**IABAC PROJECT**

**INX Future Inc Employee Performance**

**Report**

**Saurabh Kumar**

**Email:rsingh.saurabh@gmail.com**

**Table of Contents:**

|  |  |
| --- | --- |
| **Sr. No** | **Topics** |
| 1. | Project Summary |
| 2. | Business Problem |
| 3. | Requirement |
| 4. | Transformation to ML Problem |
| 5. | UNDERSTADING DATA |
| 6. | Exploratory DATA analysis |
| 7. | Statistics |
| 8. | Checking Normalization |
| 9. | Measure of Variance |
| 10. | Visualization |
| 11. | Feature Selection |
| 12. | PCA |
| 13. | Pearson correlation |
| 14. | Nominal Features Encoding |
| 15. | Scaling |
| 16. | Finding insights |
| 17. | Department wise performances |
| 18. | Top 3 Important Factors effecting employee performance |
| 19. | A trained model which can predict the employee performance |
| 20. | KNN Model |
| 21. | SVC Model |
| 22. | Decision Tree |
| 23. | Random Forest Classifier Model |
| 25. | XGBOOST Model |
| 25. | Recommendations to improve the employee performance based on insights from analysis |
| 26. | References |

**Project Summary**

**Business Problem:**

INX Future Inc , (referred as INX ) , is one of the leading data analytics and automation solutions provider with over 15 years of global business presence.

From the summary given we are able to determine that INX is facing issues with their employee performances and due to which they are facing challenges in their service delivery and customer satisfaction. These problems should be taken seriously as they can impact company’s revenue.

**Requirement:**

We have to find the core underlying causes of this performance issues. We are expected to findings of this project will help him to take right course of actions. We are also expected to find clear indicators of non-performing employees, so that any penalization of non-performing employee, if required, may not significantly affect other employee morals.

The following insights are expected from this project.

1. Department wise performances

2. Top 3 Important Factors effecting employee performance

3. A trained model which can predict the employee performance based on factors as inputs. This will be used to hire employees

4. Recommendations to improve the employee performance based on insights from analysis.

**Transformation to ML Problem:**

After analysis the problem statement and requirements, we can see we have to do descriptive and predictive analysis mainly with focus on employees’ performance. After analyzing dataset provided, we can see the employee performance is captured by performance rating.

The feature is of type categorical, dividing data between following values:

|  |
| --- |
| 1 'Low' |
| 2 'Good' |
| 3 'Excellent' |
| 4 'Outstanding' |

The other features in data set are of mix type: nominal, ordinal and continuous.

Along with descriptive analysis we have to predict performance rating.

It means we have labeled data and we will be using Supervised Machine learning.

As the outcome variable is categorical, we have to use classification ML model.

As the predictor is polytomous variable, we cannot use binary regression for classification. The models we can use are:

KNN

Decision Tree

Random Forest

XGBOOST etc.

Also, we need to do feature selection through correlation between the predictors, PCA component analysis. We need to understand data using graphs and through tools of measuring measures of central tendency and variance.

We also need to find if we have any predictor which is normally distributed or there is no normal distribution.

Before doing that, we need to prepare data, do basic checks and make the data eligible to be fed to model.

**UNDERSTADING DATA:**

The data set is present at below link:

<http://data.iabac.org/exam/p2/data/INX_Future_Inc_Employee_Performance_CDS_Project2_Data_V1.8.xls>

Below mentioned table contain describing data columns:





**Below mentioned features are Nominal:**

"Gender","EducationBackground","MaritalStatus","EmpDepartment","EmpJobRole","BusinessTravelFrequency","OverTime","Attrition"

**Below mentioned features are Ordinal:**

"EmpEducationLevel","EmpEnvironmentSatisfaction","EmpJobInvolvement","EmpJobLevel","EmpJobSatisfaction","EmpLastSalaryHikePercent","EmpRelationshipSatisfaction","EmpWorkLifeBalance".

