NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science

Bachelor's Programme "HSE and University of London Double Degree Programme in Data Science and Business Analytics"

Programming project final report

on the topic Prediction of employee churn

Student:

group БПАД182 _____ Kseniia Yakunina Pavlovna

Sign Name, Surname, Patronymic

21.06.2020

Date

Supervisor:

Polunina Polina Alekseevna

Name, Surname, Patronymic

Lead HR, Finance & Videoanalytics

Position

ООО «МВМ»

Company

Date 21.06.2020 _____ _____

Grade Sign

Moscow 2020

Contents

Abstract

The object of the study is the problem of the outflow of employees as a result of attrition. The purpose of the work is to find out which employees tend to leave most often as a result of burnout, and also build a model that will predict the departure of the employee. The work was performed using machine learning methods. Six models were build, the best result was obtained from the model Extra Tree algorithm by the following metrics: Accuracy, F1_score, RocAUC.

Keywords: *machine learning, EDA, cross-validation, ensembles of models, generalizing ability of the algorithm, retraining, quality functional.*

Introduction

This work relates to machine learning methods, the construction of binary classification models and the implementation of various algorithms.

The dataset have the main information about the employees of the company, as well as the value of the target variable. It is binary classification problem, where the first class contains employee who still work at the company, and the second one is people who decided to leave in case. of attrition.

The work involves the analysis of data on company employees, building models that predict the likelihood of an outflow of employees.

Relevance of the work

It is beneficial for the employer to keep an employee who is already working because this helps to optimise processes and reduce the cost of hiring new employees. Identifying the aspects that affect the burnout of employees will help optimise employee churn. Nowadays, more and more solutions for this problem for making predictions about employee turnover are appearing on the market, which is sold by subscription to various companies. This work will help evaluate the feasibility of solving this problem by using machine learning methods and reduce the time to develop an internal solution for the company.

The Target and objectives of the work.

The purpose of the work is to find out which employees tend to leave most often in case of attrition. and to construct binary classification models that will predict the employee's departure.

The following objectives were established for achieving the aim of the work:

1. Conduct Exploratory Data Analysis. Identify patterns in the dataset, check data on outliers, consider the correlation of features among themselves and their distribution by the target variable. It is also necessary to visualise the received data.
2. Prepare data for model construction. Remove outliers is it necessary for data, it is also important to code categorical features.
3. Use sampling methods to work with unbalanced data to increase the accuracy of the model.
4. Selection of optimal quality metrics for the binary classification problem.
5. Construct models and assessing their quality on a deferred sample.
6. Analysis of the obtained results of models based on metrics, as well as the selection of the best model.
7. Interpretation of the obtained results.

Review of sources

According to studies, finding a new employee takes up to 20% of the employee's salary. Also , for business it is important to understand the causes of employee attrition and prevent it.

In the article «Prediction of Employee Turnover in Organisations using Machine Learning Algorithm» , the author mentions that staff turnover problems negatively affect the company's long-term growth strategy. «The problem of employee turnover has shot to prominence in organisations because of its negative impacts on issues ranging from work place morale and productivity, to disruptions in project continuity and to long term growth strategies» The following author mentions the methods applicable to this problem: *Xgboost, Logistic Regression, Naive Bayes, Random Forest,*

SVC, LDA, KNN, Random Guess. The gradient beads ensemble model showed the best result in this study.[3]

However, the authors of the article “Employee Attrition Risk Assessment using Logistic Regression Analysis” mentions Logistic regression as the method that shows the closest result, and therefore recommend using it. [4]

In article «Employee Churn Rate Prediction and Performance Using Machine Learning» mentions the Random Forest classifier as the best algorithm for this type of problems. There are three significant conclusions about employee churn : « 1. If no promotion for more than five years then an employee will leave. 2. If no raise and more working hours and high salary then an employee will leave. 3. If no salary raise for employee but got promotion tend to leave an organisation.»[5]

After analysing sources, it is possible to make an assumption that one of three methods will give the best result and we have to check. it. This tree methods is : Gradient boosting, Logistic Regression and Random Forest Classifier.

Theoretical part.

Since this work is directly related to machine learning, then the python was chosen as the programming language in which this project will be written.

It is usually to use python to write programs, there are a large number of useful libraries that simplify code writing and help beautifully visualize the data received.

This task belongs to the type of classification task in machine learning. And more precisely, then to the binary classification, because we need to determine whether the employee will leave or not.

