HR Analytics Report

DataRobot

2019

Analysis of the Company Workforce & Recruiting Strategy

DataRobot Home Test | Darya Rudych

Overview

Context

HR data can be hard to come by, and HR professionals generally lag behind with respect to analytics and data visualization competency. Thus, I use HR-related mock data to address the questions that any medium to big size organization might have with respect to its HR and recruitment strategy. For the purposes of this analysis, I pretend the company I am analyzing is one of the offices of Data Robot.

Data

The data I am using for this analysis was found on Kaggle and includes the historic employee & recruiting data that contains the following:

- Employee names
- DOBs
- Age
- Gender
- Marital status
- Date of hire
- Employment Status (i.e. active, terminated, future start etc.)
- Reasons for termination (i.e new job, return to school, unhappy etc.)
- Department & Position title
- Pay rate
- Performance score
- Other

Basic Assumptions & Problem Statement

For the purposes of this analysis, we assume that DataRobot is projected to significantly grow within the coming year and is, therefore, preparing to increase its workforce to **300** employees. The company's senior management wants to gain visibility into the current employee portfolio and develop the most optimal recruiting strategy. It tasks its Analytics team to analyze historic HR data of the organization and come up with the solutions that will satisfy the following:

- Diverse profile of the organization (i.e. gender, race, age)
- High employee retention
- Reduced recruiting costs

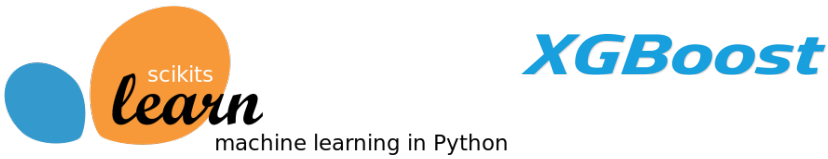
- Hiring of high-performing candidates

To ensure diversity and compliance with legal requirements, they set the following thresholds:

1. Gender disparity should be within 10% limit
2. Ratio of white to non-white employees should be 60% to 40%
3. Age should follow normal distribution

Tools & Technologies

To carry out the analysis, I am using the following:



Approach & Methodologies

Before analyzing the data, I use Python to explore and clean the core_dataset.csv. The data cleaning pipeline consists of the following steps:

1. Checking data types
2. Describing data using pandas
3. Checking for null values
4. Checking for data inconsistencies
5. Filling NaN values where necessary
6. Exploring correlations

Further data exploration is done in Tableau Desktop. I load the cleaned dataset and a dataset on recruitment spending into Tableau. Unfortunately, there's no way to join two datasets, but Tableau allows us to use unrelated datasets in the same workbook without enforcing joins or unions. I then perform analysis to produce 4 dashboards views, each exploring one issue as I have previously identified. These dashboards are:

1. Employee Diversity Dashboard (gives a snapshot of a current state of the organization)
2. Performance Dashboard (describes how different groups of employees are performing)
3. Turnover Dashboard (explores employment termination trends and reasons for termination)
4. Recruitment Source Dashboard (looks at effectiveness
5. Recommendation Dashboard (provides most insightful charts of the entire analysis)

Findings & Recommendations

Employee Diversity

1. Half of company's current employees are between 30-40 y.o. Roughly 17% is 50+ y.o. which is quite a significant number to lend a conclusion that company does not discriminate older people (Fig.1).

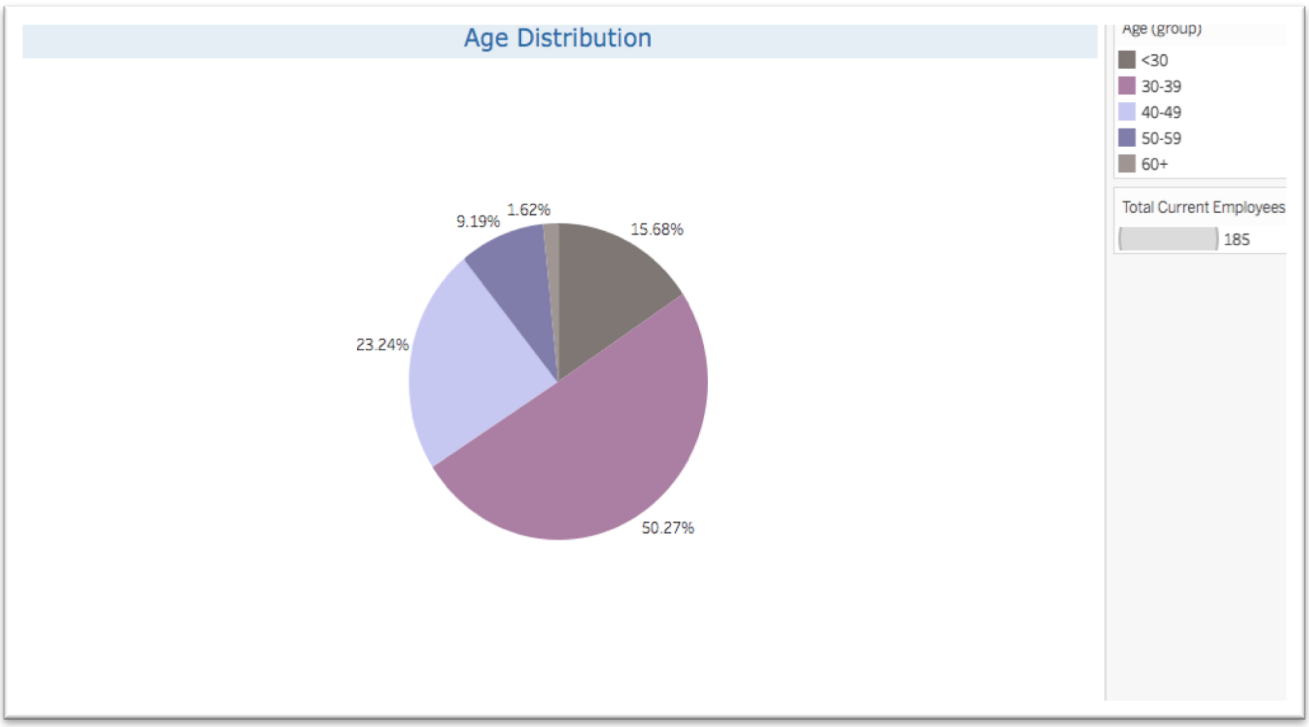


Figure 1

2. The company's employees are predominantly white. Since the company wants to keep the ratio of white to non-white employees 60% to 40% respectively, it should hire 61 white and 54 non-white people, assuming they want to bring total number of employees to 300 people (Fig.2).

