Use Case: Human Resources Analysis Predict Attrition

Bruno Candeias (M20180313), David Oliveira (M20181430), Henrique Pereira (M20181395), Manuel Oom (M20181431)

# Status Report

We choose this use case mainly because it allows us to explore different models, which will give us the opportunity to have a broader view for the problem: descriptive and predictive. In addition, human resources turnover is very present in our professional life, which also motivated our choice.

## Planning and Methodology

Regarding project planning, we have built a Gantt Chart, as show in Figure 1, in order to evaluate the progress of the project and frame it with deliverables.

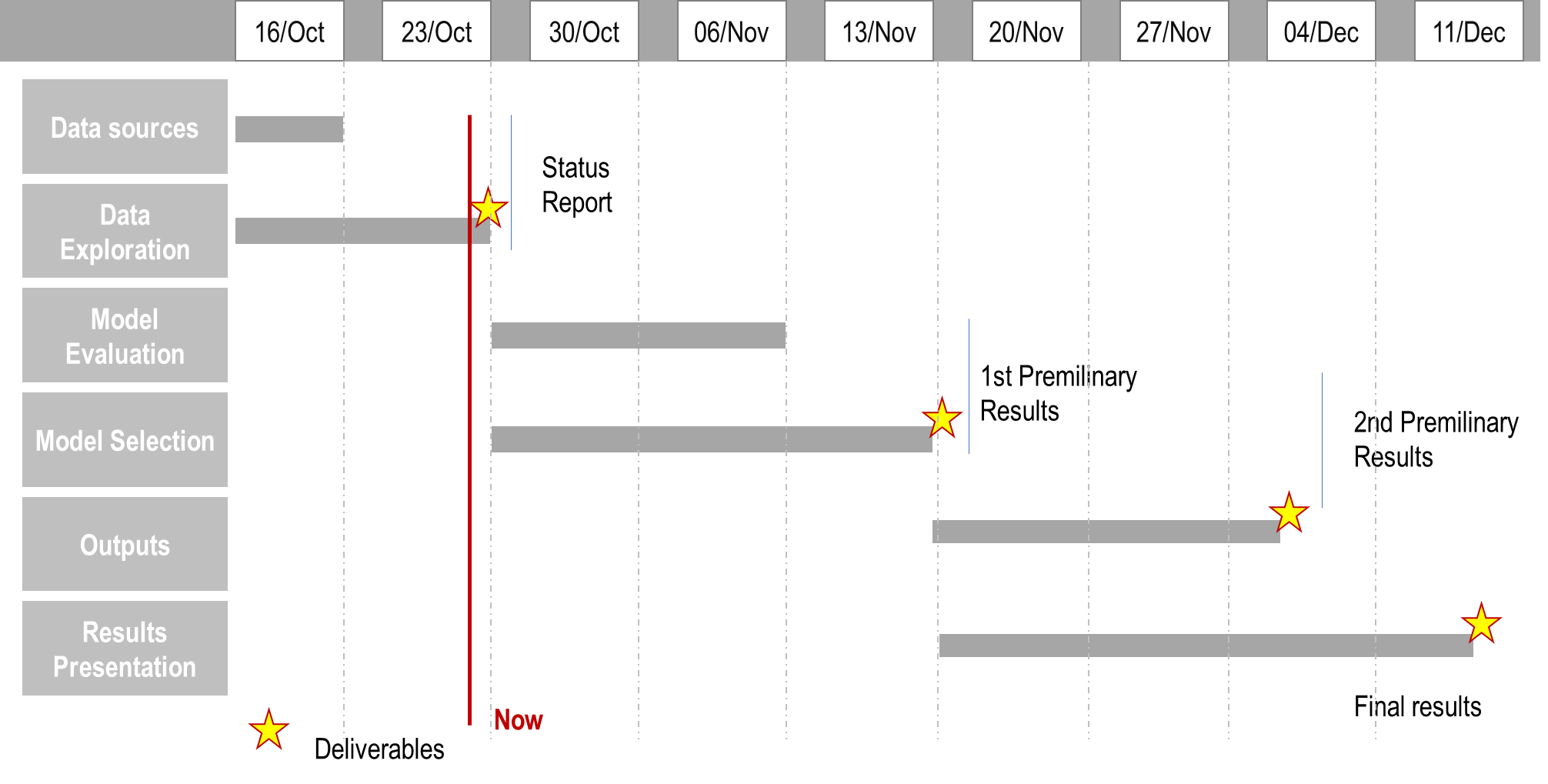


Figure 1 – Project Planning

Our approach will follow the blocks exposed in Figure 2: Data Exploration; Model Evaluation & Selection; Results Presentation. We are currently in progress in the Data Exploration block, and after a thorough knowledge of the available data, which is an important process for our project, we can move on to applying models to data.