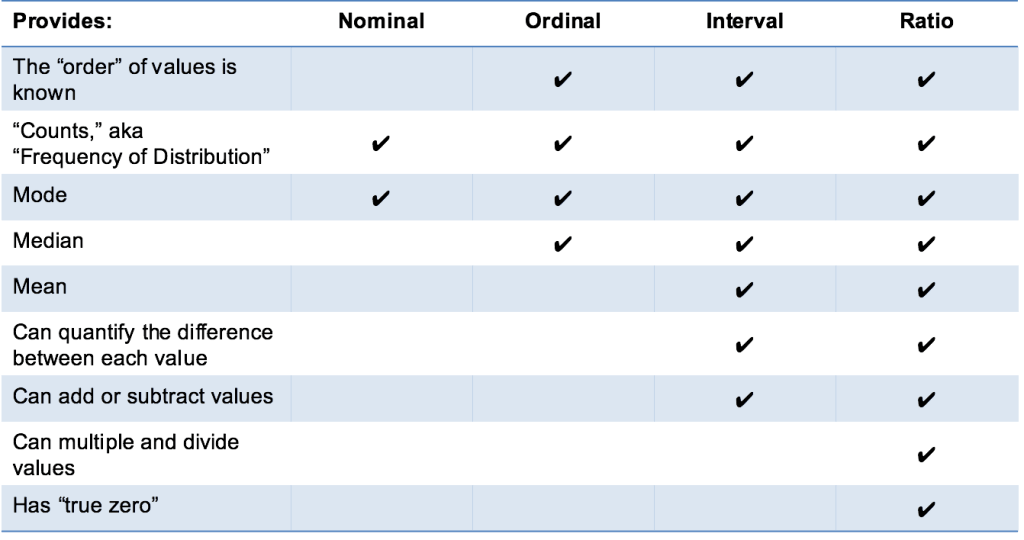
**Below mentioned features are continuous:**

"Age","DistanceFromHome","EmpHourlyRate","NumCompaniesWorked","TotalWorkExperienceInYears","TrainingTimesLastYear","ExperienceYearsAtThisCompany","ExperienceYearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager"

**Exploratory DATA analysis:**

Overall the data set has 1200 records and 28 columns.

We shall get statistics based on below figure for different scale data column.



**Statistics:**

Nominal data



Nominal data



Ordinal Data





**Checking Normalization:**

Normality test

Categorical data are not from a normal distribution. The normal distribution only makes sense if you're dealing with at least interval data, and the normal distribution is continuous and on the whole real line. If any of those aren't true you don't need to examine the data distribution to conclude that it's not consistent with normality

Still I went for finding out normality of ratio features using Shapiro Wilk Test.

I found no features is normally distributed. For reference you can find results in EDA\_DATASET.ipynb

**Measure of Variance:**

Using one side Ttest, Anova for continuous features and Chi square test for categorical data we are able to find the the variables are independent of each other. For reference you can find results in **EDA\_DATASET.ipynb**

**Checking for Null values/NaN values:**

I did not find any null value in data set.

**Visualization:**

I used count plot, frequency histogram and box plot for visualization.

Using statistics and visualization I was able to under data. Some of the insights are mentioned below:

The employees dataset have 725 males and 475 females.

Maximum number of employees have life sciences background, then from medical.

Maximum number of employees are married followed by single and then some are divorced.

Maximum number of employees are developers. Many employees travel rarely for their job. Very few employees do overtime. Out of 1200, only 178 employees have left organisation.

Highest number of employees are graduates followed by masters. Doctorates are lowest number

The employees seem satisfied with environment. Majority seem satisfied with job environment. The job satisfaction is also high.

Mean age of employees is 37 years with oldest employee at 60 years. Maximum distance of employee is 29 km.

