**Project title – Employee Turnover – Analyzation, Visualization and Prediction**

**Team members** – **Divya Dhaipullay**

**Abstract:**

This data visualization project mostly focuses on understanding what major factors generally contribute to employee turnover. Following it, analyzing, visualizing and create certain prediction models like Logistic Regression, Random Forest, KNN, Gradient boost, XGBoost to find any useful patterns in the employee behavior and draw meaningful insights from the plotted visualizations. Employee turnover can lead to lot of effort, be in terms of human resources, capital, or business performances. The turnover can also lead to reduced efficiency in projects completion and can get harder to control.

**Introduction (motivation, background, and objectives):**

**Introduction, Motivation and Contribution:** The motivation this data visualization project holds is to develop a model which strongly focuses on using visualization tools and techniques to predict the likelihood if a certain employee will leave the company or not and plan the retention of good employees so that the businesses work better and there is a greater cash flow which becomes an obvious reason for giving caution or care. This visualization project has various interesting graphs, analysis of every graph plotted, its model prediction and strategies for future are “solely” based on the plotted visualizations. The reason to get such prediction is to retain a quality work force and resourcing them to heavy technical and high-cost or revenue generating projects which generates a lot of money and enhances company’s value. The motivation behind choosing a simpler dataset is to more time in plotting the visualizations rather than cleaning the data itself

**Approachable Problem Description**

* Why do employees often leave the company?
  + Major reasons behind the customer leaving company could be
    1. Bad work culture within the company leads to no satisfaction among employees
    2. Employees may use their skills to get better offers from the company's competitors
    3. Lack of recognition for the service the employees provide
    4. Some individual specific unknown reasons like personal problems, health issues etc.
* How do we find the employee turnover?

1. Tracking their wellness, professional growth along with work
2. By analyzing complains from the employees
3. And also analyzing the competitors offers

* But how do we prevent employees from leaving a company?

1. For high-risk employees, we need to accommodate retention plans
2. Improve wellness services
3. Provide various other perks
4. And provide appraisals in addition to giving recognition and credit

**Key evidence and studies:** According to myaliat

According to 2018 data from the US Bureau of Labor Statistics, over 3 million employees have left their jobs voluntarily every month this year. In the Gallup 2017 State of the American Workplace report it was found that only 33% of US employees are engaged in the workplace. In a 2016 survey by SHRM, the average overall turnover rate was 18%. That’s nearly 2 out of every 10 employees in America. The cost of replacing an employee may be greater based upon position and skill level.

The Center for American Progress analyzed 30 case studies on the cost of turnover and found:

Entry level positions, or positions with an annual salary of around $30,000 or less, typically costs about 16% of the salary to hire and train a replacement. Mid-level positions, or those with an annual salary of around $30,000 – $75,000, typically cost about 20% of the salary to hire and train a replacement.

Key positions which require specialized training, or those which are well compensated and have high educational requirements, can be very expensive to replace. Stats show that a massive 213% of the salary of that employee is used to replace him.

Based on the numbers from the Center for American Progress above, you can calculate the potential cost to replace an employee by taking into consideration the position level and compensation for a specific job role.

Entry-level: Annual salary × .16

Mid-level: Annual salary × .20

Key position: Annual salary × 2.13

Total Turnovers ÷ Average Number of Employees × 100 = Turnover Rate

**Image courtesy** :https://myaliat.com/news/high-cost-employee-turnover-employer-know/

Graphical user interface

Description automatically generated with low confidence

Source : My Jupyter Notebook. 24% of the employees that left the company, it still is a greater number, since replacing 24% percentage of people is a greater matter for the resourcing team

**Critiques from Existing Work:**

**Example 1:** <https://thecleverprogrammer.com/2020/07/12/employee-turnover-prediction/> here is a predictive model built without even visualizing the data, it is too hard to believe that one can build a prediction model without even visualizing it. This is a terrible thing to do. Once should always learn about the data through their plotted visualizations to learn well about it and then portray it to the audience to gain wider attention, else it will become mere waste of resources in any complex projects.

