Job Shifting Prediction and Analysis Using Machine Learning

Dr. Padmaja Pulicherla, Professor, Teegala Kirshna Reddy Engineering College, Hyderabad.

Tharun Kumar, Student, TKREC, Hyderabad

Naren Abbaraju, Student, TKREC, Hyderabad

Himanshu Khatri, Student, TKREC, Hyderabad

Abstract - In today's volatile employment structure, the employees tend to shift the job in an unexpected manner. In that case the company may face issues regarding scarcity of the workforce and find problem to reemploy quickly. Thus to overcome this problem we have designed a predictive model to anticipate the chances of an employee leaving the job. In this project the train and the test datasets are taken from Analytics Vidhya site where in the algorithm used to do the prediction are Random Forest, XGBoost, CatBoost, LightGBM out of which CatBoost has performed the best and ended up giving the most accurate prediction. The datasets provided by Analytics Vidhya were structured in nature but incomplete in observance thus to fill that the missing values imputation procedure had to be performed and then the data was fed to the algorithm for prediction. Knowing the employees approach towards job shift prior would actually help the company to plan out the workforce efficiently. CatBoost is a gradient boosting technique on decision trees library made available as open source by Yandex. It is universally applied across a wide range of areas and to a variety of problems. Considering accuracy, robustness, usability, extensibility Catboost as an upper hand over the other models.

Keywords: Logistic Regression, SGDClassifier, Decision Trees, Naive Bayes, Random Forest, AdaBoost, Gradient Boosting, XGBoost, Target, Experience, Last New Job, Size of Company.

1. INTRODUCTION

1.1 In the present day IT rush, the competition between many multinational companies is at a whole new level and these companies want their best employees to stay with them to sustain in the market. For this they have to know whether their employees are happy with their work and pay or are they willing to shift to a new company.

1.2 This made us to go for a research about how the above given problem could be solved. Through many documentation and cases, it worked out that data science and machine learning can make the work less requesting and faster.

1.3 using the features present in the dataset. The dataset for this is removed from the Analytics Vidhya site. With machine learning algorithms, using python as core we can predict the chances whether an employee will stay in the company or will shift to a new company.

1.4 The aim of the project would be to train a model for prediction. The model is trained on train data set which will be validated on test dataset. The CatBoost and other algorithms are used for prediction. Exploratory analysis of data is done to analyze the dependency of the target variable on
independent variables. This work would help the corporate industries to predict the chances of an employee shifting the job. By this result, company could infer and may take required action to make their best employees stay.

2. CONCEPTS OF THE PROPOSED SYSTEM

2.1 Predictive Modeling
It can be defined as training a model that can make predictions. This includes algorithms with certain parameters to make the predictions.

Predictive modeling is briefly classified into two types ie; Regression and Classification. The analysis in regression is based on variables and trends relationship to predict continuous variables.

Unlike backslide models, the errand performed by the gathering model is to assign discrete class imprints to explicit data regard as yield of a desire. Instance of a course of action show is - The precedent gathering undertaking in atmosphere deciding to anticipate a splendid, stormy, or bone chilling day.

Sorts of Predictive Modeling Algoritms

Choice Trees-A choice tree is a calculation that utilizes a tree molded diagram or model of choices including chance event out comes, costs, and utility. It is one way to display an algorithm.

Linear Regression –A Linear Regression is an Algorithm used to establish a relation between the dependent and the independent variables.

Ensemble methods

Random Forest
The name Random Forest can be justified by is working. A Random Forest is a collection of Decision Trees taking random inputs from the datasets. In case of regression problem the average of all the Decision Tree results is considered as final result and for classification problem the majority Decision Tree results is considered as final result.

CatBoost
It is an open source machine learning algorithm introduced by Yandex and is used to automatically handle categorical data by providing effective results without extensive training of model.

3. Data Preprocessing
This method fuses procedures to oust any invalid characteristics or unlimited characteristics which may impact the precision of the Model. The standard advances include Sampling and Cleaning process which
are used for removal or fixing of some missing data (there may be data that are insufficient). Inspecting is the methodology which is used to make the data appropriate, which may decrease the running time for the figuring.

1.1 Functional Diagram of Proposed Work

It might be parceled into 4 areas:

1. Exploratory examination of Data
2. Data treatment (Missing quality and special case fixing)
3. Data Modeling

Estimation of performance of Model

**Prepare Data**

1. In this movement we need prepare data into right
course of action for examination

2. Data cleaning

Change and Analyze Variables

We may need to change the variables using one of the approaches:

1. Institutionalization or organization

2. Missing Value Treatment

Looking at (Train and Test)

- Training Sample: Model will be set up on this precedent. 70% of the data goes here.
- Test Sample: Model displays will be authentic on this precedent. 30% of the data goes here

Cross Validation

Cross-endorsement is a technique for evaluating Machine learning models by means of setting up a couple of Machine learning models on subsets of the available data and surveying them on the correlative subset of the da entirety up a model.

Model Selection

Model selection is done for finding the best fit algorithm for the given dataset. We have to compare the modeling techniques. Such as:

- Logistic Regression
- Linear Regression
- SGDClassifier
- Decision Trees
- Naive Bayes
- Random Forest
- Ada Boost
- Gradient Boosting
- XG Boost
- CatBoost

Build/Develop/Train Models

- Validate the suppositions of the picked estimation
- Train Model on Training Sample data, which is the available data (Population)
- Check Performance Of Model: - Error, Accuracy

Support/Test Model
4. Implementation

The dataset used in the endeavor is taken from analyticsvidhya.com. The dataset obtained from Analytics Vidhya is kept up and invigorated by the Organizations.

The execution of the endeavor is segregated as seeks after –

4.1. Data gathering

Bad behavior dataset from Analytics vidhya is used in CSV position.

ta. Use navigate fitting, ie, fail to

4.2. Visualization

Using matplotlib library, job shifting data set is analyzed by plotting various graphs.

4.3. Data Preprocessing

18k entries are present in the dataset. The null values are filled using train.fillna(-9999, inplace=True) and test.fillna(-9999, inplace=True) in train and test files respectively. The categorical attributes (City, Gender, Relevant Experience, Enrolled University, Education Level, Major Discipline, Experience, Company size, Company type, Last new job) are converted into numeric using One Hot Encoding.

4.4. Feature selection

The attributes used for building the model are Enroll Id, City development index, City, Gender, Relevant Experience, Enrolled University, Education Level, Major Discipline, Experience, Company size, Company type, Last new job, Training Hours, Target. This helps in building the model

4.5. Building and Training Model

After feature selection all attributes are used for training. The dataset is divided into x_train, y_train and x_validate, y_validate. The model is imported from sklearan. The model is built using model. Fit (x_train, y_train).

4.6. Prediction

Using model the prediction is made. The accuracy is calculated using roc_auc_score imported from metrics.metrics.roc_auc_score(y_validate, predicted).

RESULTS:

The results are obtained by performing various steps involved in predictive modeling.

Data Visualization: Data preprocessing is done for cleaning of data and filling the missing values which creates impact in getting the accuracy of the model. Converting string variable into numerical and it under goes further process
Data preprocessing is done for cleaning of data and filling the missing values which creates impact in getting the accuracy of the model. Converting string variable into numerical and it under goes further process

1. In the wake of disengaging the instructive accumulation presentation getting ready set and testing set the model is readied using count as referenced in the table. The precision is resolved using the limit roc_auc_score imported from metric from sklearn. The exactness is referenced in the table underneath.

As ought to be clear from the results procured from the table the estimation which can be used for the insightful showing will be Cat Boost figuring with exactness of 0.6867 highest among the straggling leftovers of the count. The least which can be used will be Gradient Boosting. For further showing using unnoticeable data there is no prerequisite for using other estimation.

Recognition

This section deals with the analysis done on the dataset and plotting them into various charts like bar, pie, disseminate. We perform Bivariate examination on various fields in regards to target variable. Some of the examination done are:

1. No. of people with respective experiences have changed or not
2. No. of people with their count of previous jobs have changed or not.
3. No. of males, females and other gender people vs their decision of shifting.
4. Size of the company with respect to the target variable
This graph shows the no. of people with their respective experiences who have shifted or not to a new job. 1 represents that people have shifted and 0 represents not shifted. For example, 968 people with experience of 3 years have not changed to a new job and 191 people with the same experience have changed.

![Experience vs. Target](image)

**Figure 2- Experience vs. Target**

The graph below shows us the no. of people and their previous jobs with respect to the target. 891 people with 4 previous jobs have not changed to a new job while 147 people with same experience have changed.

| Algorithm          | Accuracy       |
|--------------------|----------------|
| Logistic Regression| 0.63951185426383 |
| Knn                | 0.542441587914122 |
| SGDClassifier      | 0.611997932785193 |
| Decision Trees     | 0.52901252585598 |
| Naive Bayes        | 0.629882752529369 |
| Random Forest      | 0.632975866909779 |
| Ada Boost          | 0.659027575078689 |
| Gradient Boosting  | 0.505516911105973 |
| XG Boost           | 0.623399203692683 |
| Cat Boost          | 0.686761106541168 |
The graph below is a comparison of gender and target which show the no. of male, female and other gender who have changed or not. Example, 11,306 males have not changed to a new job while 1578 have changed.

**Figure 5 – Gender vs. Target**

The graph below shows the size of companies and the decision of people working in the respective company. We can observe that more people have changed from companies with less size.
Figure 6-Size of Company vs. Target

The graph below shows the types of companies and the decision of people working in them. There are 5 types namely Pvt Ltd, Funded Startup, Public Sector, Early Stage Startup, NGO and other.

Figure 7- Type of Company vs. Target

Figure 8- Education Level vs. Target

5. CONCLUSION

This model when implemented in real time will let the companies to get an insight about the chances of an employee shifting to a new company based on which the company can take necessary action to keep their best representatives from leaving their company. Using the ideas of machine learning, we have manufactured a model utilizing train informational collection that has experienced information preparing. The CatBoost display predicts results with
precision of 0.6867. Information perception helps in breaking down the objective variable regarding each reliant variable. We created numerous diagrams and discovered fascinating insights that helped in comprehension the dependency of the target variable on rest of the variables.

REFERENCES:

The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman.

Programming Collective Intelligence by Toby Segaran.

Website Links:

https://tech.yandex.com/catboost/

https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/

https://www.analyticsvidhya.com/blog/2013/04/predictive-modeling-what-case-study-part-1/