**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 to find any useful patterns in the behavior of an employee through drawing meaningful insights from the plotted visualizations. nThe turnover can also lead to reduced efficiency in projects completion and can get harder to control. This project also aims at providing retention plan to enhance the business growth.

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

The motivation this data visualization project holds is to develop a model which strongly focuses on visualization tools and techniques that predicts the likelihood of a certain employee that might or might not leave the company and plan accordingly in the improvisation of strategies that lead to 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 nvalue.

**Critique** – **Example 1:** <https://medium.com/analytics-vidhya/employee-attrition-prediction-df77255610ef> in this article “salary vs satisfaction level” in step3 , multiple box plots are shown which is too hard to understand or comprehend. Hence it is easier to show the same information as subplots of various distplot, bar or histogram, so that the comprehension would be better for the audience. The same it is with step 4, “Salary vs Monthly hours spent” is shown as various box plot and there are so multiple outliers lying on the outer whisker of the plot making it messier to perceive. The same is happening in multitude of other plots consecutively in the article and to perceive its notion is harder. I believe substituting it with an easier form of plot like distplot is better, which I will be using in my project. When the data visualization is itself so misinterpreting for the observer then it is complex to build any related work on, like creating some prediction models like in my project case.

Chart, histogram

Description automatically generated

Additionally, here in the plot the legend is clashing completely with the plot itself, which gives a bad thought to the seeker about the interest that creator is possessing in their work.

Chart, box and whisker chart

Description automatically generated

**Example 2:** <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 once 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 3:** <https://towardsdatascience.com/predicting-employee-turnover-7ab2b9ecf47e> the most worthful plot that I found here is the one for feature importance, which tells a lot about which feature contributes the most based on which relevant prediction in this project can be made. I wish there had been more content on understanding data through visualization rather than hiding it, on the visualization wise in this article else the article does an outstanding job for prediction.

Chart, histogram

Description automatically generated

**Questions or objectives:**

* I am planning on plotting, bar, scatter, swarm, box, KDE, stacked horizontal and vertical plots, histograms, line charts, heatmaps, triangular correlation matrix visualizations, ROC curves. Various other visualizations that I intend to plot include distplots, sub plots, by building various axes and figure lengths. For better understanding of what all is covered in the project, I am attaching relevant images from google to better explain the content of visualizations like distributions of “Satisfaction, Evaluation, and Monthly Hours”, “Turnover Frequency on department”, “Turnover Frequency on salary Bracket”, “satisfaction vs evaluation”, “clusters of employee turnover on basis of satisfaction vs evaluation”, “ROC curves for random forest classifier”, “ROC curves for decision trees classifier” ,” Employee Satisfaction Distribution - Turnover V.S. No Turnover” , “count of projects”to indicate how many project the turnover employee vs non turnover employees worked in, “employee department distribution”, “employee average monthly hours distribution - Turnover V.S. No Turnover”.
* As mine is a ML related project I am expecting to see the **nature of data, progressions**, if I need to up sample the data or down sample it and at what place I need to do it. If the **data is imbalanced**, I will have to balance it by seeing the **visualizations**. For example, the **graph** showing frequency of employee turnover is greater for certain department they work. If **further I want to predict** something in this project, I might need to understand what has happened and what will I need to further do to get the desired output. **This can only happen with** **visualizations** as they give a clear picture of what the data is telling which is why **visualizations** are so impactful. A 15000 x 10 rows of csv file can be collectively put into one plot or several subplots or histograms., hence is the actual power of **visualizations**.

Chart, bar chart

Description automatically generated

**Datasets and methods:**

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

* I have found data set and have attached it in the references, you may have a look. The data seems not to be too messy while I was having a look on it initially. It is a csv file. The **visualization** methods and libraries I am planning to use are – yellowbrick, seaborn, matplotlib, count plots, and many which I have elaborated above.
* One look of my data

Graphical user interface, text, application

Description automatically generated

* My data set has 10 parameters(columns) and 15000 rows which is believe is enough for a single student performing for their projects.

