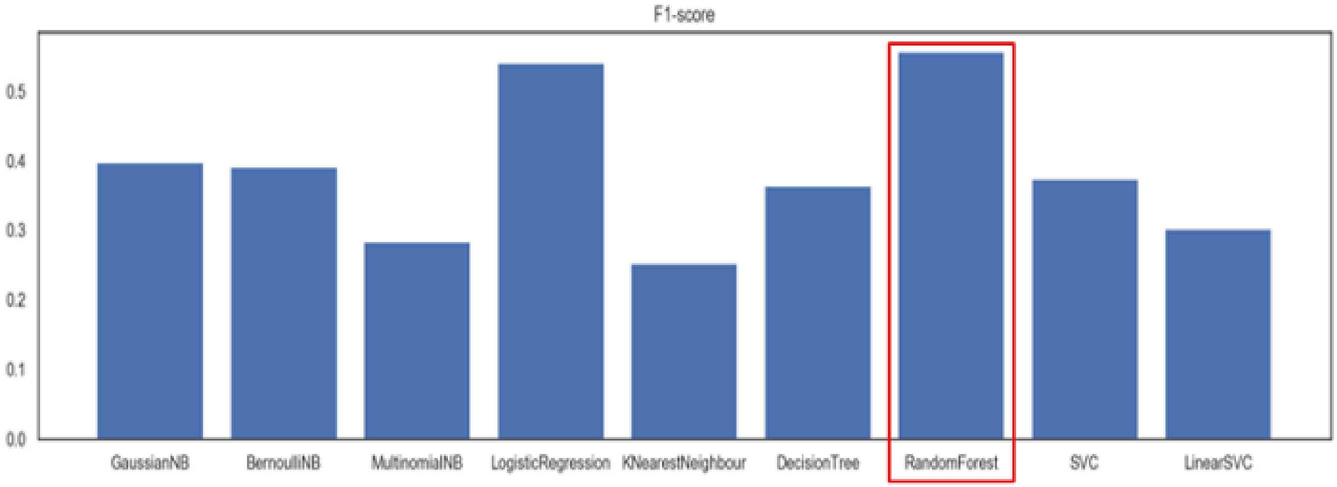
A new feature 'EduField_Dept' is created which defines whether 'JobRole' is related to 'EducationField', with the labels '0' = not related, '1' = related, '2' = somewhat related. However, HR seems to have done fabulously in job matching as this feature did not emerge useful.