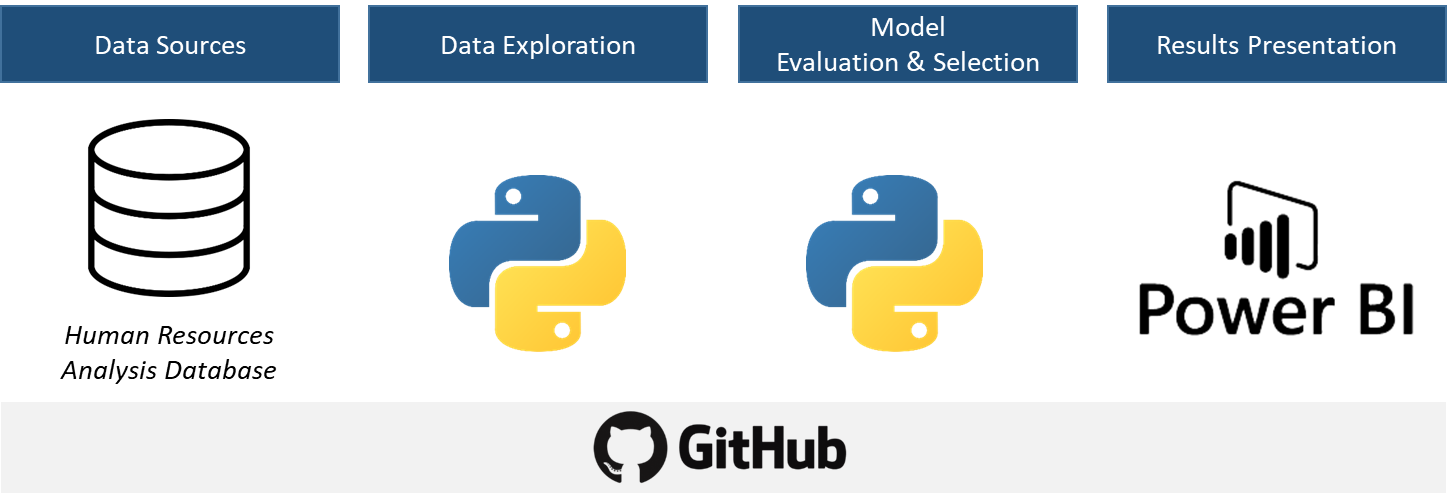


Figure 2 - Project Methodology

## Dataset Exploration

The dataset (*HR\_DS.csv*) that will be used in the use case (Human Resources Analysis Predict Attrition) contains 1470 records with 35 columns:

* *Age*
* *Attrition*
* *BusinessTravel*
* *DailyRate*
* *Department*
* *DistanceFromHome*
* *Education*
* *EducationField*
* *EmployeeCount*
* *EmployeeNumber*
* *EnvironmentSatisfaction*
* *Gender*
* *HourlyRate*
* *JobInvolvement*
* *JobLevel*
* *JobRole*
* *JobSatisfaction*
* *MaritalStatus*
* *MonthlyIncome*
* *MonthlyRate*
* *NumCompaniesWorked*
* *Over18*
* *OverTime*
* *PercentSalaryHike*
* *PerformanceRating*
* *RelationshipSatisfaction*
* *StandardHours*
* *StockOptionLevel*
* *TotalWorkingYears*
* *TrainingTimesLastYear*
* *WorkLifeBalance*
* *YearsAtCompany*
* *YearsInCurrentRole*
* *YearsSinceLastPromotion*
* *YearsWithCurrManager*

We are starting to explore the dataset to evaluate each variable and how they are correlated. Our first findings are:

* Most of our data (*DistanceFromHome, MonthlyIncome, NumCompaniesWorked, PercentSalaryHike, TotalWorkingYear, YearsAtCompany, YearsS*inceLastPromotion) shows skewness, and not normal;
* Columns like *YearsWithCurrManager* and *YearsInCurrentRole* have 2 diferent distributions with a cutoff by 5 years;
* There are several variables that have outliers: *MonthlyIncome*, *NumCompaniesWorked*, *PerformanceRating, StockOptions*, *TotalWorkingYears, TrainingTimesLastYear, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager*;
* Regarding the variable *Attrition* and how it is correlated with other variables, we analyzed the data and we realized that younger employees leave more in all categories, except SalesExecutive, ManufacturingDirector, Manager and Divorced. We also realized that the gender is important in some conditions.

In this initial part of the project, we are facing some difficulty regarding the meaning of each variable and how their value has to be interpreted. For example, it is not clear for us how we should interpret *PerformanceRating* variable, because we do not know if it is in ascending or descending order. We also noticed that the dataset is unbalanced, which can be a problem when we apply the models. However, there are some ways to solve this issue and we are evaluating which one are we going to apply.

Regarding the block Model Evaluation & Selection, we are studying several predictive and classification models to apply in the dataset related with the use case, namely:

* Logistic Regression Classifier;
* Linear Support Vector Classification;
* C-Support Vector Classification;
* Radius Neighbors Classififier;
* Random Forest Classifier;
* Classification on imbalanced data with Tensorflow Neural Network.

## Next Steps

1. Explore more deeply the dataset;
2. Data preparation and transformation;
3. Selection of the variables that are relevant for the problem and the variables that show strong correlation;
4. Apply several models to our dataset;
5. Analyze preliminary results.