**Example 2:** <https://medium.com/@aaysbt/employee-churn-analysis-be94751e4df5> The color combinations chosen here are bad, There is a lot of clashing amongst the labels of both bar plot and pie chart. The Pie chart doesn’t clearly help us read the data. Moreover, does not easily reveal exact values. Pie charts can easily be manipulated to yield false impressions

Chart, bar chart

Description automatically generated

**Example 3:** <https://www.autogrow.co/how-to-avoid-bad-hires/> This graph holds the same problem as that of the 50 shades of blue from our lecture, too hard to interpret the details. The below pie chart is also on the same lines, where we would definitely not be able to identify the section of pie chart based on the color of the legends

Map

Description automatically generated with low confidence

Chart, pie chart

Description automatically generated

The above plot uses indistinguishable colors which is hard to interpret. Sometimes pie charts might not sum up to 100

**Data:** <https://drive.google.com/file/d/1jEzDZX3XBPo-bebmioXtwk2aDehYrV8p/view>

**Contributions:**

The existing work for Employee Churn Analysis and Prediction in the field of ML mostly concentrates on using

1. Few basics algorithms like Logistic Regression, Decision Trees, and Gradient Boost which don’t clearly give all the required features in detail useful in prediction, whereas the approach used in this project is unique as it uses models like XGBoost to give the features description in exact with greater prediction visually as well.
2. The finer details that I have included in this project like visually showing the results and conclusion using the Churn Probability Plot is quite unique.

Chart

Description automatically generated

1. We did visualize the data by using RANDOM\_NOISE, so that we don’t get carried away by misconceptions of the results that we found. Despite adding noise we still found the results to be as similar to not plots that we got without adding noise and the plots that we drew for feature importance using shap package

Chart, bar chart

Description automatically generated

1. Based on the feedback given in the project proposal I have implemented certain graphs using shap package which turned out to be savior in terms of understanding the predictions visually

Chart

Description automatically generated

Chart, bar chart

Description automatically generated

1. The solutions, assumptions and the future work definition shown in this project is novel as well
2. The comparison of feature importance plots have been made suing both shap and yellowbricks packages which is new too.
3. Using both lmplot and KNN we compare the results and find them to be similar

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

**Data & Methods:**

**Process of analysis of data:** Using **Pandas** and **Numpy** I have analyzed the basic summary of the data I chose

1. **Summary stats**

A picture containing table

Description automatically generated

1. **Percentage of left employees and stayed employees**

Graphical user interface, text, application, email

Description automatically generated

1. **Department wise stayed and left**

Graphical user interface, text, application, email

Description automatically generated

1. **Correlation Data**

Graphical user interface

Description automatically generated with low confidence

1. **Data Frame for Area under the Curve**

Graphical user interface, text, application, email

Description automatically generated

**For greater details please refer the Jupyter Notebook**

**Ideas:** Based on the analyzed data I have come up with certain plots required like the

1. Various histograms for understanding the distribution of the data

Graphical user interface

Description automatically generated

1. A few bar plots to understand the data in greater detail

Chart

Description automatically generated with medium confidence

1. Understanding the top 5 departments to retain the employees from those departments Chart

   Description automatically generated
2. Some basic comparison graphs to understand the employee details

Table

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Chart

Description automatically generated

1. For salaries vs the percentage left or stayed

Chart, bar chart

Description automatically generated

1. Pairplots and heatmaps to understand the correlation

A picture containing chart

Description automatically generated

Chart

Description automatically generated

A picture containing calendar

Description automatically generated

1. Barplot for indicating correlation

Chart, waterfall chart

Description automatically generated

1. Some KDE plots to understand the comparisons of the data

Chart

Description automatically generated

1. Using KNN to plot the graphs to understand various categories of employees Scatter chart

   Description automatically generated
2. Plotting a few ROC- AUC curves

Chart, line chart, scatter chart

Description automatically generated

**Sketches and visualization methods selection:** I have used certain visualization packages like Matplotlib, Seaborn, Shap, Yellow Brick.

1. **Matplotlib:**  To draw certain basic Visualizations for that data
2. **Seaborn:** For data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics
3. **Shap:** SHAP (SHapley Additive exPlanations) is a game-theoretic approach to explain the output of any machine learning model.
4. **Yellow Brick:** Yellowbrick is a machine learning visualization library. Essentially, Yellowbrick makes it easier for you to.
5. Select features.
6. Tuning hyperparameter
7. Interpret the score of your models
8. Visualize text data

**Prototypes, Candidate visualization methods, Visualization methods selection:** On my Jupyter notebook, at every relevant plot I have explained why I used the plot and what I inferred from plotted the graph with its objectives. But as a summary of the project, I have attached a few visualizations above.

