1. FEATURE SELECTION / ENGINEERING
   1. IMPORTANT FEATURES
      1. For Visualization and Analysis
         1. Wise Department Dashboard

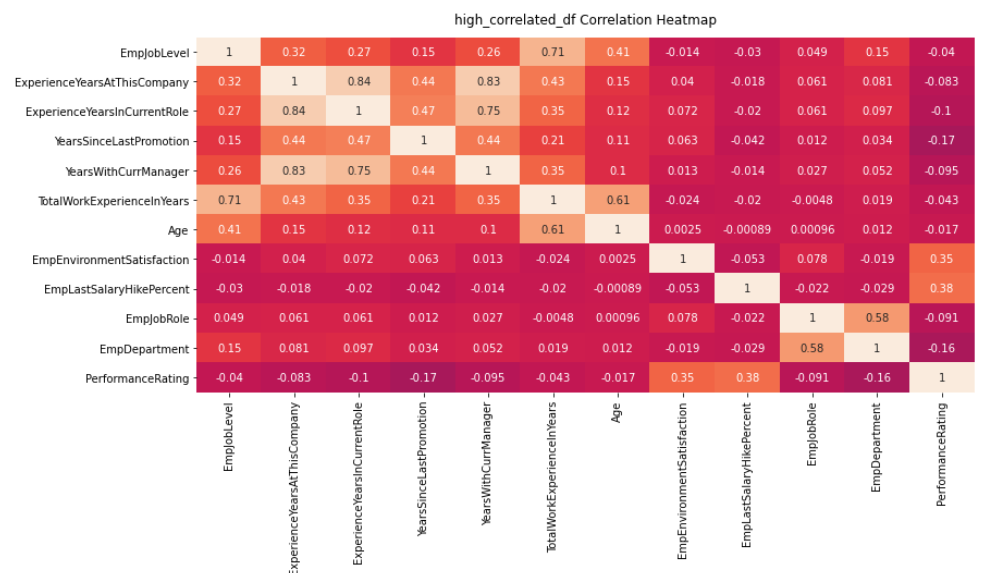
In Wise Department Dashboard, I use OverTime, EmpRelationshipSatisfaction, EmpEnvironmentSatisfaction, EmpJobSatisfaction, EmpJobInvolvement for Dimension1 features, EmpDepartment for Dimension2 feature, and PerformanceRating for specific measurement. I use that feature because from my point of view, if a department/company have great measurement on Dimension1 features that I mentioned in above, it must be great place to work for employees

* + - 1. Recommendation Dashboard

In Recommendation Dashboard, I use EducationBackround, BusinessTravelFrequency, and ExperienceYearsInCurrentRole for Dimension1 features, EmpDepartment for Dimension2 feature, and PerformanceRating for specific measurement. I use that feature because I want to find specific detail of each department which have the highest employee performance rating. So that, we can consider the insight to be next requirement to hire new employee.

* + - 1. Univariate Visualization

In Univariate Visualization that I code in visualize.ipynb, I use all features except EmpNumber to figure out visualization counts for each feature.