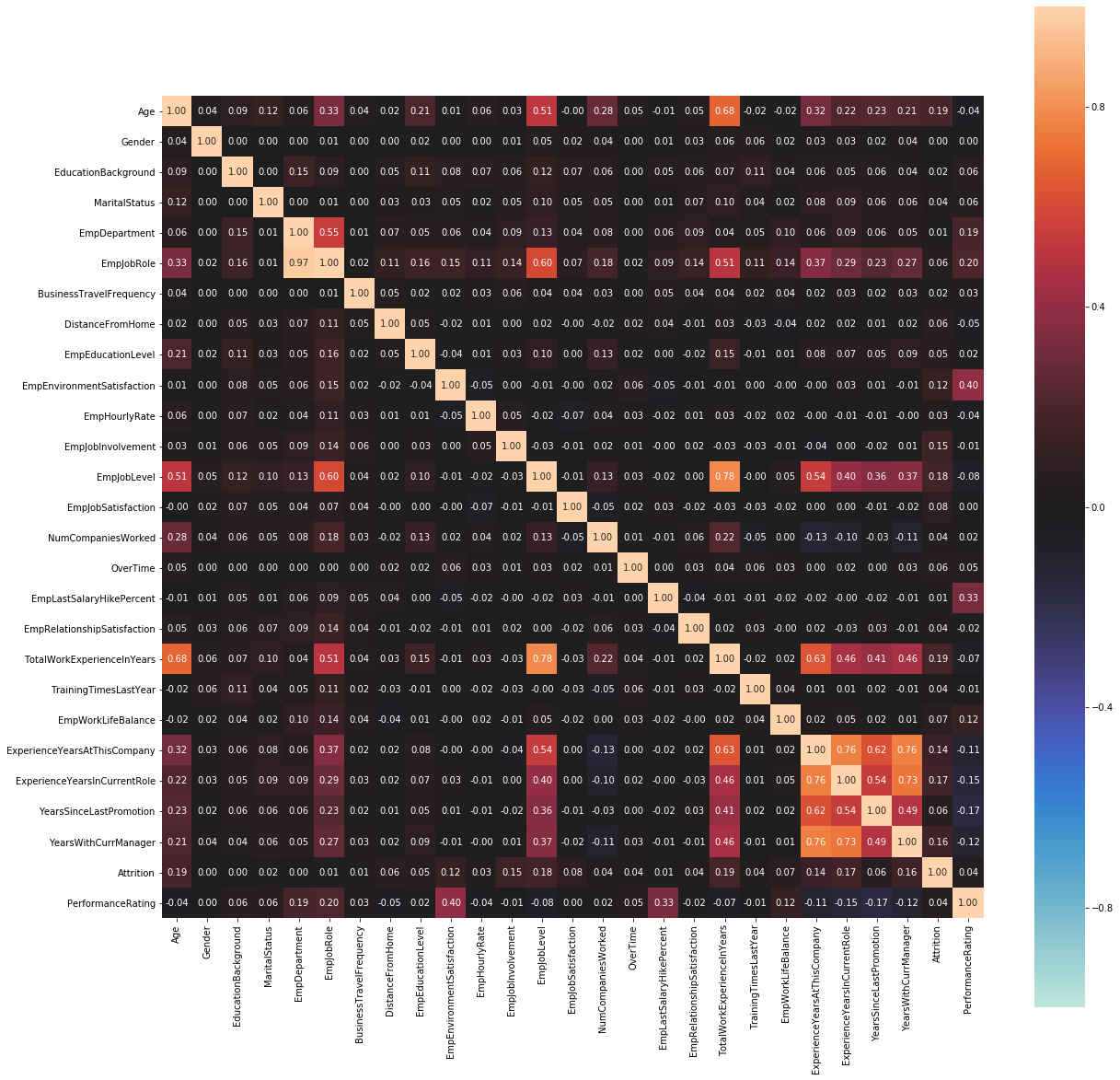
Mean hourly rate is 65 dollars. Average experience of employees is 11.3 with average experience in this company is 7.07. Average years of experience in current role is 4.2 with maximum is 18 years. Average years since last promotion is 2.1 years. Average years with current manager is 4.1 with maximum value of 17 years.

**For count plot you will find graphs in Graphs\_Count, histograms in hist folder /frequency plots and box plot in folder box/box\_ploting.**

**The notebook file is IABAC\_visualization. ipynb.**

**Feature Selection:**

First I used dython library as my predictors are of mixed type and as far as I can read about dython, it suits finding correlation between mixed predictors. Below mentioned is head map for same.