The following Python libraries were used at this stage:

• Csv – library which is works with csv files

• Numpy – library for works with arrays of data

• Pandas - library for data processing and analysis. Designed for data of different nature - matrix, panel data, time series.

• Scipy – library for make calculations

• Sklearn - library with machine learning methods

• Matplotlip – library for creating plots.

• Seaborn - library with machine learning methods

• Datetime - library helps to work with date format.

• Statsmodels.api - library helps set style for plots.

• imblearn - Library helps to work with imbalanced data and apply sampling methods

• catboost - Library for Cat Boost Classifier method

Since, dates has imbalanced classes, that is mean that we have to use sampling methods for make data more balanced. The over sampling and under sampling methods were used in this work. Both of this methods create equals share for both classes in binary classification.

Over sampling method is increased number of values in class with the smallest share by copping it values.

Under sampling method is decreased number of values class with the largest share.

Cross-validation splits our training sample into k folds, in this case there are 10 of them. The model is trained on the k-1 block and tested on the k-th.[]

There are many approaches that can be used for binary classification. In this work was used the following methods:

1. Logistic Regression - a method for constructing a linear classifier, which allows one to evaluate the probabilities of objects belonging to classes. [11]

2. The method of Supports Vectors - the support vector machine algorithm finds a hyperplane in an N-dimensional space (there is N a number of features) that classifies the data points. [1]

3. Extra Trees Classifier - the decision tree is a machine learning algorithm based on the construction of a tree where each edge is an attribute of the objective function and each

leaf is the value of the objective function. Affiliation to a class is determined by go across a tree from root to the end. [2]

4. Random Forest - «Random forest classifier algorithm consists of a large number of individual decision trees that operate as an ensemble algorithms . Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes model's prediction». [2]
5. Gradient Boosting - an ensemble model based on Decision Trees.
6. Cat Boost Classifier - Modernised version of gradient boosting algorithm based on Decision Trees

Realisation of the Project

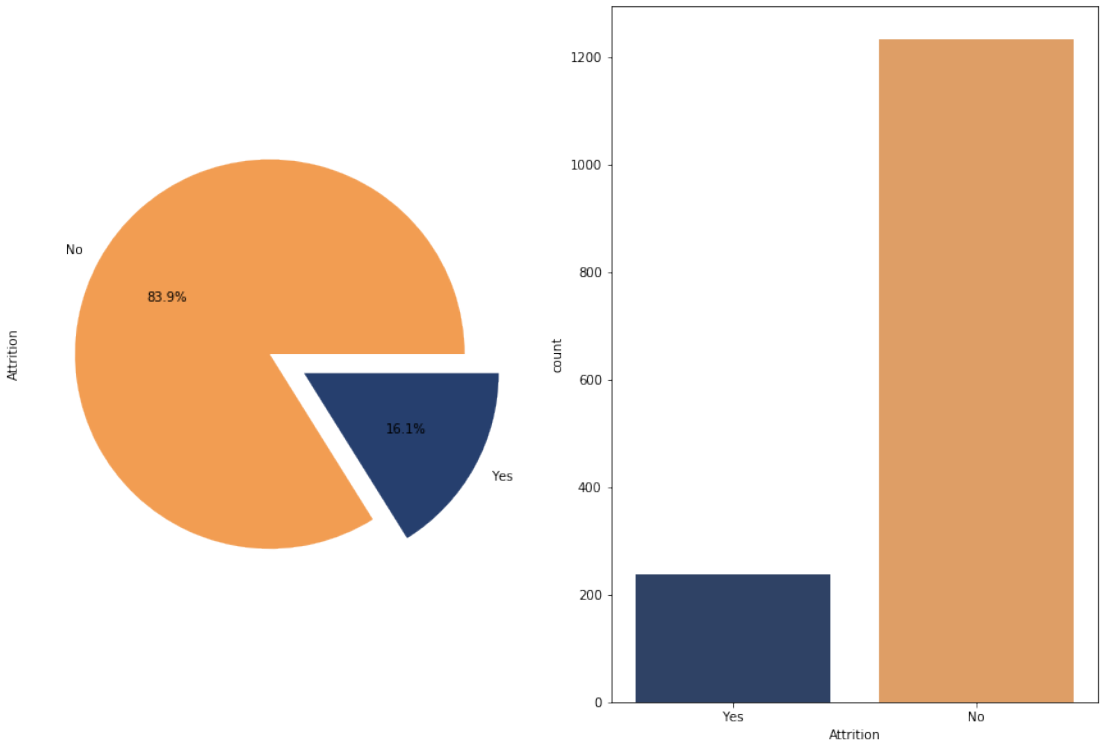
Structure of the project.

