**Approach for solving Human Resource Analytics Problem**

* **Problem Statement and Dataset**

<https://datahack.analyticsvidhya.com/contest/wns-analytics-hackathon-2018-1/>

* **Dealing with missing values**

Options available:

1. Imputers: KNN Imputer, Simple Imputer
2. Replacing with mean, mode, median of respective columns
3. Replacing with 0, abruptly high number, “Not Present” (for categorical feature) etc.

KNN Imputer is out of option as dataset contains 54,808 rows. KNN is computationally very expensive and has really high RAM consumption because of which execution is terminated.

Replacing missing values with “Mode” of respective column as those features were categorical. We can use Simple Imputer also for doing the same job.

We also have an option to drop rows which has missing values but that’s the last option we must consider as dropping columns or rows always cause loss if data which is always a red flag in Machine Learning.

* **Logical Feature Selection**

1. “employee\_id” can be dropped as individual employee’s id should not decide whether he/she should be promoted or not.

* Checking for unique values of each features in order to check if the contain any out of scope values.
* **Calculating correlation of variable**

Options available

1. Carmer’s V co-efficient
2. Phi co-efficient
3. Contingency co-efficient

We can’t sue contingency co-efficient because value range from 0 to infinity so it is difficult take judgement which one is highly correlated and which one is not.

Cramer’s V co-efficient makes use of Phi co-efficient so it’s a better option for checking correlation also the value of Cramer’s V co-efficient range from -1 to 1 similar to that of Pearson’s correlation co-efficient

* **Statistical feature importance**

Since output is also categorical Chi-square test can be used to check the feature importance of categorical variables.

* **Choice of Machine Learning Algorithm**

Basic Distance Based Algorithms

1. K-Nearest Neighbors Classifier
2. Logistic Regression

Basic Non-distance Based Algorithms

1. Decision Tree Classifier
2. Naïve Bayes Classifier

Ensemble Methods

1. Random Forest Classifier
2. AdaBoost Classifier
3. Gradient Boosting Classifier
4. XGBoost Classifier

K-Nearest Neighbors can’t be a good choice looking at the size of the dataset we are dealing with. Also K-Nearest Neighbors is supposed to be the lazy learner.

Logistic Regression can be a good choice as the problem is Two class classification problem (“is\_promoted” Yes or No). But we can also discard this option as we have already seen there are hardly any features which are highly correlated to the output variable, and none of the feature is showing a linear trend.

Still Logistic Regression is considered just to see how it performs.

* **Encoding**

Options available

1. Label Encoding
2. One-Hot Encoding

In case of encoding we have to One-Hot encode all input categorical features in case of Logistic Regression Algorithm as every class/category of every categorical feature has to get equal importance also Logistic Regression is a distance based model.

In case algorithms like Decision Tree Classifier, Random Forest Classifier we can use Label Encoding as all the algorithms are non-distance based.

* **Scaling**

Options available

1. MinMax Scaler
2. Standard Scaler

We can use any of the above two methods for scaling of dataset.

**Why do we need scaling of dataset?**

For instance many elements used in the objective function of a learning algorithm (such as the RBF kernel of Support Vector Machines or the L1 and L2 regularizers of linear models) assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

* For Logistic Regression using Backward Elimination for extracting relevant features, the method uses p-test to extract the most relevant features
* **Methods for tacking Data imbalance**

1. Under sampling Techniques

* Random Under sampling
* Cluster Centroids
* Tomek Under Sampling

1. Over Sampling Techniques

* Random Oversampling
* SMOTE

Under sampling is always bad choice to make as it removes majority chunk of your data in order to make number of instances of both the class same, which can cause huge loss of data which risky. Cluster centroids and Tomek Under Sampling can still be a descent choice among undersampling techniques, but both these methods have huge RAM consumption because of which execution is interrupted.

In case of Over sampling we can use any Over sampling technique among the available choices but SMOTE is the most preferred one as it makes use of K Nearest neighbour concept to fill makes same of number of instances for both classes.

**Business use cases:**

Generally huge amount of data about employees working in the company is present with HR department of the company who manages this data.

1. By applying predictive analysis to this data, HR is able to become a strategic partner that relies on proven and data-driven predictive models, instead of relying on gut feeling and soft science
2. HR predictive analytics enable HR to forecast the impact of people policies on well-being, happiness, and bottom-line performance. An example is the role it can play in preventing expensive employee turnover.
3. **Leveraging the Data:**  which is already available, where in the company don’t have to rely on feedback from individuals which cannot be trusted on times. The company can totally make decisions based on the factual data available with them.
4. **Predictive Analytics and Company Culture:** The predictive model can also tell you the features or factors impacting the success or failure of individual department or the overall company.
5. **Predicting Employee Turnover Offers Warnings, Solutions:** When your company predicts high-risk employee turnover demographics, you increase your enterprise’s ability to address the problem. In other words, you may be able to lessen the turnover percentage by focusing on the specific types of workers likely to resign.
6. **Predictive Analytics Can Improve Culture, Productivity, and Achievements:** Companies rightfully focus on employee retention numbers, but data analytics might also help improve a company’s culture and productivity. For example, if a company concentrates on salary’s influence on employee happiness, it might miss other employee-happiness variables, causing it to miss out on changes which might actually create a better work environment.

When employee predictive analytics becomes company culture, it cuts down on reactionary decisions. It increases the odds of company success by exposing the least costly obstacles to that success.

**Challenges Faced:**

1. Tackling data imbalance with proper choice of method.
2. Finding the perfect balance between Bias and Variance. That is regularizing models in in order to avoid over fitting.
3. Keeping in mind the that there should be a “random\_state” value assigned where “random\_state” is one of the hyper parameter of the model as the results change every time we try and run the model.
4. Using correct metric for checking the performance of different models. As dataset contains huge data imbalance “accuracy\_score” cannot be used as evaluation metric as it will always give false interpretation.
5. Performing hyper parameter tuning of the models carefully in order to find the best estimator for each model.
6. Running Ensemble techniques hyper parameter is highly time consuming as ensemble methods are computational very expensive.
7. The hyper parameter tuning has to be done in step by step manner where in we have to keep the range small otherwise there is possibility of overshooting values which are actually making the basic model a best estimator and also keeping smaller ranges will help code run faster.