**Gazing into the Data Abyss: Employee Attrition**

One thing is for sure, there is a LOT of data out there. Ever since we have built on the idea of assigning values to items, we have always been thinking about what-if scenarios on how we can make the most out of what we have.

It is no surprise that over the several years of exploratory advancements with numbers, we have at the cusp of unlocking the next big thing when it comes to Data Analysis. What we can do with the data we have and how we can optimize it to maximize our objectives, it seems almost magical.

With so many applications out there, I find it amazing to dive deeper into the realm of analytics and see what amazing insights can be taken. With that, I have chosen the concept of employee attrition and what steps we can take when looking at the signs to maximize employee retention.

Before I dive into the technical part, I would like to thank the following people form who I have gained valuable insights while implementing this project:

Randy Lao <https://www.kaggle.com/randylaosat> : Predicting Employee Kernelover <https://www.kaggle.com/randylaosat/predicting-employee-kernelover>

Tuatini GODARD <https://www.kaggle.com/ekami66> : Detailed exploratory data analysis with python <https://www.kaggle.com/ekami66/detailed-exploratory-data-analysis-with-python>

Vincent Lugat <https://www.kaggle.com/vincentlugat> : IBM Attrition Analysis and Prediction <https://www.kaggle.com/vincentlugat/ibm-attrition-analysis-and-prediction>

Janio Martinez Bachmann <https://www.kaggle.com/janiobachmann> : Attrition in an Organization || Why Workers Quit? <https://www.kaggle.com/janiobachmann/attrition-in-an-organization-why-workers-quit>

Aniket Katyal <https://www.kaggle.com/aniketkatyal> : IBM Employee Attrition Classifier <https://www.kaggle.com/aniketkatyal/ibm-employee-attrition-classifier>

Anisotropic <https://www.kaggle.com/arthurtok> : Employee attrition via Ensemble tree-based methods <https://www.kaggle.com/arthurtok/employee-attrition-via-ensemble-tree-based-methods>

With credit given where credit is due, this Exploratory Data Analysis (EDA) encompasses with the amalgamation of concepts related to cleaning, defining and analyzing data to obtain meaningful insights. So let’s dive in shall we?

**Attrition: What makes them stay? What makes them go?**

Employee retention is one of the biggest metrics that a company should have in mind when thinking of growth. It becomes all the more necessary when you consider the costs of finding, interviewing and hiring the right candidate for the various jobs within the organization. This is all the more detrimental when finding out why people leave the company at an unprecedented rate. Employee attrition is caused when the total strength of the company is greatly reduced as more employees leave the company than expected.

This data set comes from a fictitious scenario come up by the data scientists at IBM. Although it is not attuned to the real-word scenario, it still is a viable dataset to mine and build insights on. The main crux of the data is to classify what type of people would leave a company based on 35 parameters. Our categorical target for this dataset is a Yes/No classification from the attrition.

With that in mind, the various steps that we have gone into building insights from start to finish are highlighted in FOUR major blocks:

1. Collecting and Cleaning the Data
2. Exploring the data
3. Transforming the data to build insights
4. Applying machine learning models for feature engineering

With these at the forefront, it highlights the general steps that can be taken to work towards any dataset. For the given attrition dataset, I wanted to focus more on the graphical and visual representation aspect while keeping the machine-learning concepts in the back-burner.

**STEP 1: Cleaning the data**

Once you obtain the data, it is necessary to clean the dataset so that the valuable insights have a defined meaning. There are a number of ways for cleaning data, ranging from null-values to outlier detection; each with its own unique implementation.

When looking at the bare-bones of the given data, we can find a number of people with distinct categories ranging from age to number of years with their current manager. These parameters are what shape the given dataset and with a bit of leaning, we can delve deeper into the main causes of attrition.

Once cleaned, our dataset would look something like this:

**STEP 2: Exploring the data**

Since the given dataset is split into a numeric values such as Age and labeled values such as Department, it seems prudent to partition the data on quantitative and categorical values. This makes it easier to fine-tune the model for accurate results.

One way of visually communicating how data is spread is through histograms. This can give general insights in terms of data distribution. Some of the key takeaways from this dataset are shown below:

Another way of estimating relations between two values arise from the Kernel Density Estimate (KDE) plots. This makes it easier to highlight the bivariate relationship between data distributions that occur within the framework. Some of the key insights are as follows:

**STEP 3: Transforming the data to build insights**

Data in its innate nature needs to be molded so that we can mine it for insights. The main drive behind this is to fine-tune the data so that we can build actionable metrics from the data as a whole. With this, both the categorical and quantitative data ca be shaped.

When working with categorical values, there may or may not exist a quantifiable priority within the labels. Therefore, one-hot encoding is used to give a quantifiable value to the categorical labels.

A correlation matrix is the fastest way to see which metrics are related to each other. Although correlation does not mean causation, it is a great way to estimate the underlying behaviors within the dataset.

**STEP 4: Feature Engineering**

Now that the data has been prepped for analysis, it is time to apply machine learning methodologies to the dataset.

The following technique were used to implement the dataset to estimate the best model that would be used

1. Logistic Regression – Sine there are 1/0 value datasets, it makes it ideal to estimate.
2. Decision Tree – An easy to implement classification method.
3. Random Forrest – Decision trees with overfitting prevention
4. SVM – Classify data based on its location on the hyperplane
5. KNN - Classify data based on its location with similar clusters

With these implemented, we get an estimate on what would be the best approach to take.

As we can see, SVM and Random forests take the lead. We can break down the metrics to see what the most driving factors are:

**Outcomes and key takeaways**

Based on the above results, how much people earn and how old they are greatly affect their reason for changing to a different company. In a world where we constantly strive to know our worth, it goes to say that we will find greener pastures if we do not find it at a single point. Ultimately, it falls on the company to maximize their employee retention with a focus on their pay. Although it might decrease net company revenue in the short run, the cost of rehiring and re-training each and every time an employee leaves will definitely leave a mark.

All in all, there are so many ways to approach the same data problem from different angles. With practice, the end goal will be to form an effective story with the data to provide actionable insights. Thank you very much for coming along with me on this journey. This was a great learning experience for me and I hope that it would be for you as well.