1. Exploratory Data Analysis
2. Preparing data for model construction.
3. Choosing metrics.
4. Implementation of Models
5. Comparison analyses of models
6. Conclusion

Exploratory Data Analysis

We have 1,470 objects and 35 features.

The data are not balanced in relation to a target variable-ratio: 83% with people who continue working at the company to 16% with people who left.



Let’s consider in more detail the available features. We have 26 numerical and 8 object features.

Numerical Features: Age, DailyRate, DistanceFromHome, EmployeeNumber, HourlyRate, MonthlyIncome, YearsWithCurrManager, YearsInCurrentRole, TotalWorkingYears, StandardHours, NumCompaniesWorked, YearsAtCompany, PercentSalaryHike, YearsSinceLastPromotion.

Encoding Categorical features: Education, JobLevel, JobInvolvement, EnvironmentSatisfaction, PerformanceRating, WorkLifeBalance, RelationshipSatisfaction.

Categorical features: BusinessTravel, Department, Gender, MaritalStatus, OverTime, EducationField , JobRole.

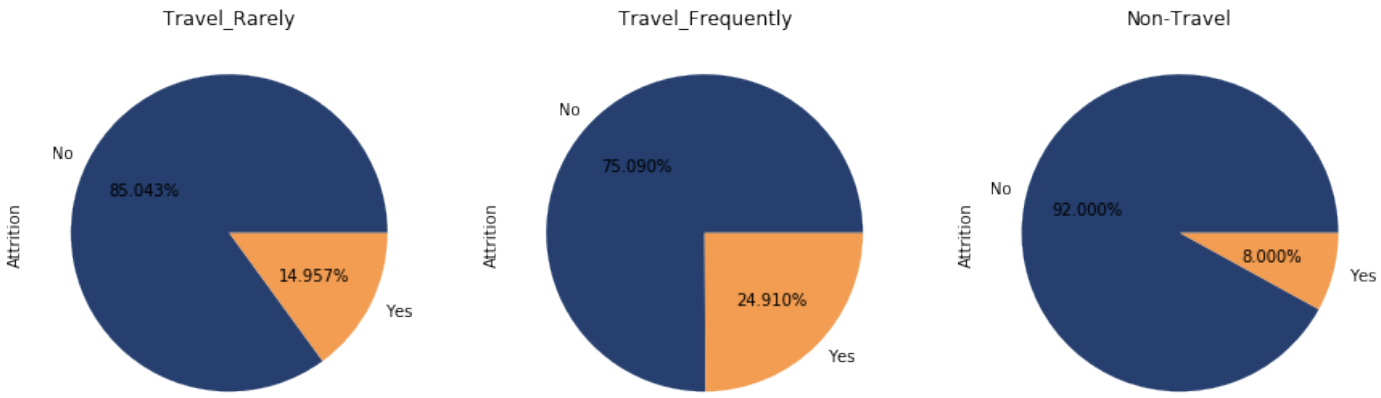
There are 3 constants in these signs that tell us that all employees are over 18 years old, all employees have a rate of 80 hours and the number of employees with the same number. These data do not carry essential information for analysis, therefore, they were deleted. After it we have 23 numerical

The employee index column is unique and not repeated. That is why it was set as index.