Figure 2

3. The company has slightly more female than male workers. To keep the ratio even, it should hire 44 female and 71 male workers in the next recruiting cycle (Fig. 3)

4. 96% of company's employees are US citizens. Non-citizens that have ever worked in the company held either managing or highly technical positions. Perhaps, the company's policy is to hire only highly skilled and experienced non-citizens (Fig.4).

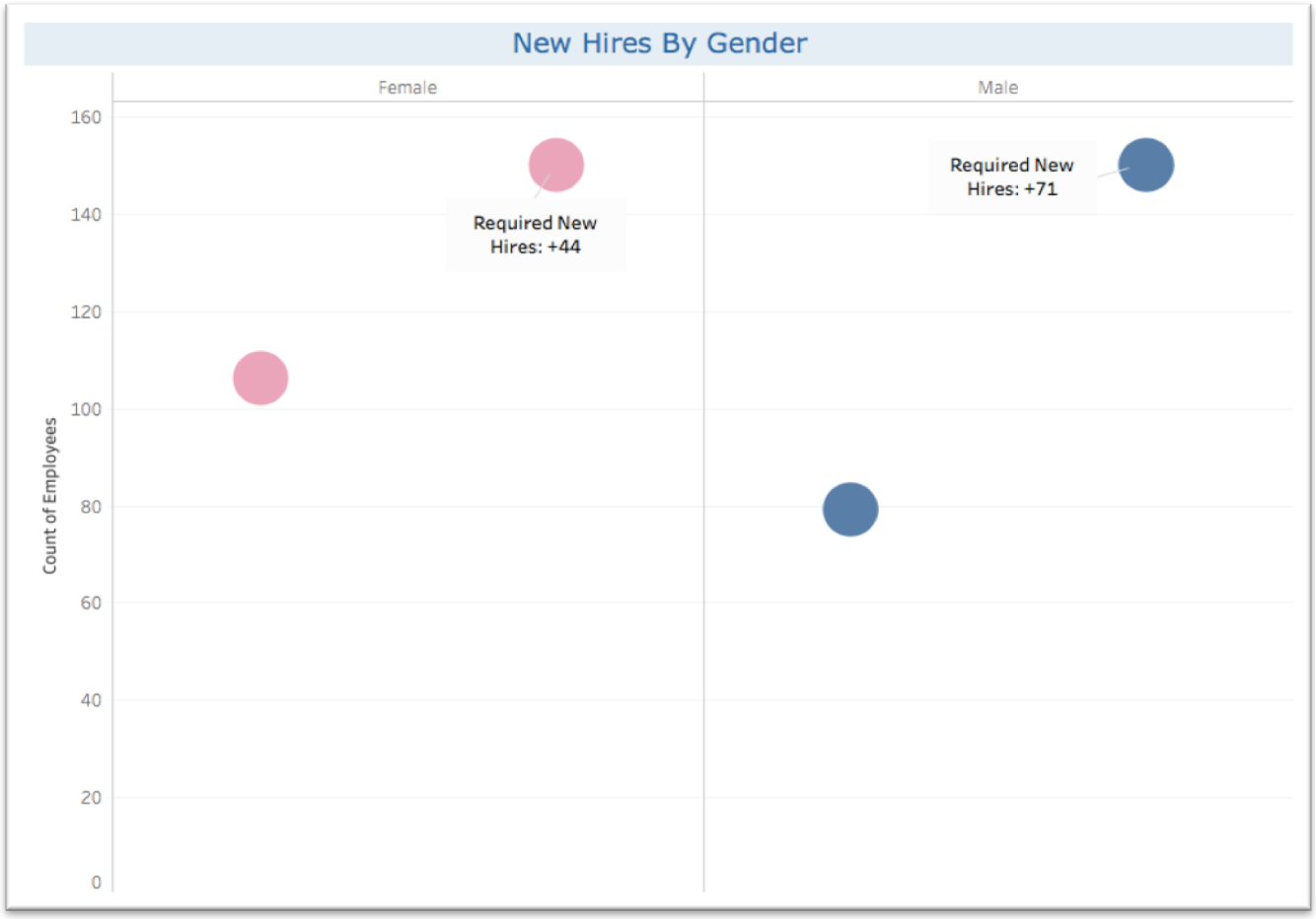


Figure 3

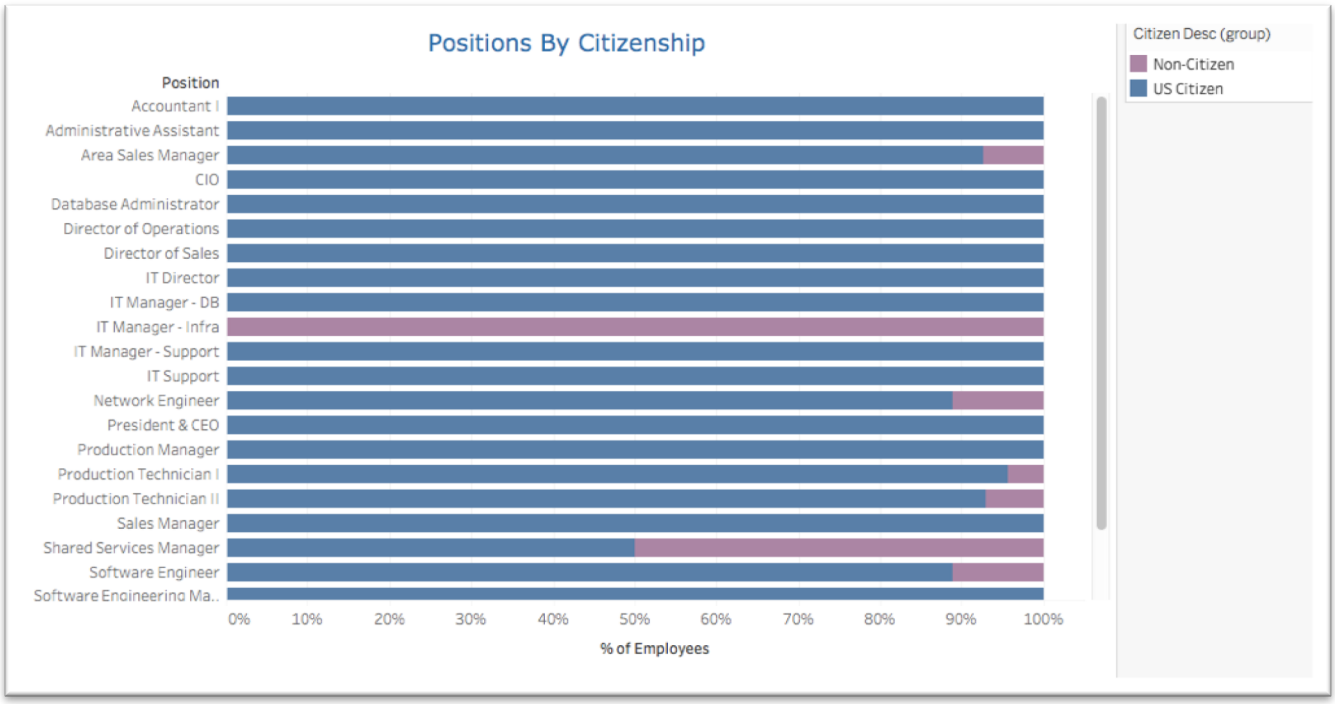


Figure 4

Employee Performance

1. Among different factors analyzed, there doesn't seem to one that has a significant visible effect on performance. However, some of the trends are: 1) the number of employees who don't meet performance requirements decreases as the age increases (relation age vs. experience); 2) most high performing employees fall under 30-39 age group (Fig.5).