* + 1. For Machine Learning Model

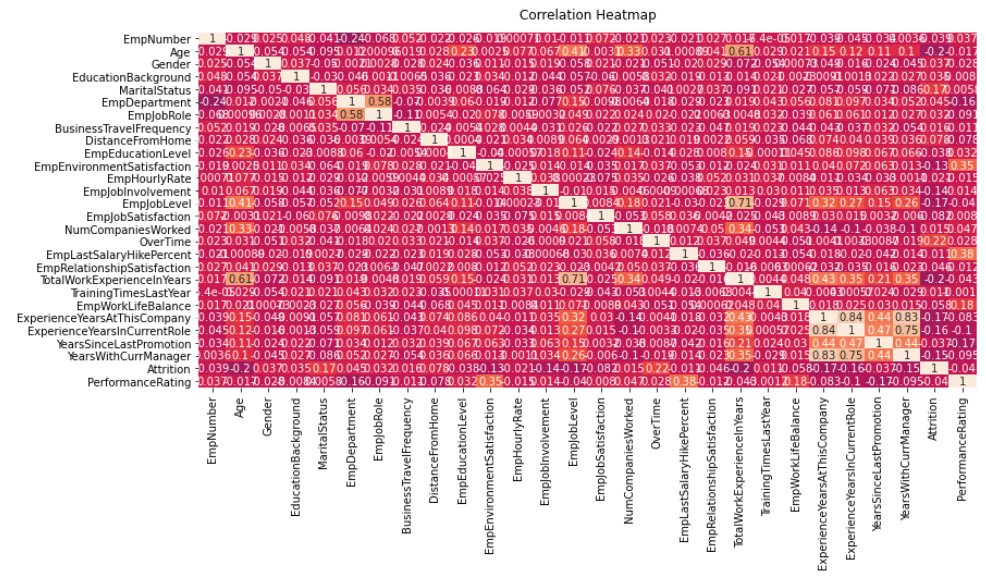
I use EmpJobLevel, ExperienceYearsAtThisCompany, ExperienceYearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, TotalWorkExperienceInYears, and Age because those features have high correlation between each other. And then I use EmpEnvironmentSatisfaction and EmpLastSalaryHikePercent because they have high correlation to PerformanceRating feature. Last, I use EmpJobRole and EmpDepartment because they have high correlation to both of them.

* 1. IMPORTANT FEATURES TRANSFORMATION

I do feature transformation on PerformanceRating features. I convert it from int64 to Object/String datatype to solving imbalance dataset through SMOTETomek Resampling method. The other feature transformation that I do is doing label encoding for categorical feature, so that the categorical feature will be converted to int64 datatype and the machine learning can learn the features.

1. RESULT, ANALYSIS AND INSIGHTS
   1. RELATIONSHIP FEATURES THAT DON’T FIT TO MACHINE LEARNING

I found that there are several features that doesn’t fit to machine learning feature, it because that most of them have low correlation between each other and to the PerformanceRating feature, such as EmpNumber, Gender, EducationBackground, and etc. The complete features that have low correlation can be viewed on all feature correlation heatmap below



* 1. IMPORTANT TECHNIQUES

There are several techniques that I used to improve the accuracy of machine learning prediction, i.e.:

* + 1. Removing Outliers

I remove the outliers because outlier can improve the correlation of each feature and it also improve learning accuracy of machine learning.