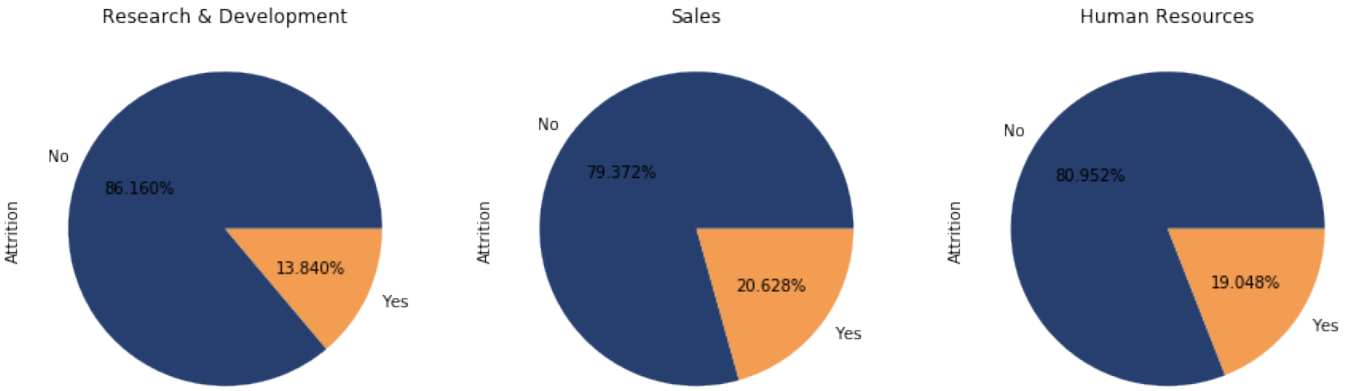
The most correlation features in dataset is : *TotalWorkingYears and JobLevel, MonthlyIncome and JobLevel, MonthlyIncome and TotalWorkingYears, PercentSalaryHike and PerformanceRating, YearsInCurrentRole and YearsAtCompany, YearsWithCurrManager and YearsAtCompany.*

After analysing the signs and objective function, the following conclusions were made:

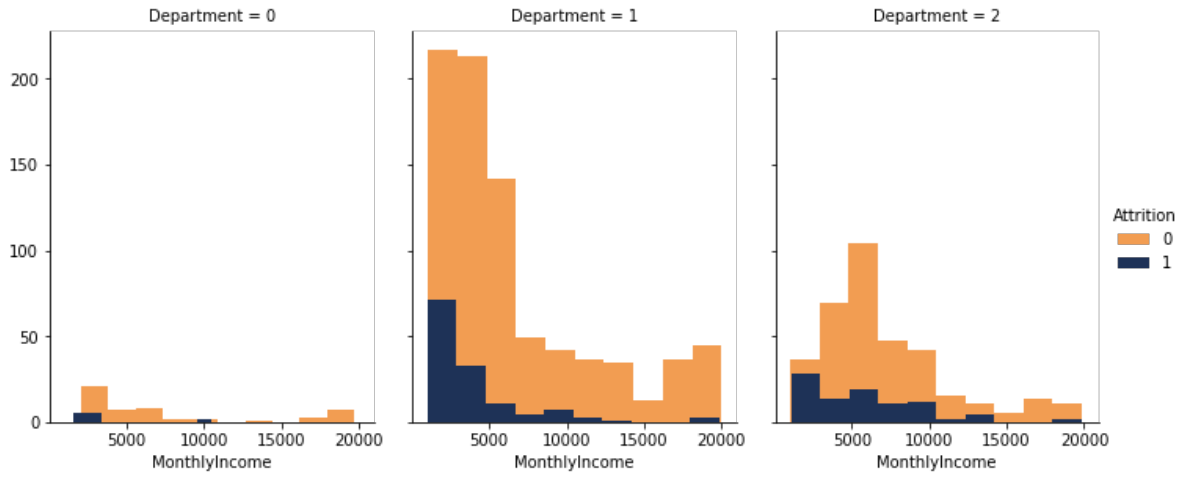
- Employee who travels often burns out more than others.