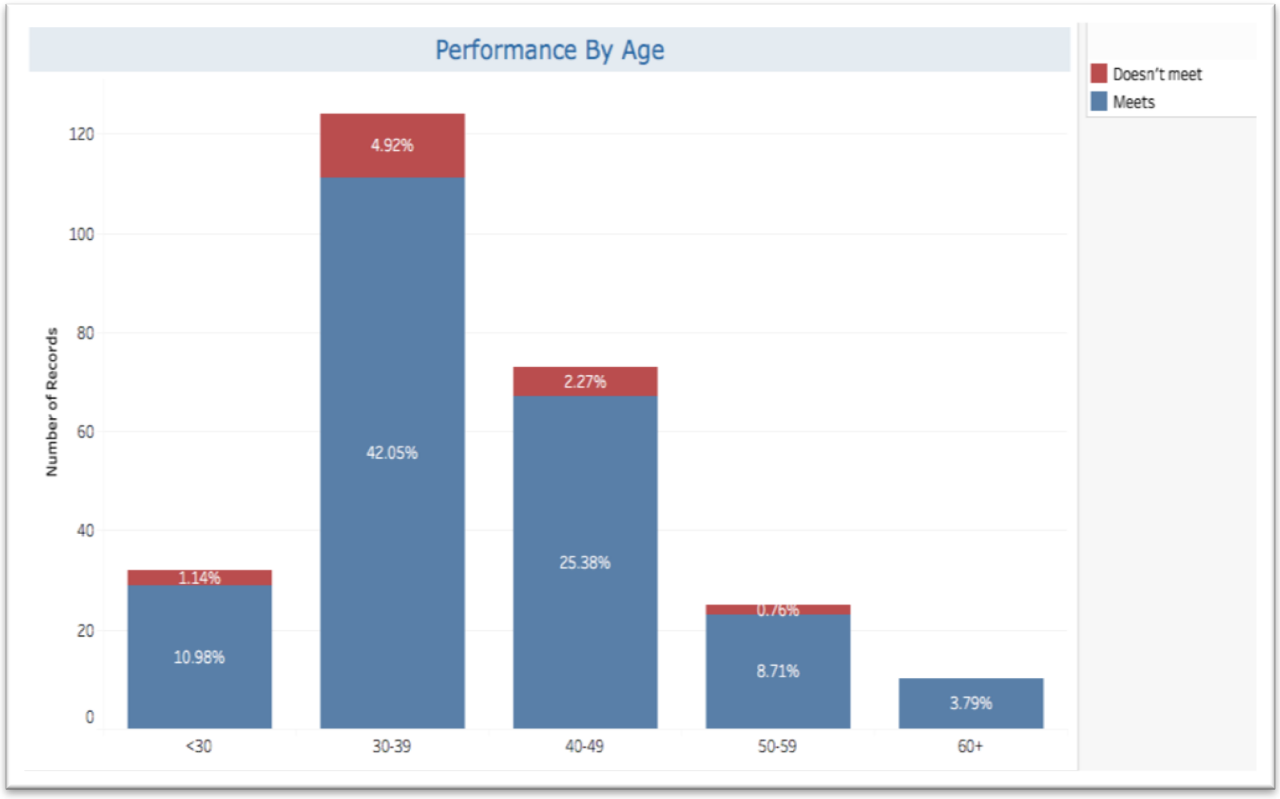


Figure 5

2. Further investigation of performance shows that employee referrals result in the highest number of high performing employees. Next come Newspaper/Magazine, Vendor Referrals, Facebook, and On-campus Recruiting. The company should prioritize these recruitment sources to target high performing employees (Fig. 6).

3. Among other factors in the dataset, department an employee works for was found to be most closely associated with the performance. This chart demonstrates this trend and indicates that Sales and Production have, historically, lower % of people who meet the requirements. Perhaps, this is associated with people who manage these teams or high performance targets in comparison to other departments. The HR department should collect more data on specific departments to investigate performance further (Fig.7).

