HR Attrition

Green Team

Ridhi Likhi, Hojin Lee, Candice Lee,
Atabay Kadiroglu, Melis Ocal

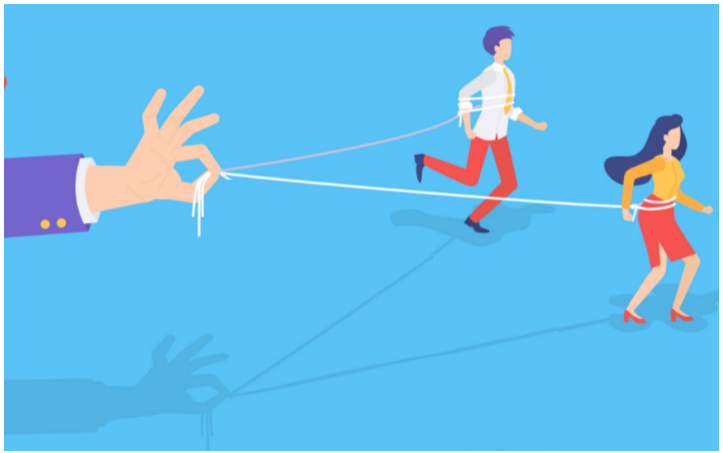
Outline

1. Business Problem & Goal
2. Exploratory Data Analysis
3. Baseline Model Comparisons
4. Feature Engineering
5. Parameter Optimization
6. Model building
7. Conclusion & Suggestion

Business Problem

A major problem in high employee attrition is **its cost to an organization.**

- Recruitment costs
- Loss of knowledge base
- Training costs
- Cultural Impact
- Lost productivity
- Potential customer dissatisfaction/Reduced or lost business
- Disruption of team dynamics



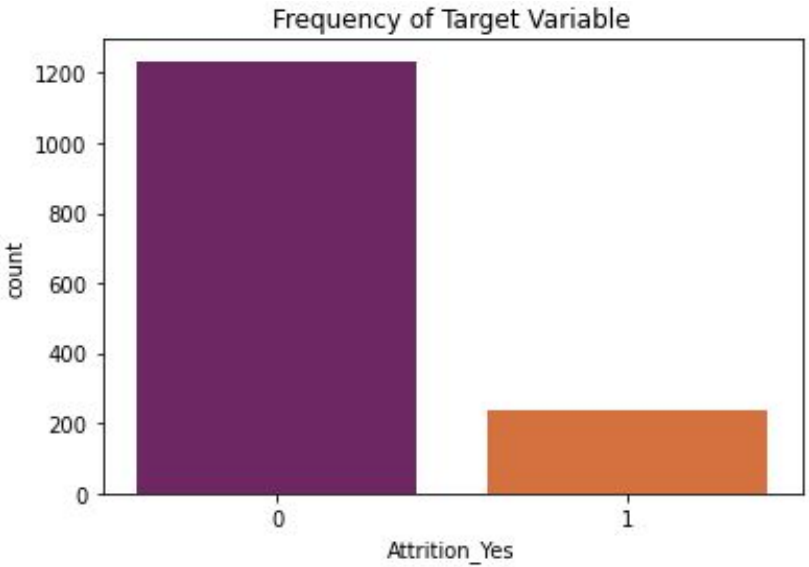
Goal

- Identify **significant factors** leading to attrition.
- Develop a **predictive** model.
- Developing strategies to **reduce turnover rates** ensuring **high return on investment** for the organization.

Exploratory Data Analysis

HR Attrition DataSet :

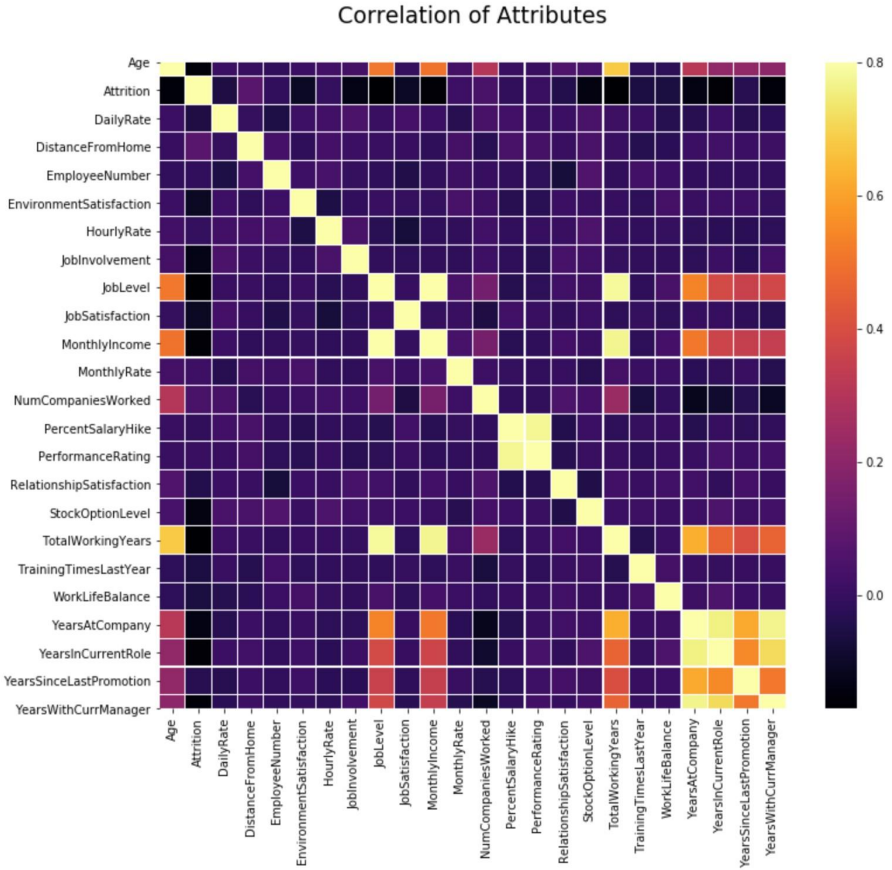
- **Instances:** 1470
- **Attributes:** 35
- **Target variable:** Attrition
  - **Attrition rate:** about 16%
- **Missing values:** N/A
- **Duplicates:** N/A
- **Outliers:** Very few