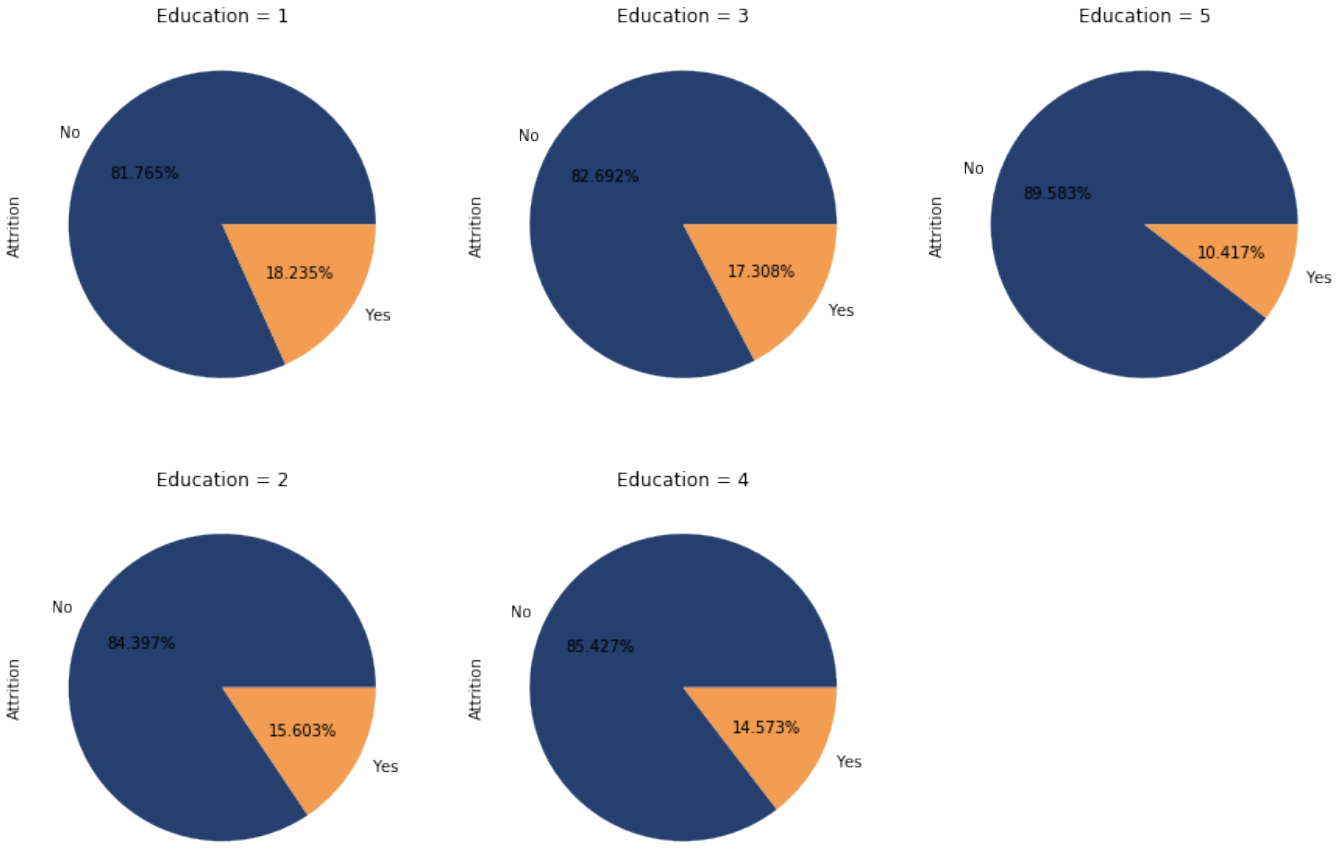
- People from Sales department left the company in case of attrition frequently from the others departments.



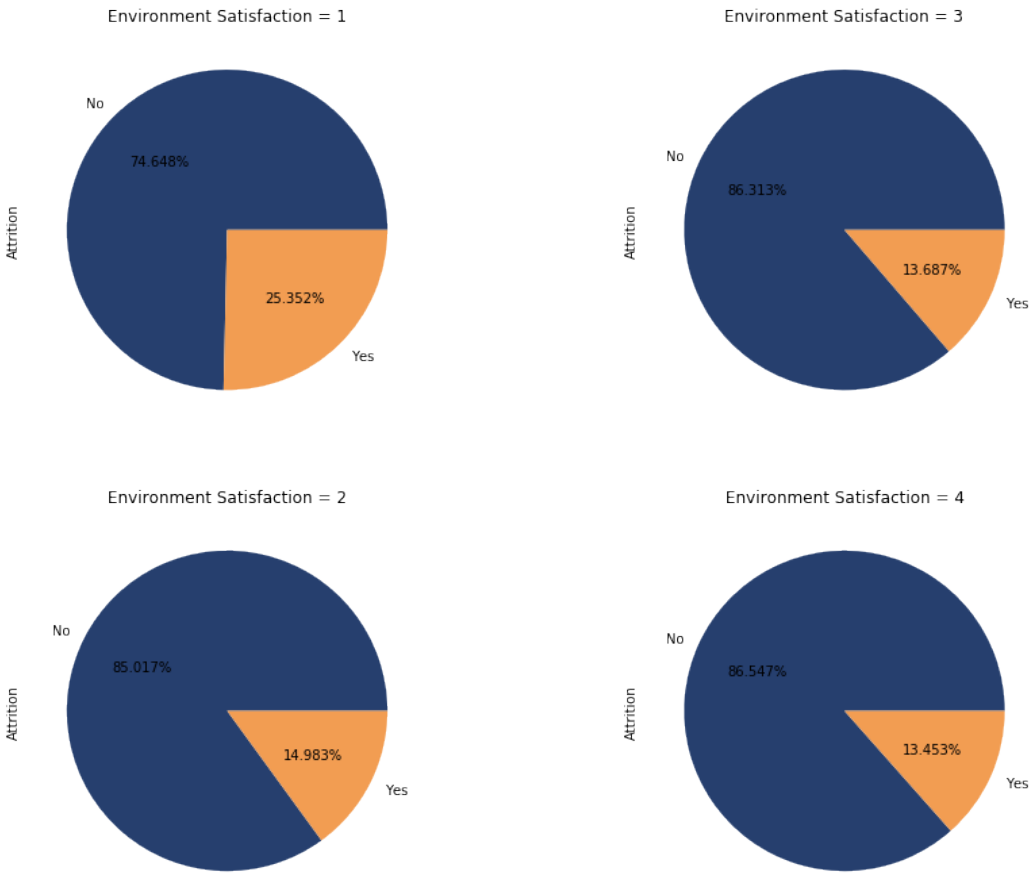
- Generally people go away with less income. the income of people who left the company often lies below 10,000.