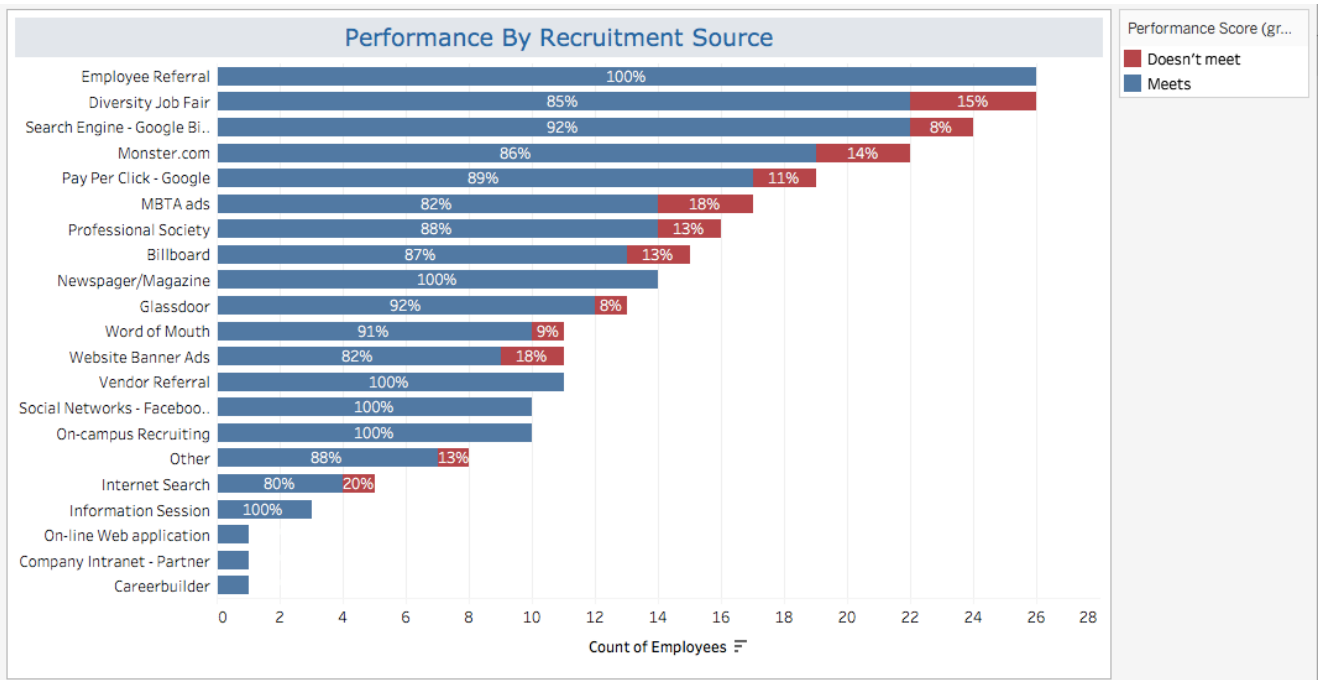


Figure 6

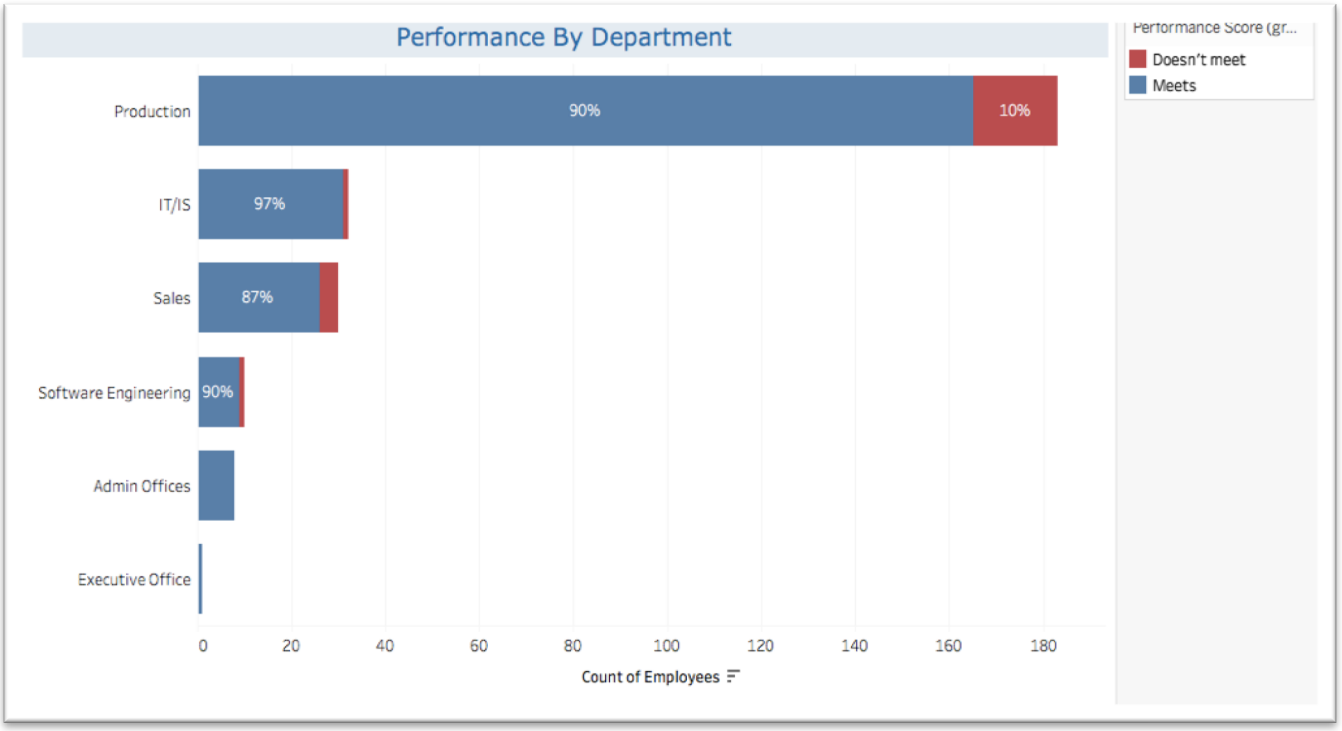


Figure 7

Employee Turnover

1. While 100% of people in the ages of 60+ meet performance requirements, they are the first to leave, which is quite natural. On the other hand, people under 30 y.o. have high performance rates and low termination rates. Empoyees in the age groups of 30-39 and 50-59 y.o. have roughly the same termination rates but the latter has higher % of well performing employees. Therefore, for the next recruiting season company should prioritize hiring people under 30 y.o. and people between 50-59 y.o. (Fig.5, Fig.8).