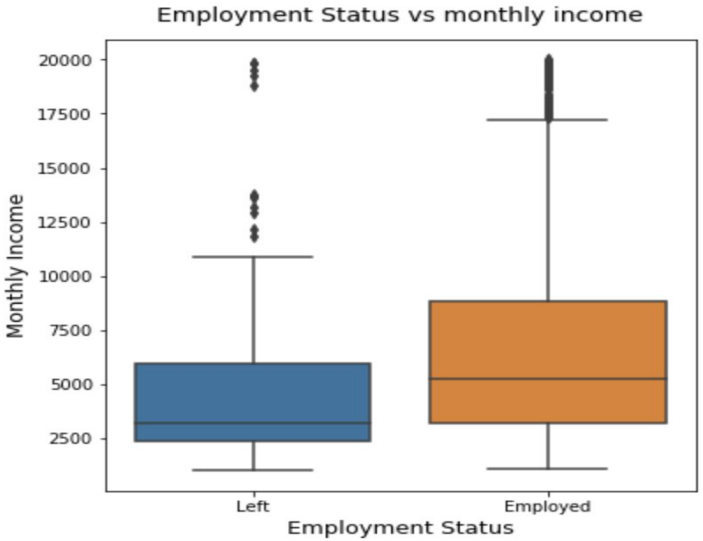
Exploratory Data Analysis

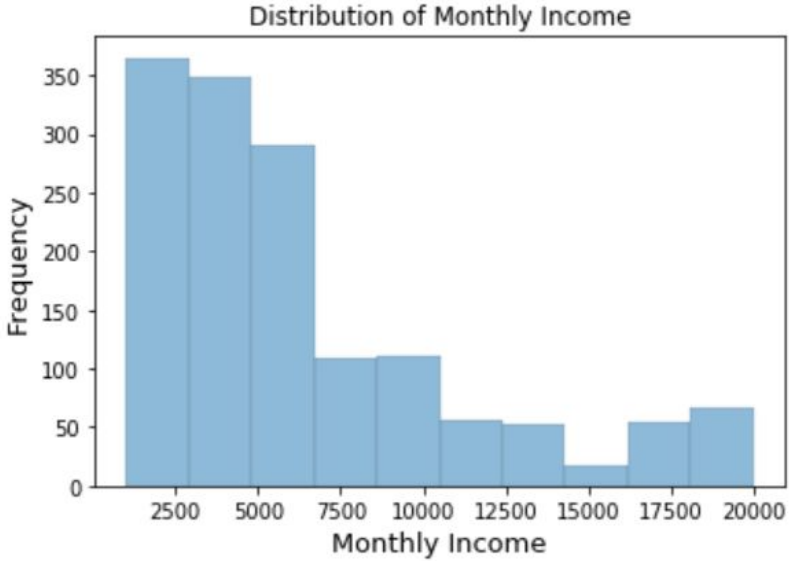
Highly Correlated Features:

- Job Level & Monthly Income
- Job Level & Total Working Years
- Monthly Income & Total Working Years
- Age & Total Working Years