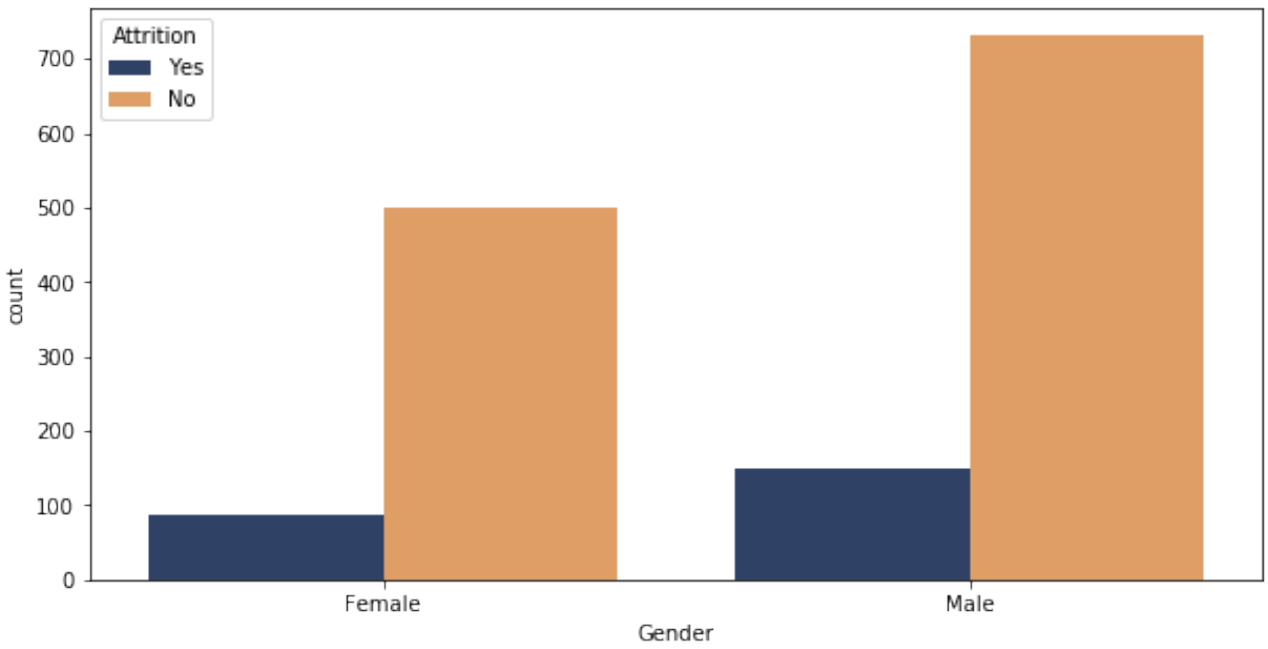
- People with Bachelor degree and with under college education have the biggest share of people who left. Category of People with Doctor degree has the smallest part of people with Attrition



- The level of satisfaction within the company shows that people who left the position frequent rated lower than those who continued to work.



- The majority of the company's employees are men. Also, men is tends to burn out more that Women.



Preparing data for model construction.

First step is to encode categorical features into numeric equivalent. In this work Label Encoder from sklearn library were used.