Graphical user interface, text

Description automatically generated

* There are no missing values in my data set and hence a majority of processing time is not wasted and mostly focused on building **visualizations**

Text

Description automatically generated

* Since the datatypes of the features are not having any inconsistencies, I find to be good enough to work with

Table

Description automatically generated

* The range of data is not bad as well. They have enough varieties in within every feature making it easier to get appealing plots

Table

Description automatically generated

* For some basic curves initially about data like univariate, bivariate analysis, I am planning to use bar, scatter, swarm, box plots so as to get a relevant idea of the underlying data in the **visualization**.
* **KDE** – *Employee Satisfaction Distribution - Turnover V.S. No Turnover*, as KDE can produce a plot that is less cluttered and more interpretable, especially when drawing multiple distributions as employees who possessed low on satisfaction levels **(0.2 or less)** left the company in larger numbers, employees who possessed low on satisfaction levels **(0.3~0.5)** left the company in larger numbers, employees who possessed really high on satisfaction levels **(0.7 or more)** left the company in larger numbers

Chart

Description automatically generated

* *Employee average monthly hours distribution - Turnover V.S. No Turnover* uses **KDE** as it is a bi-modal distribution for employees that turnover, employees who had less hours of work **(~150hours or less)** left the company more, employees who had too many hours of work **(~250 or more)** left the company, employees who left generally were **underworked** or **overworked**.

Chart

Description automatically generated

* **Stacked horizontal and vertical plots** – *Distribution of Employee Turnover, employee project count, employee department distribution***,** as we can clearly see the different data segregation for each variety /feature

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

* **Distplot** – *Distributions of Satisfaction, evaluation, and monthly hours* used basically for univariant set of observations and visualizes it through a histogram i.e. only one observation and hence we choose one particular column of the dataset.
* **Countplot** – *Time spent in the company, count of projects* to indicate how many projects or the time the turnover employee’s vs non turnover employees worked in

Chart, bar chart

Description automatically generated

* **ROC curves** - *ROC curves for random forest classifier, ROC curves for decision trees classifier* as a divergence indicator that signals a possible upcoming trend change.

Chart, line chart, scatter chart

Description automatically generated

* **Scatter plot** - *Clusters of Employee Turnover, clusters of employee turnover on basis of satisfaction vs evaluation*  so as to obtain a clear scattered group of data points forming a cluster

Chart, scatter chart

Description automatically generated

* **Bar plots** - *Distribution of Employee Turnover, employee project count, employee department distribution*, as we can clearly see the required value and map it accordingly due to different data segregation for each feature

Chart, bar chart

Description automatically generated

* **Sub Plots** - Distribution of Satisfaction, evaluation, and monthly hours, which individually are distplot. As subplots give a combined view of a set of similar features it gets easier to compare

Chart, histogram

Description automatically generated

* **Cross Tab** - *Turnover Frequency on department, Turnover Frequency on salary Bracket* as it takes two or more lists, pandas series or dataframe columns and returns a frequency of each combination by default.

Chart, bar chart

Description automatically generated

* **Lmplot** - *Satisfaction vs evaluation* to plot distinct clusters for employees who left the company. First group is of those who were hard-working and sad which is satisfaction **below 0.2** and evaluations **greater than 0.75**. Second group is of those who were less-working and sad which is satisfaction **between 0.35 and 0.45** and evaluations **below 0.58**. Third group is of those who were hard-working and happy which is satisfaction **between 0.7 and 1.0** and evaluations above **0.8.**

Chart, scatter chart

Description automatically generated

* **Heatmaps** – Applied on the entire data correlation initially, **prediction and actual stayed and left employees for various classifiers and algorithms methods**. We get to know how correlated the data and how we do the prediction on such data based on **confusion matrix**, had it been highly uncorrelated data we might need to use other methods for prediction.

Chart

Description automatically generated

Chart

Description automatically generated

* **Triangular correlation matrix visualizations** – Applied on the entire data set to summarize a large amount of data where the goal is to see patterns.

Chart

Description automatically generated

* **Histograms** –Applied on the entire data to analyze it initially to summarize discrete or continuous data

Chart, calendar

Description automatically generated

* Since by now you have got o know my entire data set and the **visualization project plan,** I feel, the data I chose is not too large nor too small, I possibly can draw as many plots as possible for many of the above distributions that I mentioned, I believe the data is enough sufficient for me to get a good **visualization** pattern every time I think of plotting one data story either using full data set or when I choose a few parameters in the set.

**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>