I got strong correlation between few predictors but as I am not subject matter on Human Resource domain, I was sceptical of removing all and I am not sure whether they will increase noise or they will add any information.

**PCA:**

I thought of feature extraction using PCA which I have studied and is a good model for feature extraction. PCA reduces the number of features to a few principal components without losing much of the information. IT is most effective dimensionality reduction technique works fairly good even with categorical data. It assumes that PCA factors are linear factors of original features. IT works on variance as measures of how important is a particular dimension.

As I have nominal string data, I need to do encoding, I choose **label encoding** as the values in my features are not very large in range. I could have gone for one hot encoding or binary encoding but they would increase number of features, I already have decent number of features with just 1200 records.

After that I did **scaling** to bring all features on same scale. I used Standard Scaling. (will be discussed further in other section).My data was not having very large difference on scale but still scaling can help in generalizing my model.

After doing PCA, I found my variance was still being explained by most of variables and it did not help in feature selection or feature extraction.

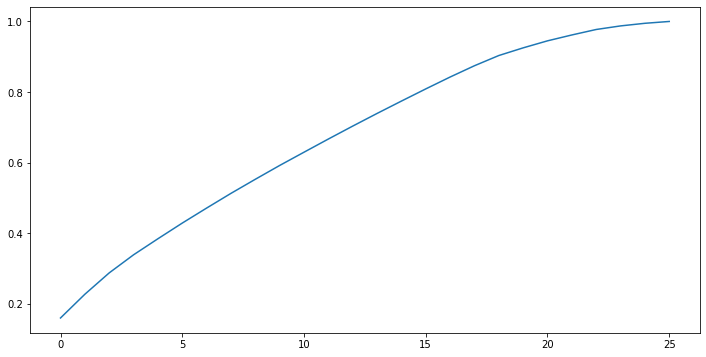
My parameters and graph explained PCA does not help.

Hence I cannot use PCA.

My variance ratio explained:



Below mentioned is my explained variance graph:



My feature selection file is **feature extractions, feature selection. Ipynb**

**Pearson correlation:**

Now after encoding the I was able to do. Pearson correlation

I found below mentioned features which were highly correlated

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **TotalWorkExperienceInYears** |  | 0.680886 |
| **EmpDepartment** | **EmpJobRole** |  | 0.568973 |
| **EmpJobLevel** | **TotalWorkExperienceInYears** |  | 0.784229 |
| **TotalWorkExperienceInYears** | **ExperienceYearsAtThisCompany** |  | 0.633555 |
| **ExperienceYearsAtThisCompany** | **ExperienceYearsInCurrentRole** |  | 0.764102 |
| **ExperienceYearsInCurrentRole** | **YearsSinceLastPromotion** |  | 0.5406 |
| **ExperienceYearsInCurrentRole** | **YearsWithCurrManager** |  | 0.728973 |
| **YearsSinceLastPromotion** | **ExperienceYearsAtThisCompany** |  | 0.62023 |
| **YearsWithCurrManager** | **EmpEducationLevel** |  | 0.088988 |
| **YearsWithCurrManager** | **ExperienceYearsAtThisCompany** |  | 0.759258 |
| **YearsWithCurrManager** | **ExperienceYearsInCurrentRole** |  | 0.728973 |

I decided to remove mentioned features, as they are the one who are correlated with many other features:

|  |
| --- |
| **YearsWithCurrManager** |
| **ExperienceYearsInCurrentRole** |
| **TotalWorkExperienceInYears** |

**Nominal Features Encoding:**

There are many techniques foe encoding the nominal variables. The oldest being one host encoding (dummy variables), Binary encoding, label encoding where we will give numerical values to the string values.

Binary encoding will create extra features which will increase our features and can add noise in our model. Label encoding too can cause your model to add non relevant information if the value count in your nominal variable is too high. In our dataset We do not have too many value count in variables ,hence I opted for Label Encoding .For binary or dummy encoding we have data set of 1200 rows with 28 features, I do not want to increase more features.