* + 1. SMOTETomek

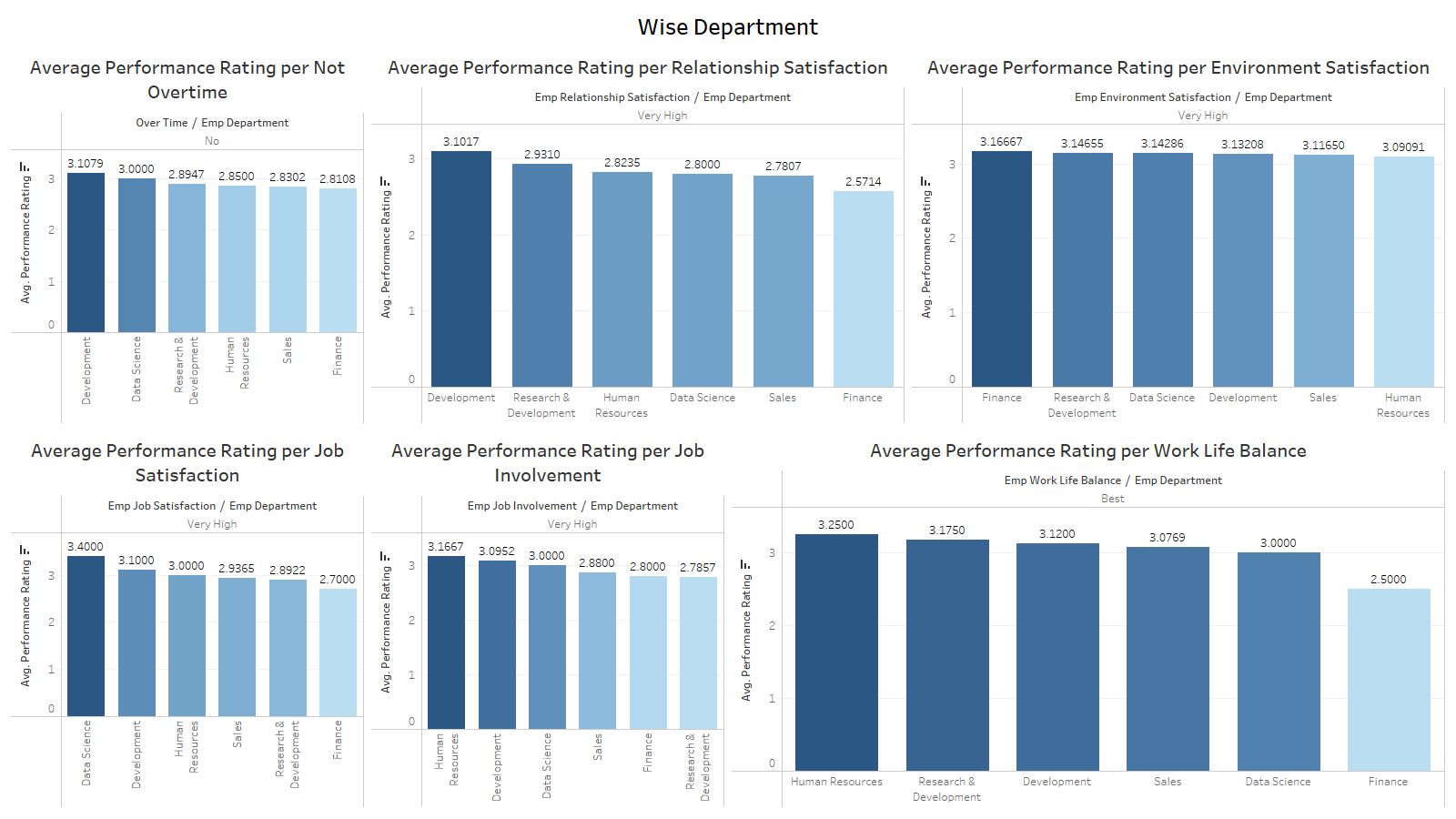
SMOTETomek is one of resampling method to solving imbalance data in the feature. If the dataset has balance ratio, the machine learning model will learn accurately.

* + 1. RandomizedSearch CV

RandomizedSearch CV is one of method in hyperparameter tuning model to improve accuracy of learning in the model through finding the best parameter for the machine learning model itself

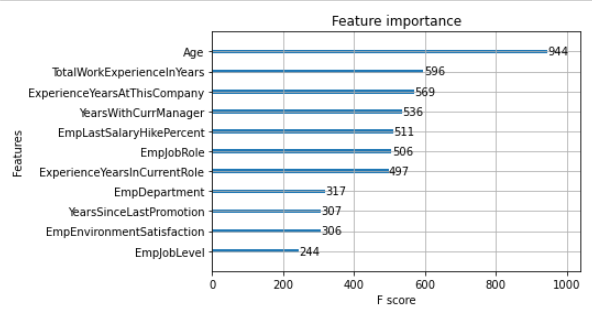
* + 1. Feature Selection through Correlation Heatmap

I use correlation heatmap to check and figure out about which features that have high correlation to each other and to the PerformanceRating. Because, the higher correlation between feature, the higher accuracy model would be.