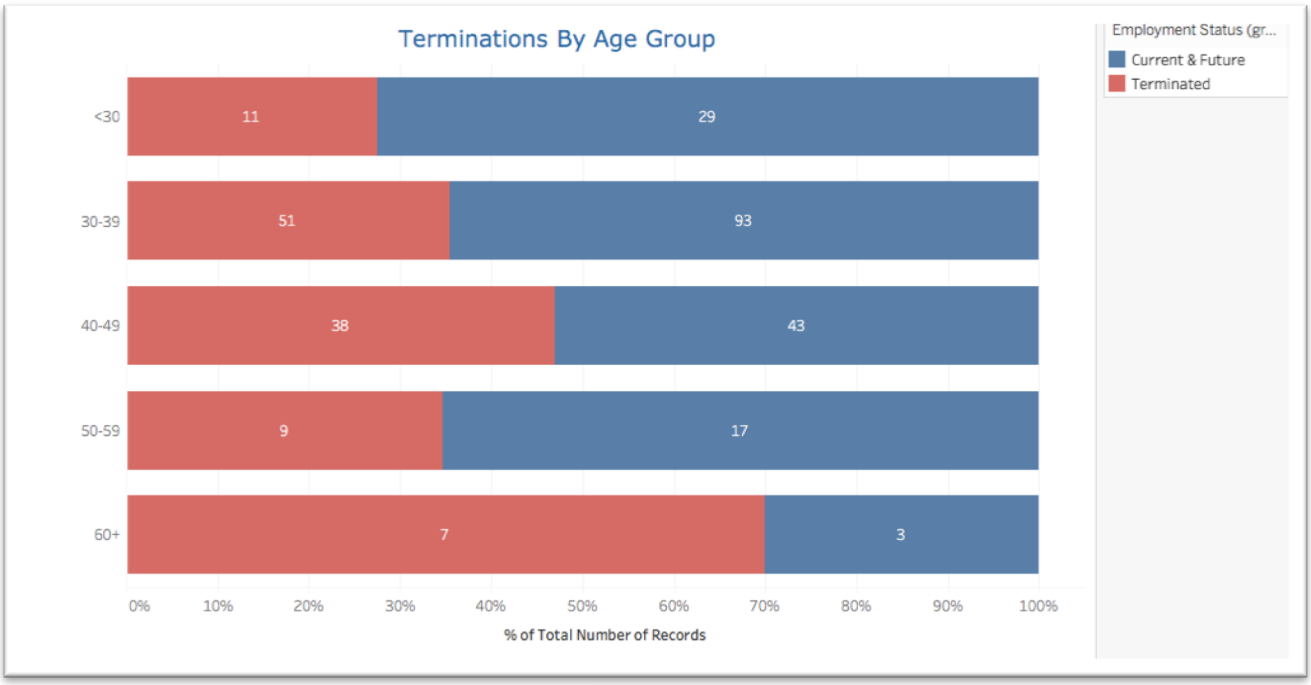


Figure 8

2. Since the company has very low % of employees who are non-citizens it should consider hiring more because non-citizens, as the chart suggests, tend to stay longer with the company (Fig.9).

3. The number of months seems to be closely associated with the likelihood of employment termination. Most people leave within the first 10 months of employment with the company. Employees between 30-49 y.o. tend to be the group that is more likely to leave no matter how long they have been with the company (Fig.10)

4. To increase retention rates the company might want to investigate further the reasons for employment termination. As analysis suggests, employees are enticed over to other company that offer them better position and pay. Company's HR should also be alarmed by the number of people who left due to "unhappiness". This is something that needs to be investigated further.