**Scaling:**

Real world dataset contains features that highly vary in magnitudes, units, and range. Normalisation should be performed when the scale of a feature is irrelevant or misleading and not should Normalise when the scale is meaningful.

Examples of Algorithms where Feature Scaling matters

1. K-Means uses the Euclidean distance measure here feature scaling matters.

2. K-Nearest-Neighbours also require feature scaling.

3. Principal Component Analysis (PCA): Tries to get the feature with maximum variance, here too feature scaling is required.

4. Gradient Descent: Calculation speed increase as Theta calculation becomes faster after feature scaling.

Note: Naive Bayes, Linear Discriminant Analysis, and Tree-Based models are not affected by feature scaling. In Short, any Algorithm which is Not Distance based is Not affected by Feature Scaling.

**Finding insights:**

1. **Department wise performances**

Below mentioned are the department wise performance ratings mean.



From the table above we are able to analyse that Development department employees have delivered best. Data Science department is a near second. In middle lies Human Resource and R&D. The bottom department in terms of performance **Finance and sales** which are pretty important departments and we need to improve performance in these departments.

The analysis done for department wise performance is present in **Department wise performances.ipynb**.

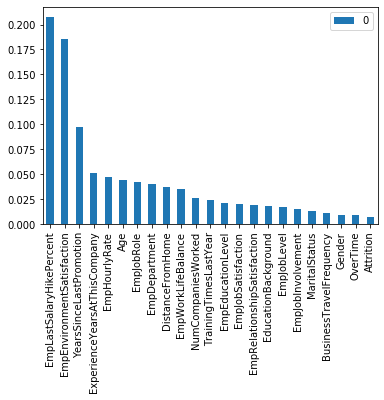
1. **Top 3 Important Factors effecting employee performance**

I used tree and tree ensemble models to find the feature importance, as I am not domain expert and I have not removed all the features with correlation. These models are not based on any presumption on data being normal or any other distribution and data is not suppose to be parametrised. I used Decision tree feature importance, Random Forest and XGBOOST. My RANDOM FOREST model was most accurate.

From all the models I received same important features.I have used feature\_importance\_ parameter of trained model.



The graph plotted below will give more clarity.



The three most important features in terms of employee performances.

EmplyeeSalaryHikePercent

EmpEnvironmentSatisfaction

YearsSinceLastPromotion

The files used to reach this conclusion are **Random\_Forest\_Feature\_importance.ipynb, XGBOOST\_feature\_selection.ipynb and DT\_feature\_importance.ipynb**

1. **A trained model which can predict the employee performance:**

As we have already discussed while transferring our business problem the problem is of classification supervised machine learning. We did tried models used for classification of polytomous variable. Let’s analyse KNN first.

**KNN Model:** KNN model is based on classifying or grouping the values based on distance. For KNN model we need to scale down data as finding distances towards centroid based on same scaled data makes sense. After scaling data I trained my model and got an accuracy 74 percent.

Then I cross validated using K fold validation and analysed 74 percent seem fine. Then I did hyper parameter tuning GridSearch Cross Validation and was able to improve my model to only 76 percent.

The file is **Modeling\_KNN.ipynb**

**SVC Model:** Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.

In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

Then, we perform classification by finding the hyper-plane that differentiate the two classes very well. We have to do tuning for SVM for kernel trick and computation wise it is costly. Major applications of SVM are face recognition, text and hypertext categorization etc.

My SVC model was not able to predict Good and Outstanding category at all. Also, the accuracy of model was 73 percent.

 Moreover, it depends on doing bunch of counts It can train on small set. It can learn features independently but cannot learn relationship between features.

The file is **SVM Model.ipynb**

**Decision Tree:** Decision tree is Supervised Algorithm considered to be one of the finest and mostly used for classification problem. It works well both categorical and continuous input and output variables empowers models with high accuracy, stability and ease of interpretation. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision node and leaf node. It is easy to understand, we need not do do very much data cleaning, also it is considered as non-parametric method, it can b of any distribution. But we need to take care as it can over fit data and it results in low accuracy for continuous variables. Also, it has been proven that it helps in Business Decision Support. To manage overfitting, we do pruning.