* 1. ANSWERED BUSINESS PROBLEM
     1. DEPARTMENT WISE PERFORMANCE

From dashboard in above, we can conclude that:

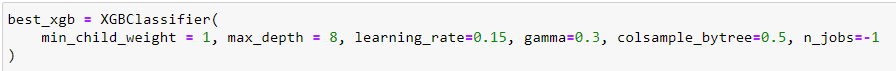
* Department which has high performance rating per Not Overtime is **Development Department**
* Department which has high performance rating per Relationship Satisfaction is **Development Department**
* Department which has high performance rating per Environment Satisfaction is **Finance Department**
* Department which has high performance rating per Job Satisfaction is **Data Science Department**
* Department which has high performance rating per Job Involvement is **Human Resources Department**
* Department which has high performance rating per Work Life Balance is **Human Resources Department**
  + 1. TOP 3 IMPORTANT FACTOR EFFECTING PERFORMANCE RATING

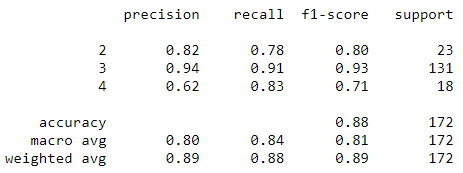


Based on graph in above, Top 3 Important Factor which effecting Performance Rating are Age, TotalWorkExperienceInYears, and ExperienceYearsAtThisCompany features.

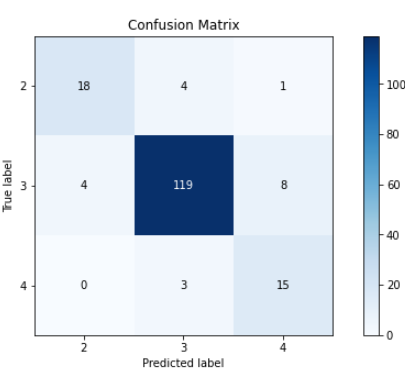
* + 1. RECOMMENDATION OF MACHINE LEARNING

Based on accuracy result of the model, I recommend 2 type of machine learning because the models have highest accuracy, i.e.:

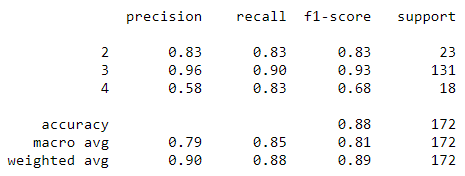
* + - 1. XGBoost
         1. Best Parameter
         2. Classification Report

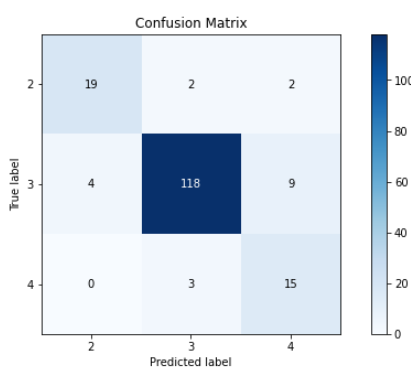
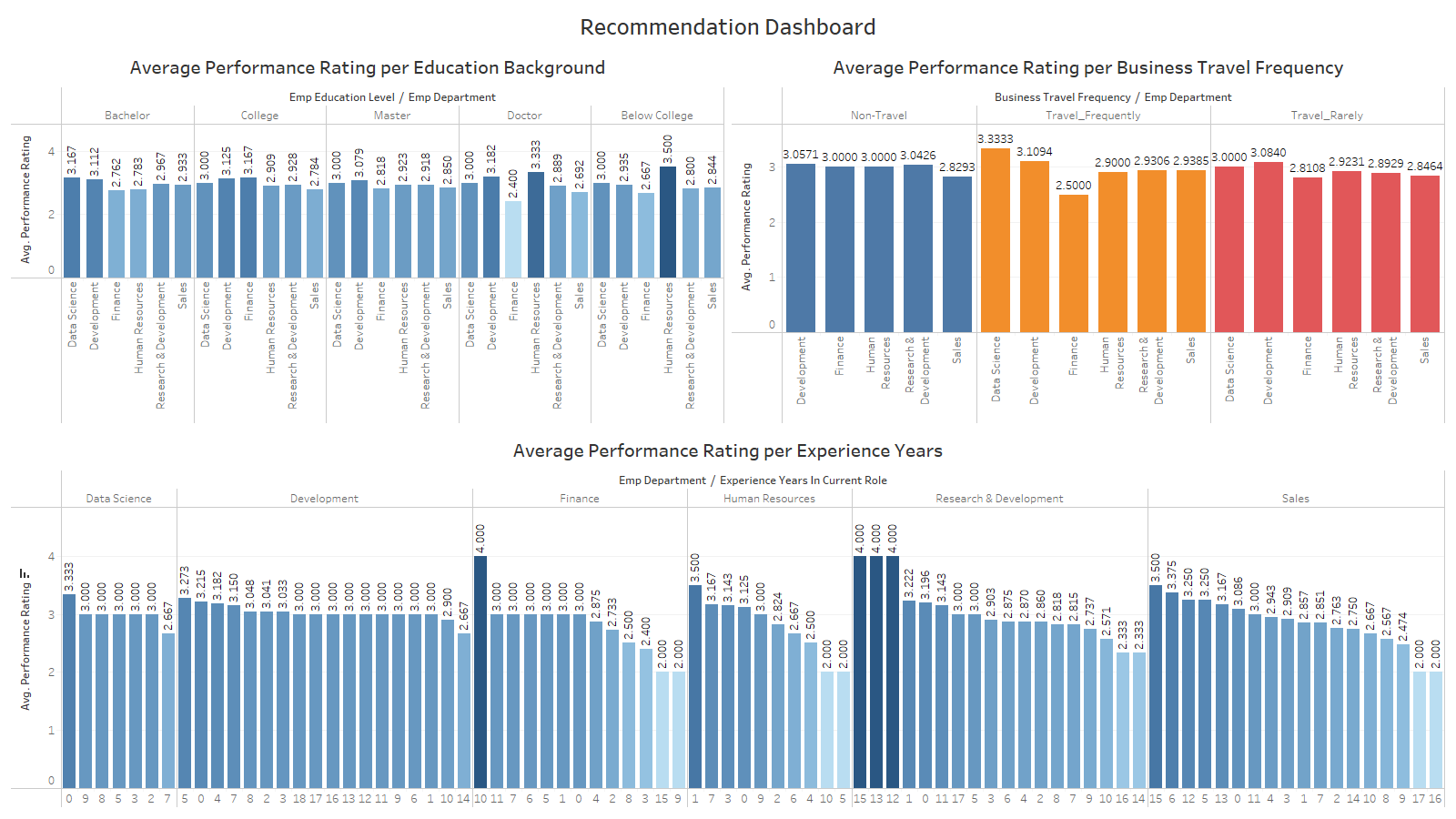


* + - * 1. Confusion Matrix



* + - 1. RandomForest
         1. Best Parameter
         2. Classification Report



* + - * 1. Confusion Matrix
    1. RECOMMENDATION TO IMPROVE PERFORMANCE RATING

Based on Recommendation Dashboard in above, there are 2 recommendation way to improve Performance Rating of Employee, i.e.:

* + - 1. We should hire/improve employee of each department with specification as below:
         1. **Data scientist** which has best performance rating are from Bachelor Degree and from 0 (zero) year experience.
         2. **Development** which has best performance rating are from Doctoral Degree and 5 year experience.
         3. **Finance** which has best performance rating are from College Degree and 10 year experience.
         4. **Human Resources** which have best performance rating from Below College Degree and 1 year experience
         5. **Research & Development** which has best performance rating are from Bachelor Degree and 15,13 and 12 year experience
         6. **Sales** which have best performance rating are from Bachelor Degree and 15 year experience.
      2. We should reconsider about business travel frequency of each department as below:
         1. **Data Scientist** should travel frequently to get high Performance Rating
         2. **Development** should travel frequently to get high Performance Rating
         3. **Finance** should not to travel (Non-Travel) to get high Performance Rating
         4. **Human Resource** should not to travel (Non-Travel) to get high Performance Rating
         5. **Research & Development** should not to travel (Non-Travel) to get high Performance Rating
         6. **Sales** should travel frequently to get high Performance Rating