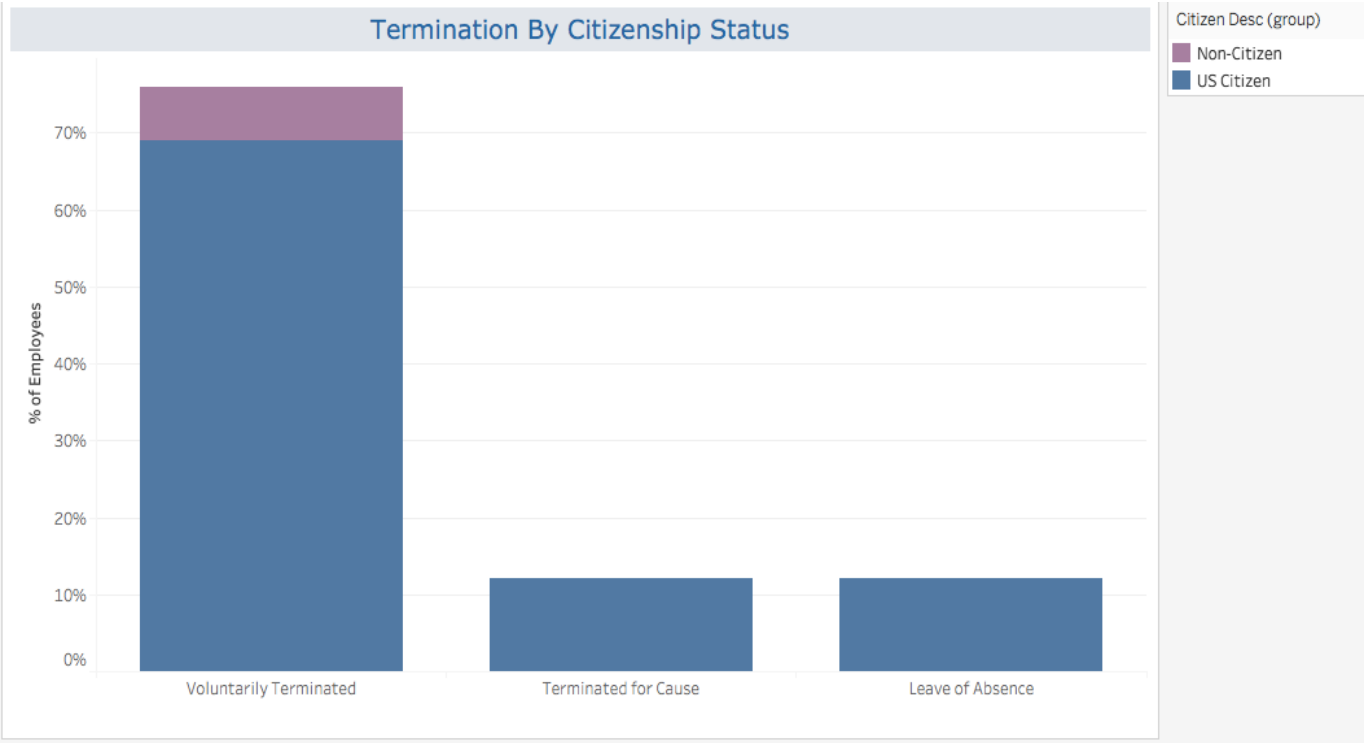


Figure 9

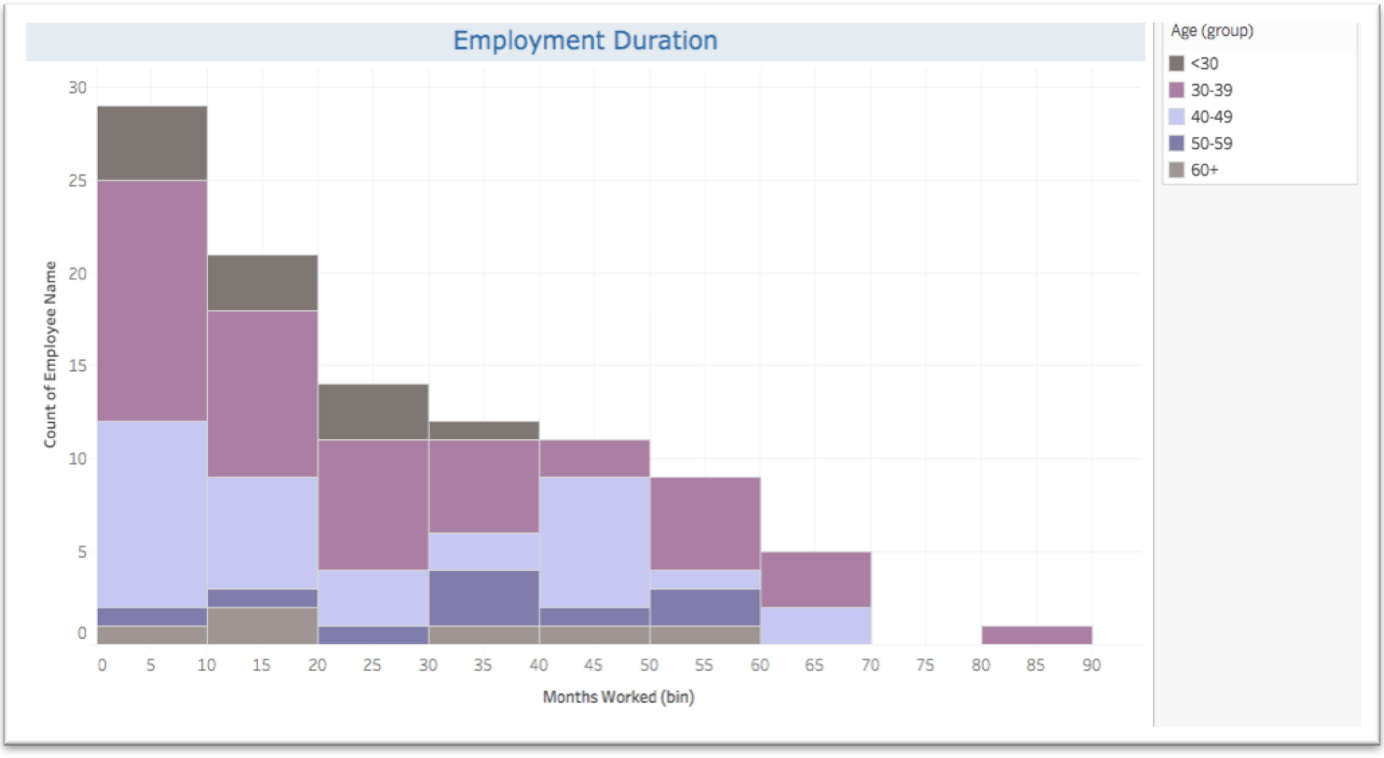


Figure 10

Recruitment Source Effectiveness

1. The scatter plot (Fig. 11) suggests that the company should leverage recruitment sources that are positioned in the upper left corner of the chart (i.e. the sources that are least expensive and have attracted highest number of employees). Next, the company should prioritize sources in the upper right corner of the chart - the sources that are costly but generate a lot of hires. So far, Employee Referrals and Diversity Job Fair are two most effective resources. The company should definitely cut spending on recruitment sources that yield low number of hires (BillBoard, Banner ads, Social Networks, MBTA ads, On-capus recruiting etc) or redistribute its budget towards more effective sources. For instance, it could stop spending on sources that have attracted less than 15 employees and introduce referral bonuses instead. The Employee Referrals is the best source also because 100% of employees that were referred have met the performance requirements (Fig.6)