I trained model and cross validated and did hyperparameter tunings, I was able to find the accuracy of model is 87 percent.

I pruned the tree at max depth of 10. Though I have to admit my resources were not enough to do hyper parameter tuning very efficiently.

The file I used is **IABAC\_DT (3).ipynb**

**Also to mention the dataset is unbalanced towards Excellent being majority class and good & outstanding being minority class.**

**This can cause issue of my model tilting towards majority class even with high accuracy without being accurately predicting minority class.**

**Hence, I used SMOTE technique of balancing the dataset which I found better than under sampling. I used SMOTE technique in all my tree-based models.**

**Random Forest Classifier Model:**

We have already achieved 87 percent accuracy through decision tree. Now we will try using ensemble model based on random decision tree --random forest.

It will help us to get rid of overfitting problem of decision tree. It operates by constructing a multitude of decision trees at training time and then selecting the classes based on majority voting.

It builds random subsets of trees and then combine the subtrees. It can be slower than DT. As our problem is of classification, we will use Random Forest classifier.

The accuracy of my Random forest model is 94.4 percent.

The file is **Random Forest.ipynb**

**XGBOOST Model:**

We have achieved 94.4 percent accuracy through Random Forest. XGBoost is normally used to train gradient-boosted decision trees and other gradient boosted models. Random forests use the same model representation and inference, as gradient-boosted decision trees, but a different training algorithm.

XGBoost combine decision trees, but start the combining process at the beginning, instead of at the end. If you carefully tune parameters, gradient boosting can result in better performance than random forests.

However, gradient boosting may not be a good choice if you have a lot of noise, as it can result in overfitting. They also tend to be harder to tune than random forests.

The accuracy score of my model is 93.3 percent.

I used cross validation and hyperparameter tuning for both Random Forest and XGBOOST model. With my laptop machine I faced challenge finding best parameters for both my Random Forest and XGBOOST model.

The file is **XGBOOST.ipynb**

**After analysing all model, I choose Random Forest as it gives me highest accuracy. Also, XGBOOST takes lot more time while training as compared to Random Forest.**

1. **Recommendations to improve the employee performance based on insights from analysis:**
2. As we have found Employee salary hike percent, employee environment and time since last promotion.

So, in order to improve employee performance, we need to improve office environment. We need to critically analyse how the company is doing hike and they need to re think about their promotion policy.

If we look at Office environment variable, nearly 40 % of people are not very much satisfied.

1. The job satisfaction too shows same reading as environment with about 40% people not satisfied much.
2. When we will improve environment, we will also see improvement in relationship satisfaction which again has about 40% people reporting they do not have good relationship.
3. We need to work on improving these parameters Department wise, with initially more focus on Finance and Sales department which are very important for revenue generation of company.
4. Another thing which We analysed from dataset is that People are not being rated at lowest level. This show the bell curve is not being met and people who are relatively performing low are rated one rank above. These people will not improve and hence will not create a space to hire new efficient people in their place.

**References:**

1. Most of the knowledge has been acquired though DataMites.
2. <https://machinelearningmastery.com/calculate-feature-importance-with-python/>
3. <https://pythonfordatascience.org/anova-python/>
4. <https://www.geeksforgeeks.org/python-how-and-where-to-apply-feature-scaling/>
5. <https://discuss.analyticsvidhya.com/t/dimensionality-reduction-is-good-or-bad/2444/4>
6. <https://towardsdatascience.com/building-a-k-nearest-neighbors-k-nn-model-with-scikit-learn-51209555453a>
7. <https://towardsdatascience.com/k-nearest-neighbors-its-purpose-and-how-to-use-it-36fa927acc64>
8. <https://www.researchgate.net/post/Is_there_a_specific_reason_that_using_PCA_gives_worse_results_than_without_using_it_in_SVM_classification>
9. <https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35>
10. <https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18>
11. <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sort_values.html>