**Integrated CA 2**

**Louis Wilkie**

**Introduction**

This project will analyse a dataset: Employee\_attrition.csv.

This project will follow the CRISP-DM framework.

**Business Understanding**

The main task of this project is to analyse the data to improve employee satisfaction and productivity. Throughout the evaluation of the data the influence of data preparation, statistical techniques, and machine learning modelling are to be understood.

The main task of this project is to predict two business objectives which are present in the data as features. This will involve two separate supervised learning modelling efforts.

The company that owns the data wishes to understand the influence of data preparation, statistical techniques, and machine learning techniques on the analytical outcome, possibly to refine their future data science objectives. So, importantly here in the business understanding phase, some baselines are established. These baselines are established before data preparation, statistical analysis and machine learning models are refined.

The target variables are float numbers, it does imply that they are evaluated from other metrics and are more detail has been used in generating them than a 1 - 5 scoring system. It makes this a viable regression task, as the objective variables are floats and possibly have relations with other features.

One line of enquiry could be to bin the target features and perform classification modelling, but this won’t be performed for the baseline model.

This project will utilise sklearn machine learning models, which require data in a certain format, thus minimum preparation will be performed to align with this requirement in creation of a baseline model. This includes minimum strategies for missing data and encoding data.

As two target variables have been identified, JobSatisfaction and PerformanceRating, it might be useful to examine further the relationship of these variables in an effort to understand if increasing JobSatisfaction can increase PerformanceRating, and how linked these variables are.

**Data Understanding**

Business objectives are represented by features in the data.

Basic Data Exploration.

Target Variables.

Before much data exploration is done, a baseline models are created. These models use minimum data preparation and models the two target variables, using Linear Regression. The targets are modelled including the second target and without, giving four total models. The modelling report is below. The baseline model isn’t a good model of the data, and gives a foundation for examining data preparation, statistical approaches and machine learning strategies.

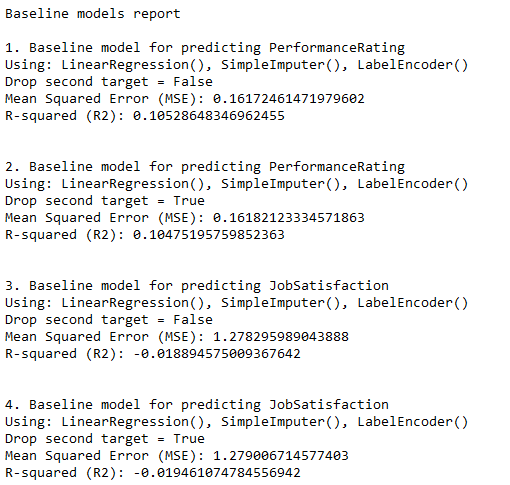


Figure 1: Baseline model report

Most features are uncorrelated. The highest correlation is *MonthlyIncome* and *JobLevel* at 0.92. Some features that describe years working under certain circumstances have strong correlation. Age has some correlation with a few features in the dataset.

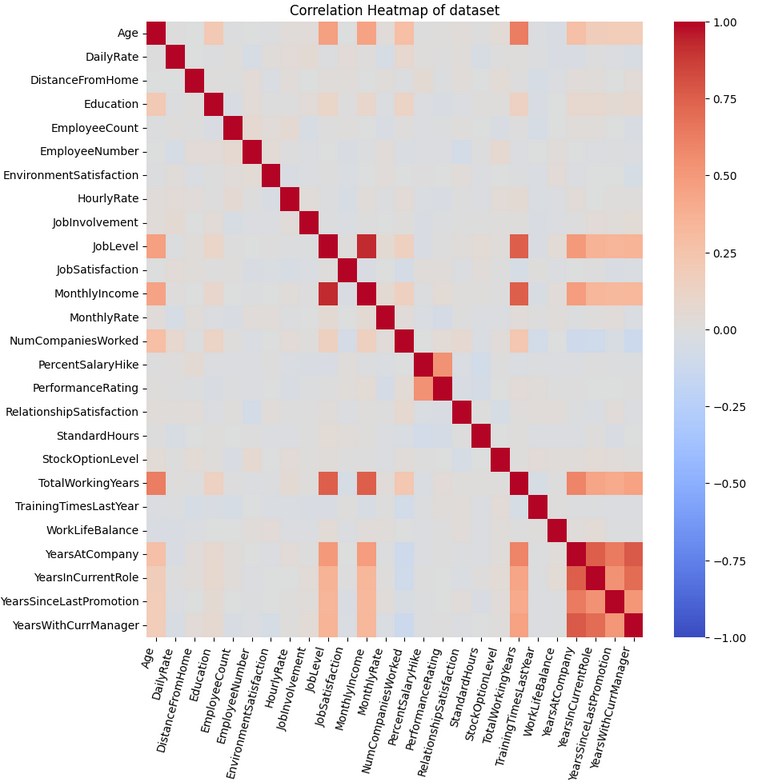


Figure : Correlation Heatmap of numeric features.

One task of the data is to

**Data Preparation**

Missing data strategy

**Conclusion**