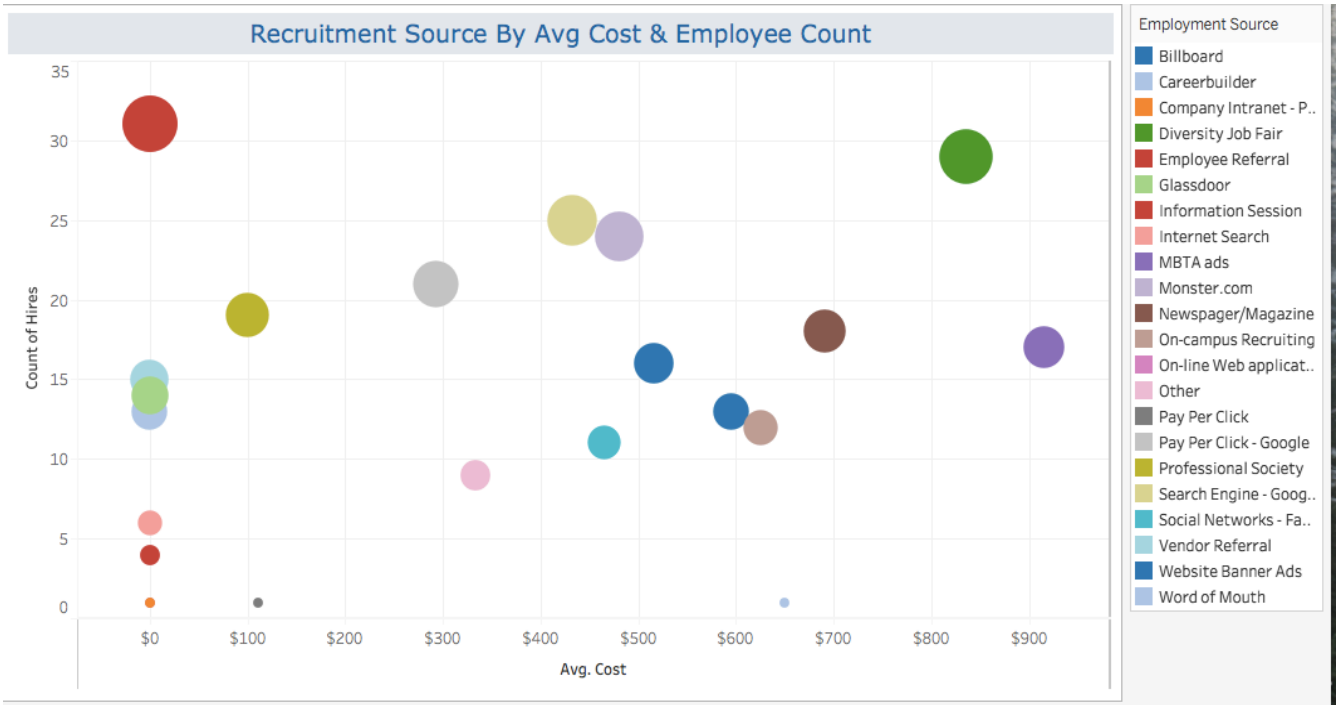


Figure 11

Predicting Performance

Since the company is interested in mitigating the risks of hiring people who will not meet performance requirements, I use company's data to build a decision tree model to see if we can accurately predict one's performance. I encode Performance Score as 0 = doesn't meet requirements and 1 – meets requirements to make it a binary classification problem. I build two different models (regressor and classifier) and evaluate their accuracy.

Model 1: Sklearn Decision Tree Regressor

This model uses linear regression and outputs a predicted number. It uses MSE to measure the quality of a split and minimize the loss using the mean of each terminal node (Fig. 12)

```
In [26]:
#Build and fit decision tree model
dtree = DecisionTreeRegressor(random_state=123, max_features=len(enc_data.columns))
dtree.fit(X_train, y_train)
```

```
Out[26]: DecisionTreeRegressor(criterion='mse', max_depth=None, max_features=15,
                               max_leaf_nodes=None, min_impurity_decrease=0.0,
                               min_impurity_split=None, min_samples_leaf=1,
                               min_samples_split=2, min_weight_fraction_leaf=0.0,
                               presort=False, random_state=123, splitter='best')
```

Model accuracy is 85.4% and R^2 is 1.0. The high R^2 score basically means that we are overfitting data and match it too closely to the training data.

Model 2: XGBoost Decision Tree Classifier

XGBoost library implements machine learning algorithms under the Gradient Boosting framework known to be more efficient and flexible than its alternatives. In my model, I use XGB model to build a decision tree that uses logistic rather than linear regressor and outputs probability.

This model's accuracy is 89.02% and RMSE is 0.31 which indicates a relatively good fit. Overall, this model is more accurate and proves that logistic regression is better suited for classification problems even when the outcome is binary.

In [34]:
```
# Compile an XGB model
xg_class = xgb.XGBClassifier(objective ='binary:logistic', colsample_bytree = 0.3, learning_rate = 0.1,
                n_estimators = 20)
```

In [35]:
```
#Fit the model to train data
xg_class.fit(X_train,y_train)

#Compute accuracy and rmse
preds = xg_class.predict(X_test)
accuracy = accuracy_score(y_test, preds)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
rmse = np.sqrt(mean_squared_error(y_test, preds))
print("RMSE: %f" % (rmse))
```

```
Accuracy: 89.02%
RMSE: 0.331295
```