**Failed experiments:**

I did try initially plotting with Neptune and dtreeviz but was harder for the inexperience I had in developing the visualizations, I rather tried shap (suggested by the professor in the project proposal feedback) and yellowbrick for visualizing since I had learnt them earlier. There could be a broader scope to add a few more retention plans or stronger results and conclusions, as a 1 person this was the maximum I could do

**Results:**

**Novelty and Insights:**

1. The features that we predicted work well for Decision Tree Classifier are shown below

A picture containing bar chart

Description automatically generated

1. The features that we predicted work well for Random Forest Classifier are shown below.

Chart, bar chart

Description automatically generated

1. The features that we predicted work well for Gradient Boosting Classifier are shown below.

Timeline

Description automatically generated with low confidence

1. The features that we predicted work well for XGBoost Classifier are shown below

Chart, bar chart

Description automatically generated

1. Comparing the feature importance plots for XGBoost both shap and yellow bricks packages

Chart

Description automatically generated with medium confidence

Graphical user interface, application

Description automatically generatedChart, bar chart

Description automatically generated

Chart

Description automatically generated

1. Churn Probability Plot for the Test Data

Chart

Description automatically generated

We can conclude from the prediction that most of the employees would stay with around 97% probability. However, there will still be a few employees who would exit from the company, this can be told with the same probability.

**Conclusion**:

**Craftsmanship and Details:**

In the dataset that I got I just got a few factors that affect the employee churn, while in actual there are many others which need to be thought about like the location and complexity of work, working conditions, gender discrimination, client relationships with the company, competitors offering benefits, the industry in which the company belongs, whether the salary is credited on time or not.

In this project, I did try to divide employee churn analysis and prediction problem into various steps like exploring, profiling, plotting visualizations, clustering, various model selection, many evaluations and a few retention plans. Based on such detailed analysis, resourcing team can get an alert to analyze high risk employees before they exit. This model is people oriented, we should refrain from consider it as a typical numeric model. Instead, we can use its probability and treat each case specifically.

Based on these models’ comparison and evaluation metrics, the up sampling that we did on the given data works better mostly during training process, we are skeptical that it might not work this well on the unseen data (based on log loss score). One of the various reasons might tend to appear that there might be data leakage in cross\_val\_score step.

**Discussion:** The calculated log loss for the original dataset remains same with training dataset as well as testing dataset.

Chart, sunburst chart

Description automatically generated

* **Safe Region or New Normal (Green)** – Employees of this region are safe and are not planning to leave.
* **Low Risk Region (Yellow)** – Employees of this region are planning to leave and have some turnover.
* **Medium Risk Region (Orange)** – Employees of this region are at risk of turnover. Remedies to be taken immediately.
* **High Risk Region (Red)** – Employees of this region are highly planning to leave and have the highest turnover. Remedies to be taken immediately.

**Possible Solutions for Retention:**

1. To save employees from leaving, we would target certain departments and rank employees by their probability of leaving, then allocate a limited incentive budget to the highest probability instances.
2. Develop learning programs for managers, then use analytics to gauge their performance and measure progress.

* Be a good coach. Empower the team and do not micromanage
* Express interest for team member success
* Help team with career development

**Assumptions:** We don't know if the employees that left are interns, contractors, full-time, or part-time. These are important variables to take into consideration when performing a machine learning algorithm to it. There might be a bias in evaluation. Evaluation can be subjective, so can be erroneous too, depending on the evaluation metrics

**Future work:** With the help of various interesting libraries and packages for visualizations like Neptune, comet, Tensorboard or MLflow. Also, in the KNN section, a few K-Elbow plots could be added to make the work a lot more interesting.

**References-**

**Data -** <https://drive.google.com/file/d/1jEzDZX3XBPo-bebmioXtwk2aDehYrV8p/view>

**Problems of turnover:** <https://www.accountingweb.com/practice/practice-excellence/three-problems-caused-by-employee-turnover-and-implications-of-an>

**Visualization methods -** <https://pypi.org/project/yellowbrick/>

**Critiques - Example 1:** <https://medium.com/analytics-vidhya/employee-attrition-prediction-df77255610ef>

**Example 2:** <https://thecleverprogrammer.com/2020/07/12/employee-turnover-prediction/>

**Example 3:** <https://towardsdatascience.com/predicting-employee-turnover-7ab2b9ecf47e>

**Example 4:** https://www.autogrow.co/how-to-avoid-bad-hires/

**Image courtesy** :https://myaliat.com/news/high-cost-employee-turnover-employer-know/

**Shap:** <https://github.com/slundberg/shap>

https://towardsdatascience.com/introduction-to-yellowbrick-a-python-library-to-explain-the-prediction-of-your-machine-learning-d63ecee10ecc