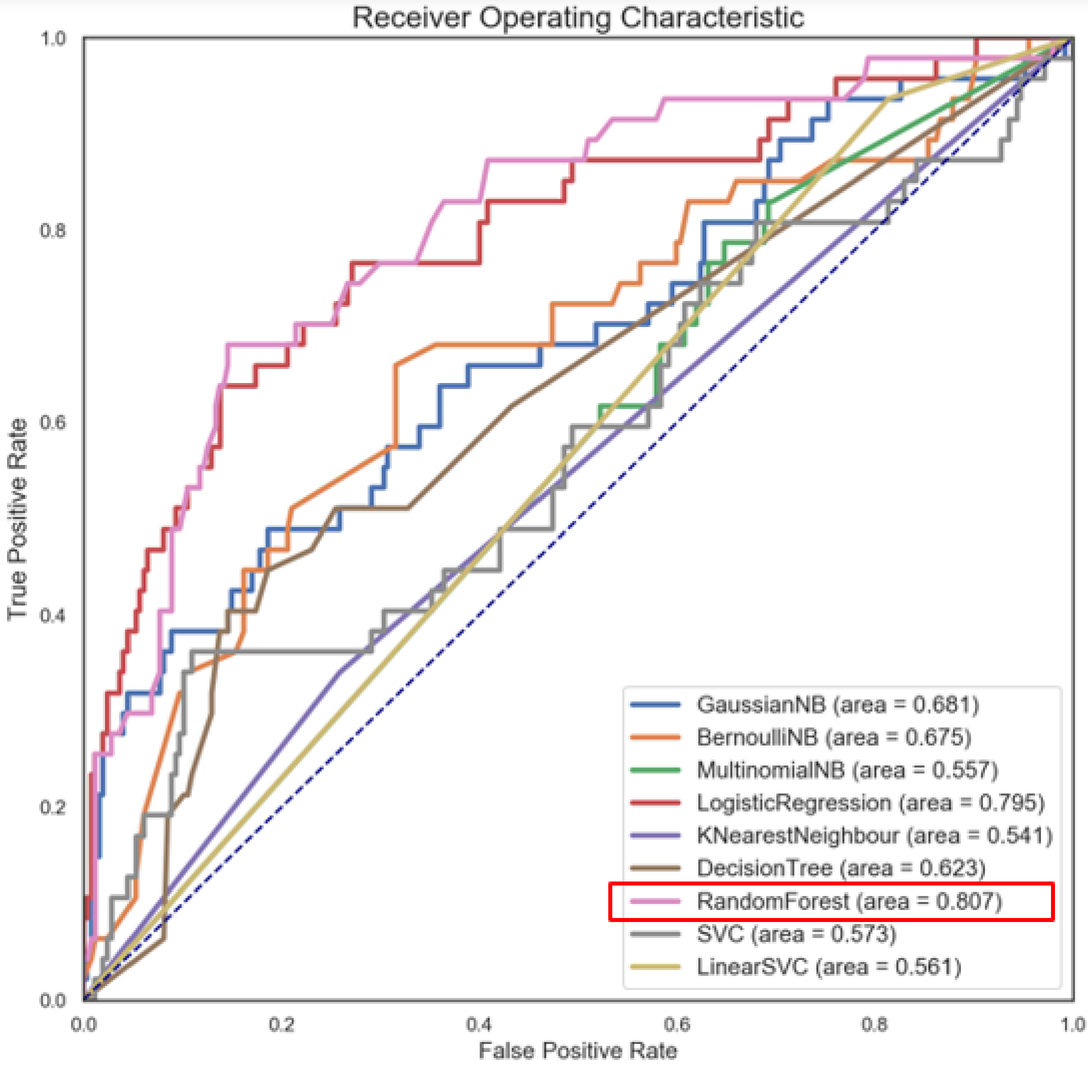
Another set of new features are also created based on my own people observation and intuition:
· Job_Coy = JobLevel / (YearsAtCompany + 1)
· Edu_Age_TotalWorking = Education / (Age + TotalWorkingYears)
· CurrMgr_Coy = YearsWithCurrManager / (YearsAtCompany + 1)
· CurrRole_CurrMgr = YearsInCurrentRole / (YearsWithCurrManager + 1)

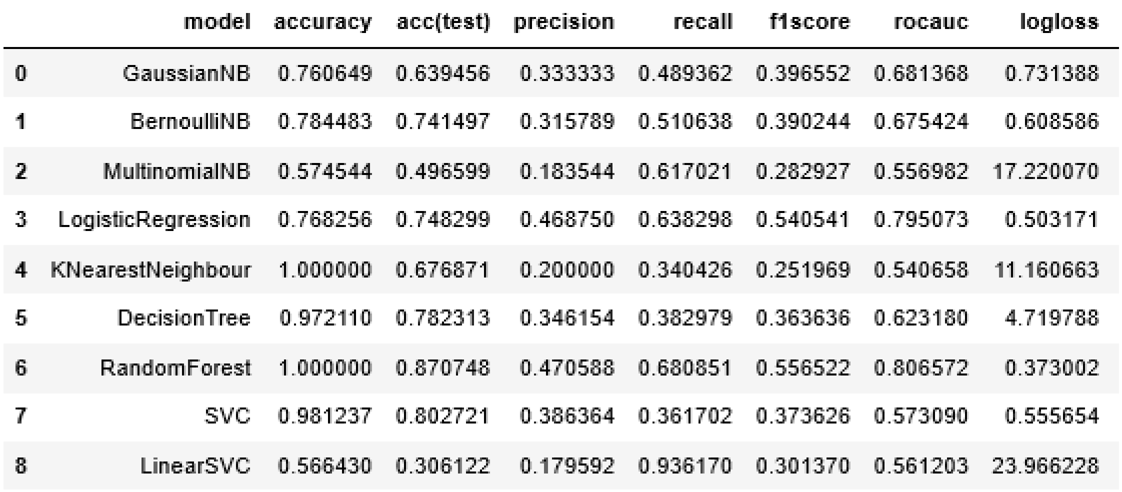
With these new features, let's run the model performance codes again including tuning hyperparameters (look for a long bunch of code). This time, Random Forest Classifier emerged with F1-score 49.6% and **Recall 61.7%**.