Since our data is not balanced, we can try some modern solution which is make data more balanced. We will use over-sampling and under-sampling methods solution for this.

All models will be done with three types of data just data, data with under sampling and data with oversampling.

In addition to it, logistic regression and support vector machine algorithm works with scaled data, that is why we have to use «Standard Scaler» method additional for this models.

Metrics

The most famous metric for the classification problem is accuracy, but it does not take into account the class imbalance, therefore we will use the F1 - score (harmonic mean between accuracy and completeness).

That is why we were used three metrics for model quality assessment : accuracy, RocAuc score, f1-score.

Models Implementations

We divided the data into 2 parts: training (80% off all objects) and test samples (20% of all objects) using «train_test_split» method and make a cross validation on train data and test stability of the model on test data. The models was constructed on three types of data : original, oversampled and under sampled.

1. Logistic Regression parameters : random_state=100, class_weight = {0:0.4,1:0.6}
2. Random Forest parameters : max_features = 10, n_estimators=1000, random_state = 63
3. Extra Trees: no special parameters
4. Support Vector Machine: np special parameters
5. Cat Boosting parameters : iterations: 1000, learning_rate : 0.7, random_state : 63, verbose: 0

6. Gradient Boosting Parameters: n_estimators: 1500, max_features: 0.9, learning_rate: 0.5, max_depth: 5, min_samples_leaf': 4, subsample: 1, max_features : sqrt, random_state : 63, verbose: 0.

Summary table of models results

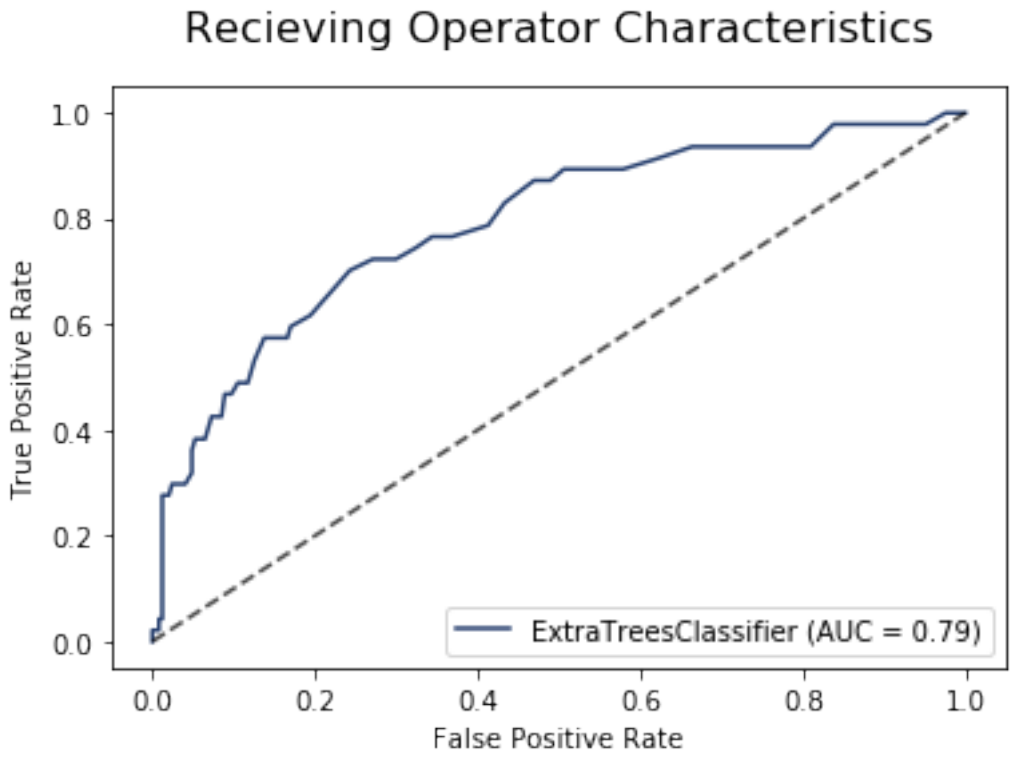
		Logistic Regression	Random Forest	Extra Trees	Support Vector Machie	Cat Bossting	Gradient Boosting
Cross Validation							
	Accuracy	0,856	0,860	0,852	0,855	0,852	0,861
	F1-score	0,465	0,294	0,56	0,250	0,379	0,380
	RocAUC	0,807	0,802	0,790	0,798	0,746	0,777
Test							
	Accuracy	0,863	0,850	0,863	0,840	0,874	0,738
	F1-score	0,523	0,290	0,285	0	0,463	0,483
	RocAUC	0,703	0,583	0,583	0,47	0,658	0,749
Cross Validation. Undersampling							
	Accuracy	0,710	0,715	0,731	0,584	0,671	0,683
	F1-score	0,730	0,697	0,726	0,639	0,666	0,665
	RocAUC	0,781	0,747	0,773	0,636	0,740	0,730
Test Undersampling							
	Accuracy	0,648	0,712	0,712	0,606	0,723	0,723
	F1-score	0,702	0,721	0,703	0,666	0,74	0,734
	RocAUC	0,648	0,712	0,712	0,606	0,72	0,723
Cross Validation Oversampling							
	Accuracy	0,817	0,914	0,933	0,586	0,916	0,924
	F1-score	0,825	0,913	0,930	0,620	0,916	0,923
	RocAUC	0,905	0,972	0,982	0,655	0,970	0,978
Test Oversampling							
	Accuracy	0,795	0,809	0,825	0,609	0,823	0,710
	F1-score	0,809	0,784	0,799	0,674	0,802	0,703
	RocAUC	0,795	0,809	0,825	0,609	0,823	0,710