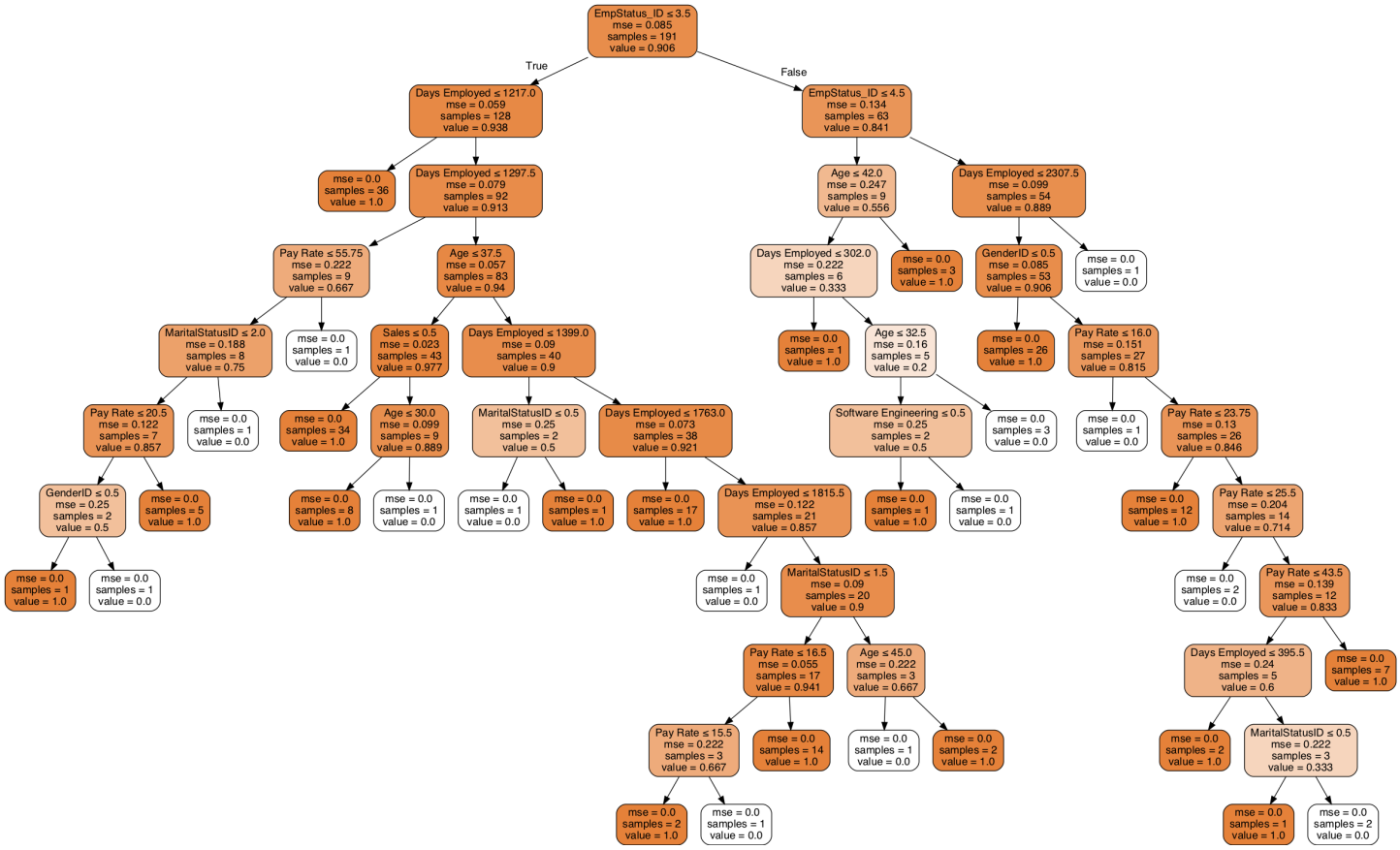


Figure 12

Findings of predictive modeling

Using XGBoost library we can also compute feature importance. This is useful to identify which feature determines performance the most. According to Fig.13 , number of days employed, pay rate, and employment status are the 3 factors have the most significance in predicting one's performance.

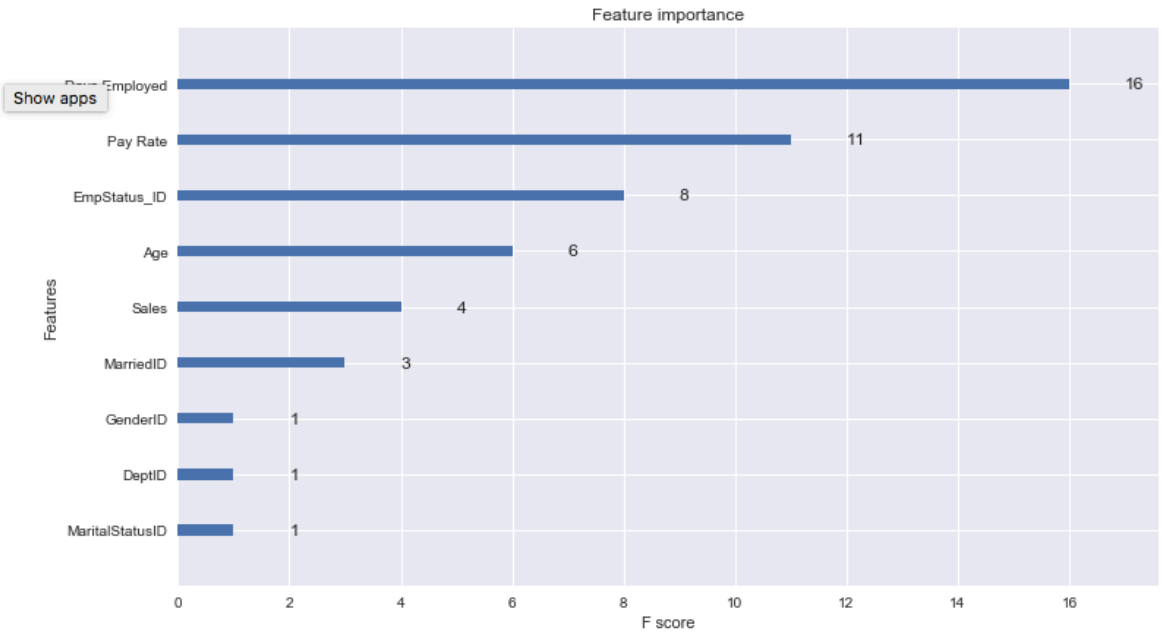


Figure 13

Final Recommendations

Based on the performed analysis of employee diversity, past performance, employment termination, and recruitment sources, it is recommended that the company:

- Develops better employee onboarding & retention program to ensure people do not leave before 1 year of employment
- Hires more non-white workers, males, and non-citizens who are under 30 years old to keep company's profile diverse and maximize performance across company
- Prioritizes employee referrals and introduces referral bonuses to encourage current employees to attract new hires
- Offers better growth/career opportunities to discourage employees from leaving the company