Exploratory Data Analysis

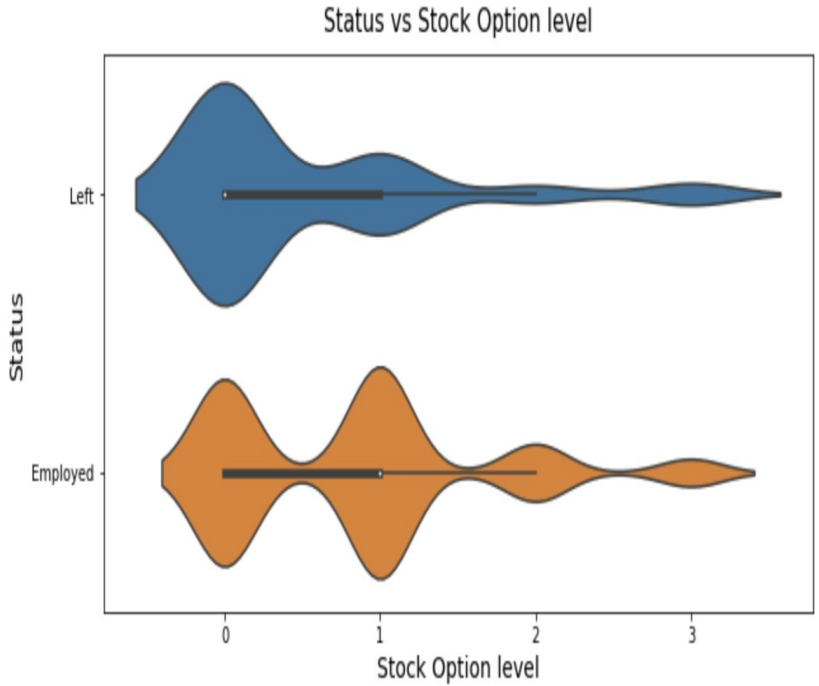


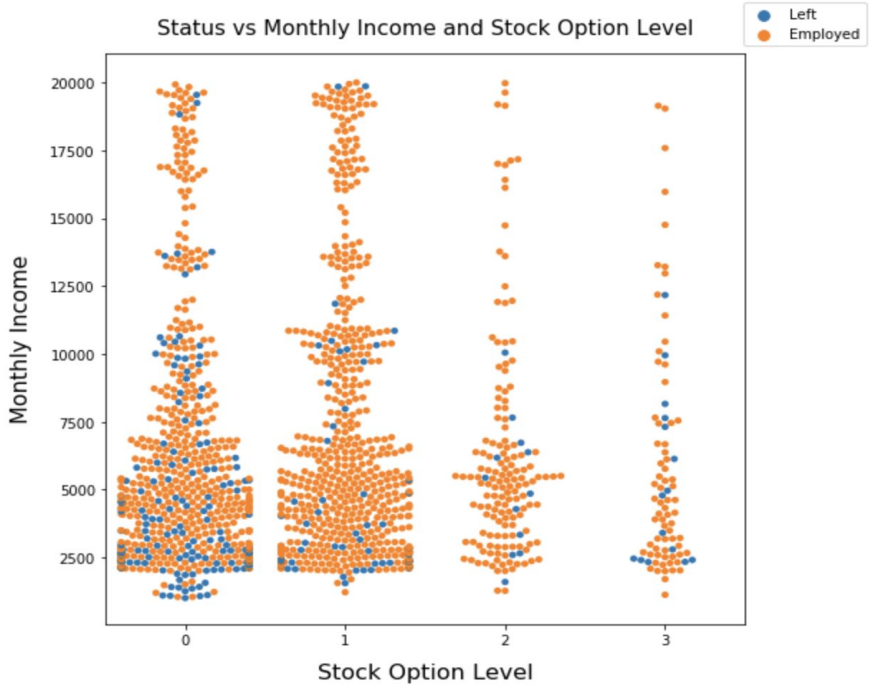


- The median monthly income is lower for employees left.

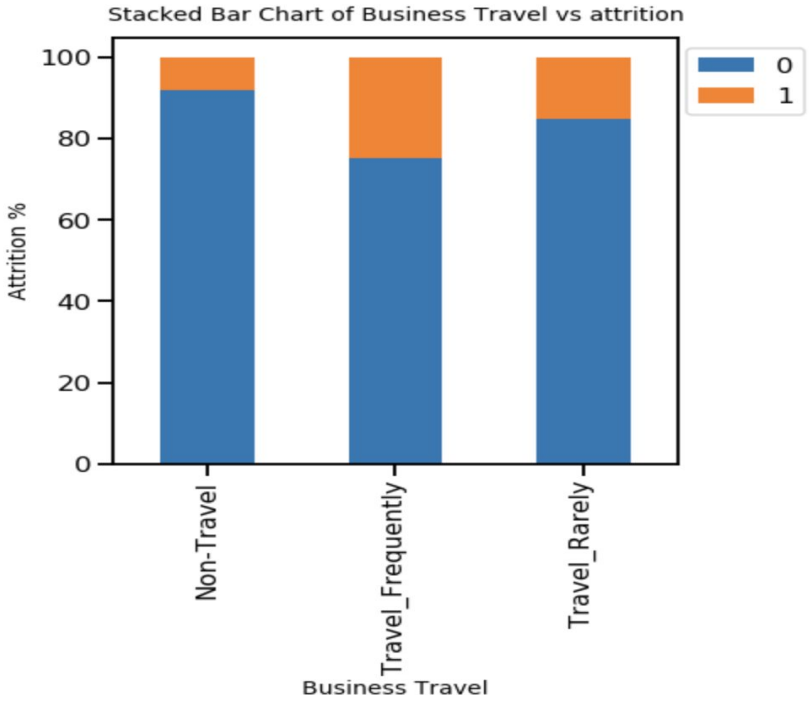
	Department	min	max	mean
0	Human Resources	1555	19717	6654.507937
1	Research & Development	1009	19999	6281.252862
2	Sales	1052	19847	6959.172646

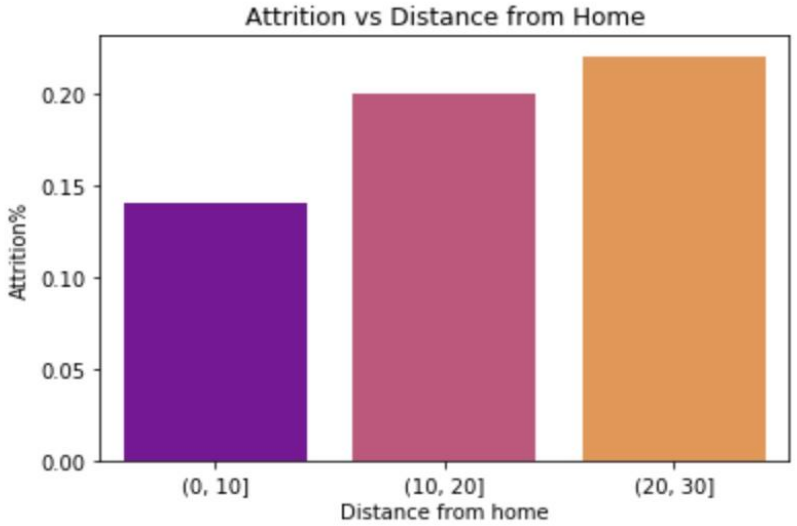
Exploratory Data Analysis



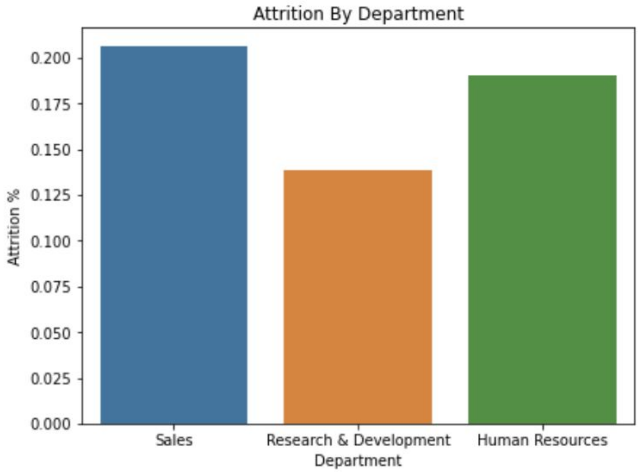


Exploratory Data Analysis

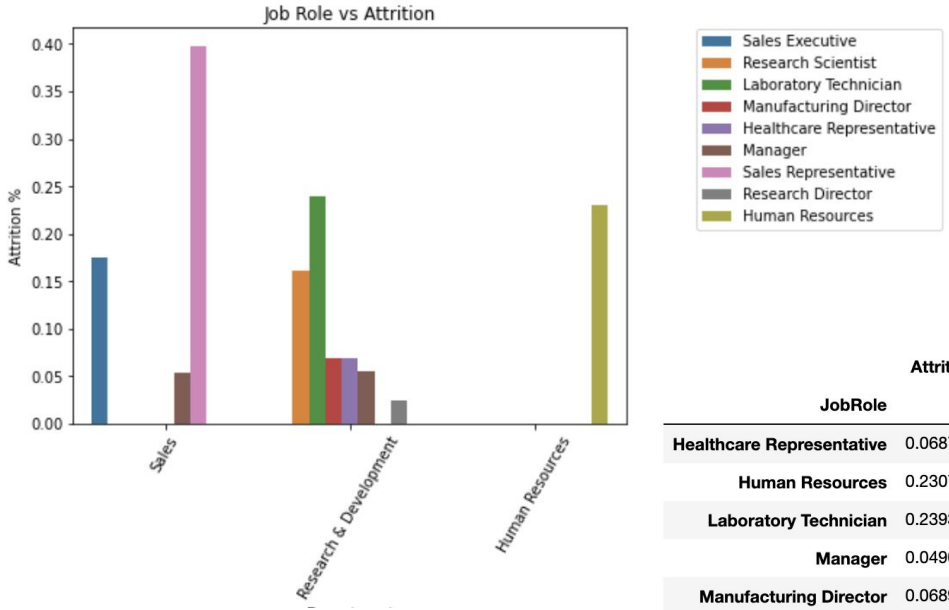




EDA Continued

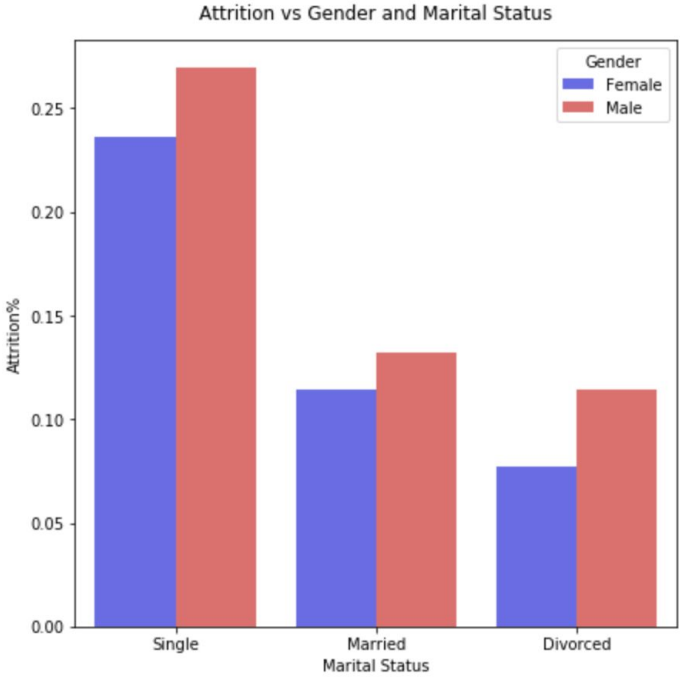


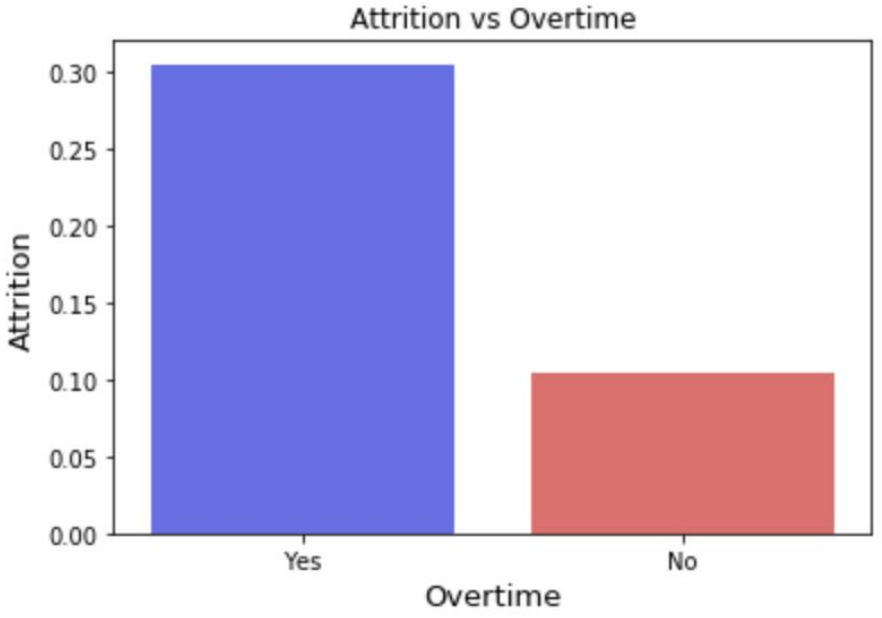
Department	Attrition
Human Resources	0.190476
Research & Development	0.138398
Sales	0.206278



JobRole	Attrition
Healthcare Representative	0.068702
Human Resources	0.230769
Laboratory Technician	0.239382
Manager	0.049020
Manufacturing Director	0.068966
Research Director	0.025000
Research Scientist	0.160959
Sales Executive	0.174847
Sales Representative	0.397590

EDA Continued



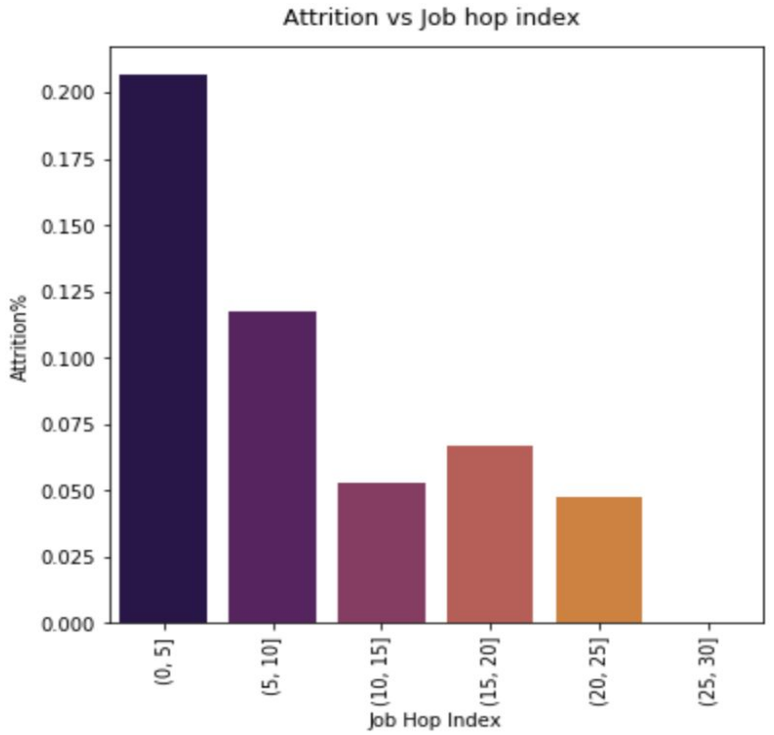


Feature Engineering

Job Hop Index:

- An index assigned to individuals to determine whether or not they have been ‘job-hopping’ frequently.

$$\text{Job Hop Index} = \frac{\text{Total experience}}{\text{Number of companies worked}}$$



Comparison of Models

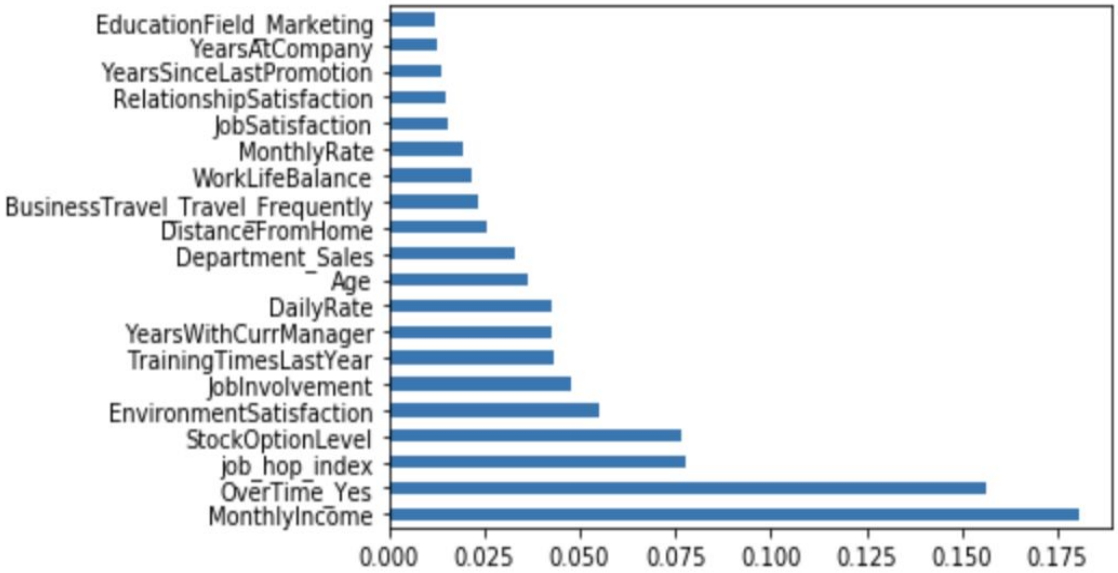
	Classifier	Precision	Recall	F1 score	Accuracy	AUC
0	Decision Tree Classifier	0.54	0.34	0.42	0.83	0.69
1	K-Nearest Neighbor	0.29	0.06	0.09	0.80	0.57
2	Naive Bayes	0.43	0.64	0.51	0.79	0.78
3	Random Forest	0.77	0.13	0.22	0.84	0.82
4	Gradient Boosting Classifier	0.54	0.43	0.48	0.84	0.81

We have an imbalanced dataset. Therefore, F1- Score and Area Under the Curve(AUC) will be the metrics for our model evaluation. We choose a classifier with the **highest F1-score and AUC.**

Gradient Boosting Classifier is the ideal choice.

Important Features

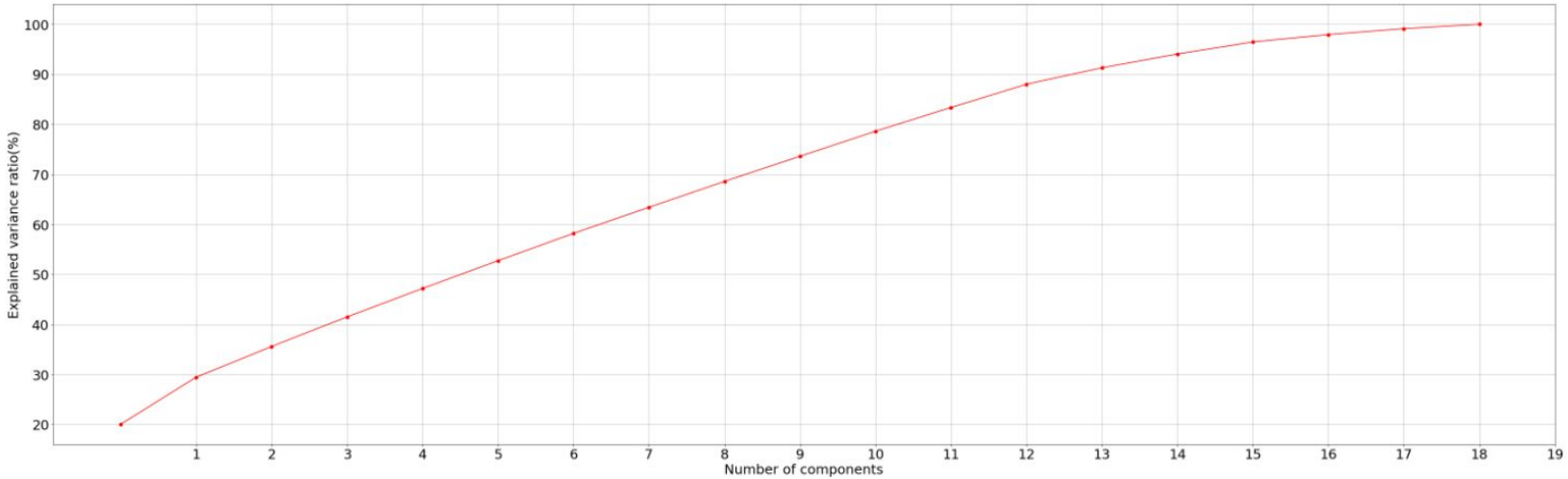
Gradient Boosting Classifier :



Dimension Reduction

Principal Component Analysis (PCA):

- The first 14 components contribute to **91% of the variance**.
- Post PCA, the following features with low weights in the components were removed:
  - *YearsSinceLastPromotion*
  - *WorkLifeBalance*



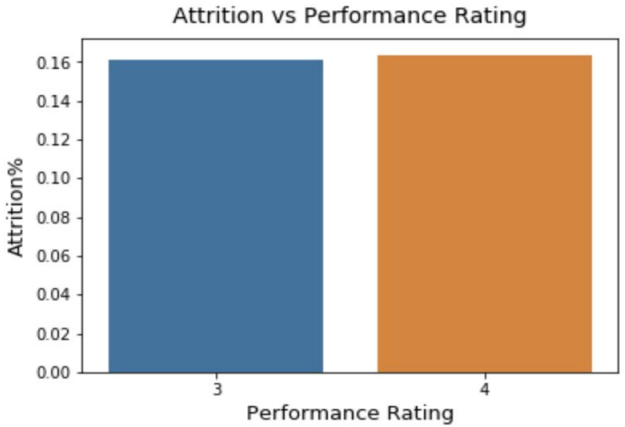
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
Age	0.234743	-0.011482	0.567719	0.296846	-0.028304	-0.071575	0.164872	0.073133	-0.081188	-0.003998	0.000488	0.243468	-0.0
DailyRate	-0.011889	0.017143	0.251464	-0.143783	0.483158	-0.295286	-0.196524	0.205490	0.345363	-0.310874	0.256704	-0.204628	0.4
DistanceFromHome	0.012573	0.062340	0.106955	-0.183368	-0.018989	0.618943	-0.145282	-0.361810	0.328559	-0.273194	0.189819	0.420580	0.0
HourlyRate	-0.015363	-0.008393	0.330950	-0.451121	-0.297154	0.052841	0.125925	0.225758	0.093406	-0.290236	0.087537	-0.373592	-0.5
MonthlyIncome	0.331000	-0.031592	0.366052	0.266621	-0.033116	-0.057400	0.159209	0.073168	-0.067699	-0.063160	-0.010733	0.147875	0.0
MonthlyRate	-0.005143	-0.017032	0.111321	0.419921	-0.237048	0.352586	-0.117245	0.098557	0.491817	0.185860	-0.277322	-0.472611	0.1
PercentSalaryHike	-0.012926	0.701203	0.041798	0.030200	-0.016510	-0.050843	-0.031270	-0.003692	0.005960	0.029844	-0.007471	0.001459	-0.0
TrainingTimesLastYear	-0.005515	-0.018956	-0.183745	0.019645	-0.158972	-0.436522	0.327810	-0.084629	0.686896	0.187188	0.141051	0.268098	-0.1
YearsAtCompany	0.472352	-0.008386	-0.098204	-0.062854	0.013026	0.005769	-0.009505	-0.009214	0.009131	0.022160	0.007840	-0.050108	-0.0
YearsInCurrentRole	0.421580	0.024493	-0.194713	-0.134639	0.029220	-0.004078	-0.112176	0.015725	0.054558	-0.002955	-0.041358	-0.103124	0.0
YearsSinceLastPromotion	0.368807	0.000745	-0.130386	-0.051245	-0.019395	0.017504	-0.067448	-0.030595	0.009155	0.083192	0.041110	-0.112096	-0.0
YearsWithCurrManager	0.416896	0.013547	-0.193851	-0.183708	0.023331	0.008949	-0.085984	-0.002781	0.021303	0.072058	-0.036329	-0.083268	-0.0
job_hop_index	0.351162	0.011299	0.030234	0.044486	0.027206	0.059547	0.087434	-0.035677	0.035711	-0.120925	0.036174	0.073430	0.0
EnvironmentSatisfaction	0.003116	-0.055547	-0.085774	0.325547	0.014026	-0.056480	-0.682732	0.309042	0.044356	0.000788	0.267302	0.193438	-0.4
JobInvolvement	-0.005544	-0.041593	0.298177	-0.362722	0.079050	-0.183105	-0.363923	-0.072991	0.124691	0.225479	-0.674684	0.240462	-0.0
JobSatisfaction	-0.006150	0.024128	-0.096062	0.280506	0.590318	0.023534	0.154610	-0.317853	0.084567	-0.292992	-0.294109	-0.179193	-0.4
PerformanceRating	0.008774	0.700813	0.017166	0.019065	-0.043856	-0.048258	-0.037673	-0.003138	-0.020483	0.041286	0.002278	-0.022711	-0.0
RelationshipSatisfaction	0.012864	-0.068689	0.262492	0.042586	-0.141336	-0.253313	-0.261472	-0.725333	-0.092728	0.187931	0.308176	-0.311869	-0.0
WorkLifeBalance	0.016521	-0.006897	-0.183998	0.162835	-0.463420	-0.305455	-0.158696	-0.133334	-0.023870	-0.684881	-0.281457	0.046043	0.1

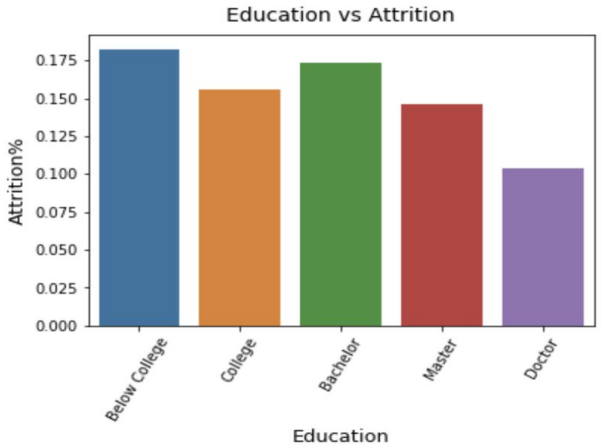
Feature Reduction (Categorical Variables)

Filter-Based Methods:

1) Removing some categorical features based on plots

We remove **“PerformanceRating”** and **“Education”** which are not good predictors.





Education	Attrition
Below College	0.182353
College	0.156028
Bachelor	0.173077
Master	0.145729
Doctor	0.104167

Feature Reduction (Categorical Variables)

Filter-Based Methods:

2) Removing features that are highly correlated

We remove **“YearsInCurrentRole"** with high **p value**

	MonthlyIncome	MonthlyRate	PercentSalaryHike	YearsAtCompany	YearsInCurrentRole	YearsWithCurrManager	job_hop_index
MonthlyIncome	1.000000	0.034814	-0.027269	0.514285	0.363818	0.344079	0.408088
MonthlyRate	0.034814	1.000000	-0.006429	-0.023655	-0.012815	-0.036746	-0.006971
PercentSalaryHike	-0.027269	-0.006429	1.000000	-0.035991	-0.001520	-0.011985	-0.010788
YearsAtCompany	0.514285	-0.023655	-0.035991	1.000000	0.758754	0.769212	0.599723
YearsInCurrentRole	0.363818	-0.012815	-0.001520	0.758754	1.000000	0.714365	0.432492
YearsWithCurrManager	0.344079	-0.036746	-0.011985	0.769212	0.714365	1.000000	0.437485
job_hop_index	0.408088	-0.006971	-0.010788	0.599723	0.432492	0.437485	1.000000

Feature Reduction Using Wrapper Method (Forward Selection)

Wrapper Method:

Using **Forward Selection** technique, we were able to **reduce the features to 20**.

This allowed us to overcome the *Curse of Dimensionality*.

```
['OverTime_Yes',
 'MaritalStatus_Single',
 'JobLevel',
 'JobInvolvement',
 'EnvironmentSatisfaction',
 'JobSatisfaction',
 'JobRole_Sales Representative',
 'BusinessTravel_Travel_Frequently',
 'YearsWithCurrManager',
 'DistanceFromHome',
 'JobRole_Laboratory Technician',
 'Department_Research & Development',
 'YearsSinceLastPromotion',
 'EducationField_Technical Degree',
 'WorkLifeBalance',
 'Age',
 'RelationshipSatisfaction',
 'TrainingTimesLastYear',
 'YearsInCurrentRole',
 'BusinessTravel_Travel_Rarely']
```

HyperParameter Tuning

```
Initial parameters: {'max_depth': 25, 'min_samples_split': 40, 'n_estimators': 100}
Improved parameters: {'max_depth': 3, 'min_samples_split': 18, 'n_estimators': 94}
```

Model Evaluation

Performance of Gradient Boosting Classifier:

	Precision	Recall	F1	Accuracy	AUC
Baseline	0.54	0.43	0.48	0.84	0.81
After optimization	**0.69**	**0.48**	**0.57**	**0.87**	**0.83**
Change (%)	**15%** ▲	**6%** ▲	**9%** ▲	**3%** ▲	**2%** ▲

Business Solution: Retention Strategy

We can use the model to **predict probability of an employee leaving.**

Based on probabilities, we create buckets of employees- high risk category, medium risk category, low risk category and no-risk category, and develop strategies accordingly to avoid attrition.

Probability_of_leaving ≥ 0.8	**High Risk**
0.6 ≤ Probability_of_leaving < 0.8	**Medium Risk**
0.5 ≤ Probability_of_leaving < 0.6	**Low Risk**
Probability_of_leaving < 0.5	**No Risk**

Strategy

High Risk	Medium Risk
A. If the employee is a high performer and has high potential, *immediate* action planning is needed. B. Initiate an open conversation with the employee to better understand their perspective and future goals C. Ask the employee's manager to have a conversation and explore the engagement levels and concerns, if any	A. Medium term action planning. B. Initiate one-on-one discussions between employee and manager or HR. C. Keep track of any behavioral change
Low risk	**'No risk'**
A. Long-term action planning B. Keep tracking for any behavioral change C. Have open house discussions within departments or groups.	No action required.

Suggestions

- Adopt a **Work From Home** strategy
- Investigate high levels of attrition within specific departments
- Holding 'Town Hall's with the overall company to provide transparency for employees
- Have HR regularly inform employees that they are open to any and all sorts of discussion

Legal Aspects

While Data Scientists are combing through HR databases, they have to collaborate with HR Compliance and Legal teams to ensure a seamless implementation of retention strategies and suggestions.

An **aggregate prediction** may be useful for several purposes, but there is no way we can make a decision about placing an employee on a long term assignment after predicting that he/she will leave based on his/her commute or another grievance.

Conclusion

We found the factors which are most important to employees, and if are not fulfilled, could lead to attrition. We built a **predictive model with highest possible F-measure** which can be used to predict which employees are likely to leave the organization. Moving forward, strategic measures can be introduced to prevent this and reduce turnover.

The model needs to be tuned from time to time as and when new dataset is received. In case any new input variable is introduced, it is important that the information is retrieved for the employees who participated in the initial study.

Thank you

Q&A