Springboard Data Science Career Track

Capstone Project 2

HR Data Analysis Report

By Loveleen Bangar

Oct 2023

The HR department of TechSolutions Inc. wants to understand key factors influencing employee job satisfaction. HR wants to predict employee job satisfaction based on historical performance data, employee characteristics, and other relevant variables.

The purpose of this project is to come up with a model for employee satisfaction based on multiple factors. HR analytics is a data driven approach to manage human resources. It involves gathering and analyzing data related to employees, such as recruitment, performance, engagement, and retention, to derive insights and make informed decisions. Main objective of this analysis is to understand factors influencing employee job satisfaction and key predictors of job satisfaction. This project aims to build a predictive model for job satisfaction based on factors such as salary, commute, benefits etc. This model will be used to provide guidance for the HR department to focus on factors which increase job satisfaction.

About Data:

Analyzed data contained 1470 rows and 35 columns. JobSatisfaction was chosen as the target column and all other columns were features to evaluate job satisfaction. Each row represents one employee per row because all the employee ids were unique in the employeeid column.

Data Wrangling:

Cleaned data to prepare for exploratory testing. Checked data type of columns and missing values and duplicate rows. There were no missing values found in any column and there was no duplicate row. Analyzed unique values in all the columns and count of records for each value in the column. Data was collected from three departments: Sales, Research & Development, and Human Resources. There are 9 job roles Sales Executive, Research Scientist, Laboratory Technician, Manufacturing Director, Healthcare Representative, Manager, Sales Representative, Research Director, Human Resources from 12 education fields.

There were four levels of job satisfaction. 569 employees selected 1 and 2 levels, around 901 employees selected 3 and 4 levels.

EDA:

EDA gave some deeper insight about employee job satisfaction.

Guiding questions for EDA were how Job Satisfaction related to other features like Business Travel, Department, EducationField, Gender, Marital Status etc. Does OverTime increase Job Satisfaction? Is there a relation between BusinessTravel and Job Satisfaction?

In data 9 columns were categorical columns. To perform analysis dummy features were created and saved in a data frame with dummy features for future use. For the analysis purpose job satisfaction values 1, 2 considered as employee not satisfied and 3, 4 considered as satisfied.

To understand feature correlation heatmap was created.

According to heatmap, JobLevel and MonthlyIncome were highly correlated. Other positively correlated columns were Joblevel, TotalWorkingYears, MonthlyIncome and Age. YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager. Target column JobSatisfaction was not correlated to any column in the dataset.

Further analysis was performed based on department and age. In this dataset, the majority of employees were in the Research and Development department (65.4%), and Sales had 30.3% and Human Resources had 4.3%.

In Research & Development 62% employees chose job satisfaction level 3 or 4, which means they were satisfied with their job and 38% chose 1 or 2 level, which means they were not satisfied with their job.

In Sales 61% employees chose job satisfaction level 3 or 4, and 39% chose 1 or 2 level. Human Resources was a small department with the least number of employees (63). In Human Resources 51% employees chose job satisfaction level 3 or 4, and 49% chose 1 or 2 levels.

Pre processing:

I have approached this problem in two ways, Linear regression using the job satisfaction column as percentage and Logistic regression by encoding 0 and 1 values for the job satisfaction column. First processed data then built a model using scikit-learn. In pre-processing, split data into train and test parts by 70 / 30 percent, then built classification and regression models logistic, KNN, Decision tree, random forest, linear regression and random regressor. Linear regression and random regressor performance was not good, It has MAPE around 50%, means models results are not promising. We can't use this model for future predictions.

I created 4 different classification models: Logistic regression, KNN, Decision Tree, and Random Forest. Even though classification models showed better accuracy scores, performance of models was still not good. Recall score of all the models was very low. Means model results show more false positives. Employees who were not satisfied were showing satisfied per model results. Which model to use depends on the client's decision, whether the client is ok with false negatives or false positives. Further model improvements were done as part of modeling.

Modeling Results:

I used a dataset from pre-processing with job satisfaction column values 0 and 1 and explored more with classification models using 80/20 train test split. In the dataset 0 values are 39% and 1 values are 61%, it was an imbalanced dataset. To create unbiased model results, I resampled training data using different techniques - Over sampling, SMOTE over sampling, ADASYN over sampling, KNN under sampling and random under sampling. Then created logistic regression and XGBoost classification models using resampled data. Main goal was to increase recall, which was very low without resampling. Out of all the models, Logistic regression with random under sampling and XGBoost with over sampling showed better results. It showed a good balance of false positives and false negatives. These model results will be discussed with the client, accordingly deciding which model to use for future predictions.

Future Steps:

These model results can be improved by choosing optimal threshold for the ROC Curve and Precision-Recall Curve directly or manually choosing threshold value for recall and precision. To make this model more useful, create a UI interface, so that this model can be run from UI without looking into code and share it with the functional team.