To Handle Imbalance Data

This two-class dataset is imbalanced (84% vs 16%). As a result, there is a possibility that the model built might be biased towards to the majority and over-represented class. After applying Synthetic Minority Oversampling Technique (SMOTE) to over-sample the minority class, some improvement in both F1-score & Recall can be observed.

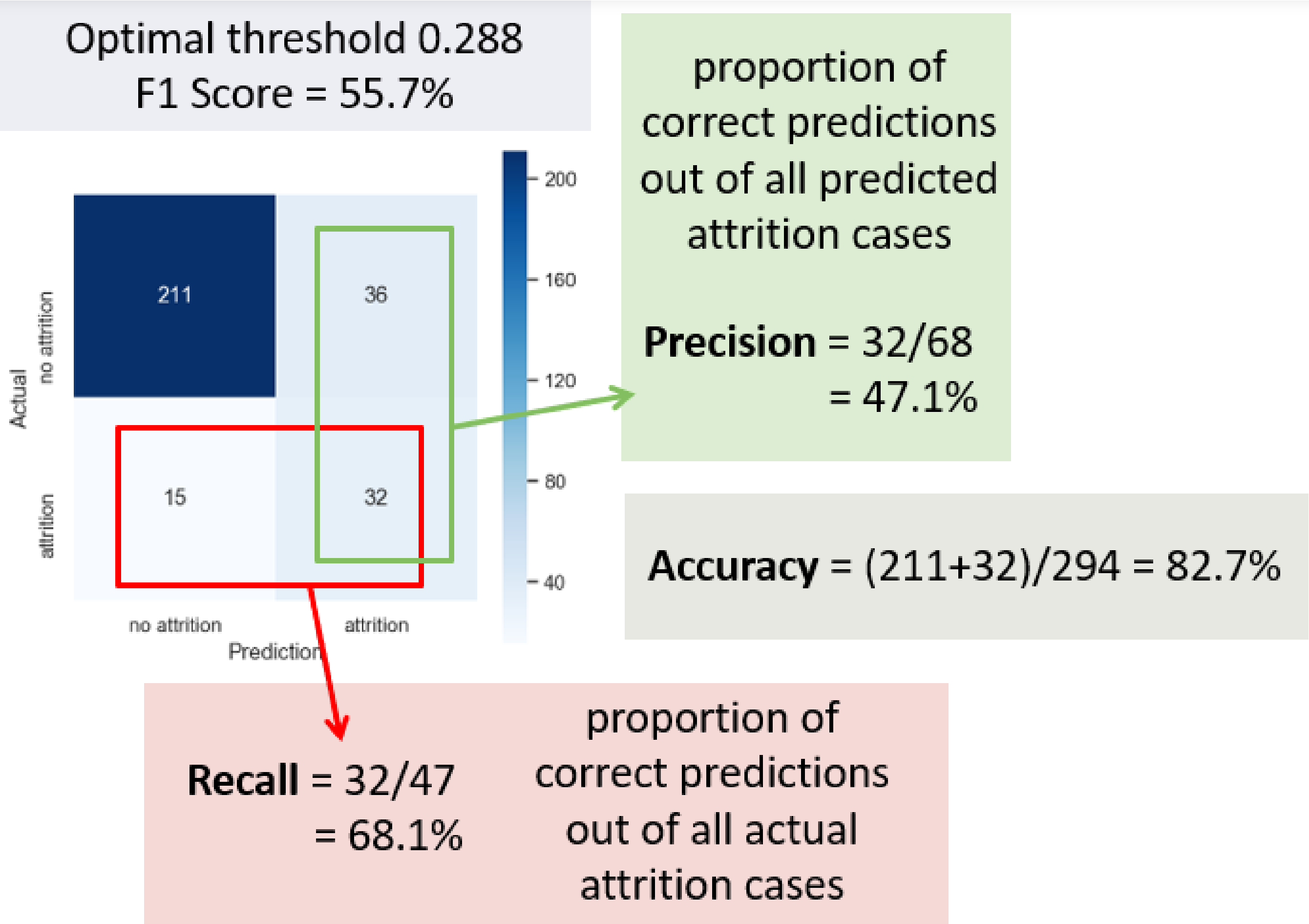






	model	accuracy	acc(test)	precision	recall	f1score	rocauc	logloss
0	GaussianNB	0.760649	0.639456	0.333333	0.489362	0.396552	0.681368	0.731388
1	BernoulliNB	0.784483	0.741497	0.315789	0.510638	0.390244	0.675424	0.608586
2	MultinomialNB	0.574544	0.496599	0.183544	0.617021	0.282927	0.556982	17.220070
3	LogisticRegression	0.768256	0.748299	0.468750	0.638298	0.540541	0.795073	0.503171
4	KNearestNeighbour	1.000000	0.676871	0.200000	0.340426	0.251969	0.540658	11.160663
5	DecisionTree	0.972110	0.782313	0.346154	0.382979	0.363636	0.623180	4.719788
6	RandomForest	1.000000	0.870748	0.470588	0.680851	0.556522	0.806572	0.373002
7	SVC	0.981237	0.802721	0.386364	0.361702	0.373626	0.573090	0.555654
8	LinearSVC	0.566430	0.306122	0.179592	0.936170	0.301370	0.561203	23.966228

Random Forest Classifier has emerged as the final winning model with F1-score 55.7% and highest **Recall 68.1%**. This could be the highest possible score achieved with the inherent limitations in the dataset.



Optimal threshold 0.288
F1 Score = 55.7%

proportion of correct predictions out of all predicted attrition cases

Precision = 32/68
= 47.1%

Accuracy = (211+32)/294 = 82.7%

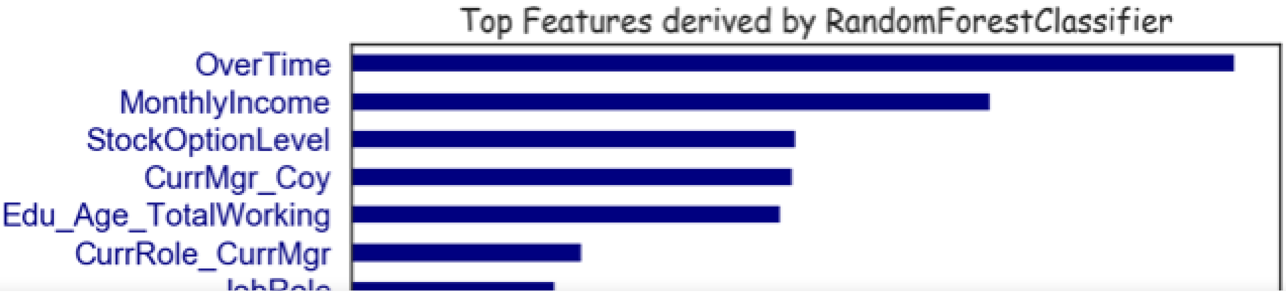
Recall = 32/47
= 68.1%

proportion of correct predictions out of all actual attrition cases

Final model performance

Key Findings (based on hypothetical dataset)

The top factor for employee attrition in this hypothetical organisation seems to be **monetary**, as 'OverTime' and 'MonthlyIncome' emerged at the top. This could be due to a bad compensation process or causing a poor work-life balance. The next important factor seems to be **personal relationships** with fellow workers, where current manager and job role could be the main contributing reasons for attrition. Finally, **employee engagement** is a critical satisfaction factor, and the organisation should keep employees constantly involved and motivated.



Top 10 features discovered by the model

Path Forward

Machine learning models are as good as the data you feed it, and more data would strengthen the model. For example in this dataset, the feature 'PerformanceRating' has been restricted to scores of 3 and 4 only. More insights could be generated if the full spectrum of performance ratings are included. In the real life situation, getting the right data is often more challenging than the analytics itself.

Conclusion

HR Analytics is gaining traction in organisations that embrace digital transformation. The scope has expanded from analytics of employee work performance to providing insights so that decisive improvements can be made to organisational processes. While some level of attrition is inevitable, it should be kept at the minimal possible level.

Python codes with inline comments are available on my GitHub, do feel free to refer to them. And the presentation file is here.

JNYH/employee_attrition

The key to success in an organisation is the ability to attract and retain top talents. It is vital for the Human…

github.com

Thank you for reading!
I hope you have enjoyed reading my stories. To get unlimited access to quality content on Medium, join as a Medium member!