So, the maximum result we have that the highest results we get with oversampling methods. The best models is Extra Trees, Cat Boosting and The Gradient Boosting. The most stable throw all testis Extra Trees Classifier algorithm. Mentionable, that Under-sampling scores is less than the original data almost in all models. It is logically understandable because, when we delete data from the dataset we decreased quality.The model with the lowest results is Support Vector Machine.

The results of the project

The highest predictive result of the project was Extra Tree Model on Oversampling data with value on metrics : Accuracy — 0,933, F1-score — 0,930, Roc Auc — 0,982. The highest result on original data was on Extra Tree Model with the value of metrics : Accuracy — 0,852, F1-score — 0,56, Roc Auc — 0,790.

Roc Auc on test original data on Extra Tree Model:



One more interesting fact that, features important method was made to all models, and we get that the features that have huge influence on prediction for all algorithms is: Job Satisfaction, Stock option level, Monthly Income.

Repositories with code

https://github.com/yaxenia/Prediction-of-employee-churn

Conclusion

To summarise all mention above, we can say that the problem of employee churn is relevant for many companies, because it is cheeper to save an employee rather than to hire a new one. The best result is mentionable on data with oversampling method. The best models are Gradient boosting, Extra Trees and Cat Boosting. Also, if we return to sources we can see, that almost all suggestion was right : Decision Trees and Gradient boosting except Logistic Regression models shows better results than others. The most influenced features is: Monthly Income, Job Satisfaction, Stock Option Level.

References

[1] Support Vector Machine https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[2] Random Forest and Forest Decision Tree - https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[3] Prediction of Employee Turnover in Organizations using Machine Learning Algorithms/ *Pankaj. Ajit, BITS Pilani Goa, India*
https://pdfs.semanticscholar.org/fa49/19810eaee67e851ad13775b78c94217a7908.pdf

[4]. Employee Attrition Risk Assessment using Logistic Regression Analysis/ *Rupesh Khare, Dimple Kaloya, Chandan Kumar Choudhary & Gauri Gupta Hewitt Associates*. India. p. - 53. https://pdfs.semanticscholar.org/1300/36abec4904111725d7b8c788b80e173aaeb2.pdf

[5]Employee Churn Rate Prediction and Performance Using Machine Learning/*Aniket Tambde, Dilip Motwani.*
https://www.ijrte.org/wp-content/uploads/papers/v8i2S11/B11340982S1119.pdf

[6] Metrics for Classification- https://github.com/esokolov/ml- course-hse/blob/master/2019-fall/lecture-notes/lecture04-linclass.pdf

[7]Cross-validation with upsampling data - https://kiwidamien.github.io/how- to-do-cross-validation-when-upsampling-data.html

[8]. Feature extraction - http://datareview.info/article/otbor-priznakov-dlya- mashinnogo-obucheniya-na-python/

[9]. https://github.com/kondratevakate/machine-learning-with-love

[10]. Алгоритм Oversampling- https://kite.com/blog/python/smote-python-imbalanced- learn-for-oversampling/

[11] Logistic Regression - https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc