Business Case Study and Report

QuantSpark Second Round Analyst Assessment

This presentation has live translations.

QuantSpark

Presented by:
Joaquim Bolós Fernández

Date Submitted:
March 16th, 2023

QuantSpark

Overview

- Problem Statement
- Project Outline
- EDA - I
- EDA - II
- EDA - III
- EDA - IV
- EDA - V
- Recommendations - EDA
- Predictive Model - I
- Predictive Model - II
- Findings & Recommendations - Models
- Challenges
- Further Steps

Problem Statement

Back to Agenda

01

Retention of high performing employees

Strategic Objective

02.1

Capacity to predict exactly which employees are at most risk of leaving

02.2

Capacity to determine changes to the current operating model

Tactical Objectives

03.1

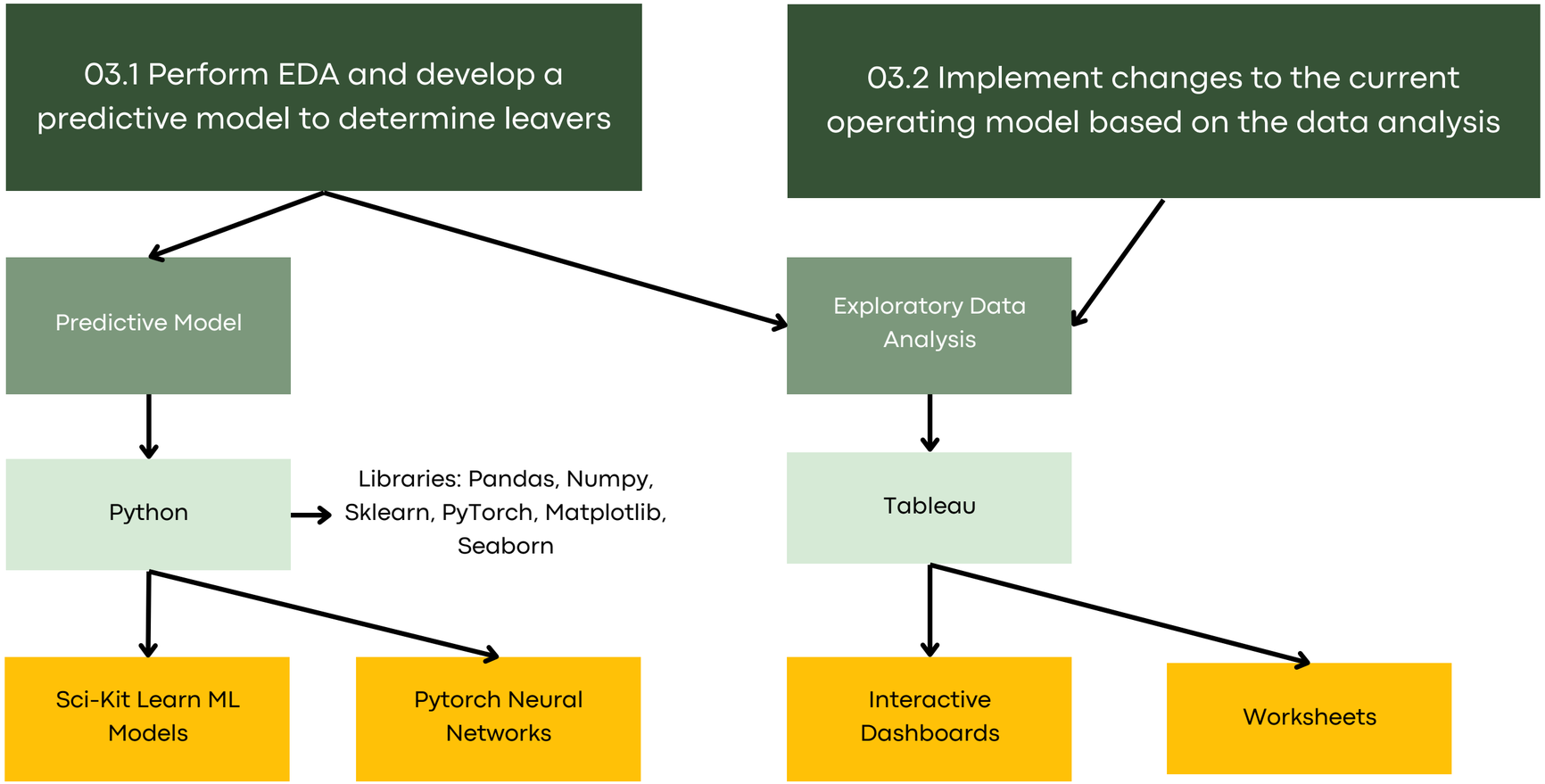
Perform EDA and develop a predictive model to determine leavers

03.2

Implement changes to the current operating model based on the data analysis

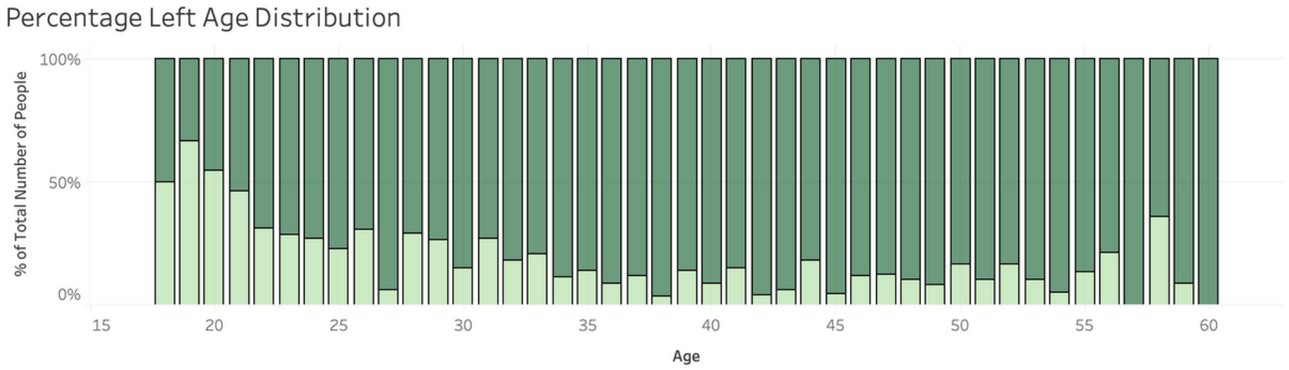
Operational Objectives

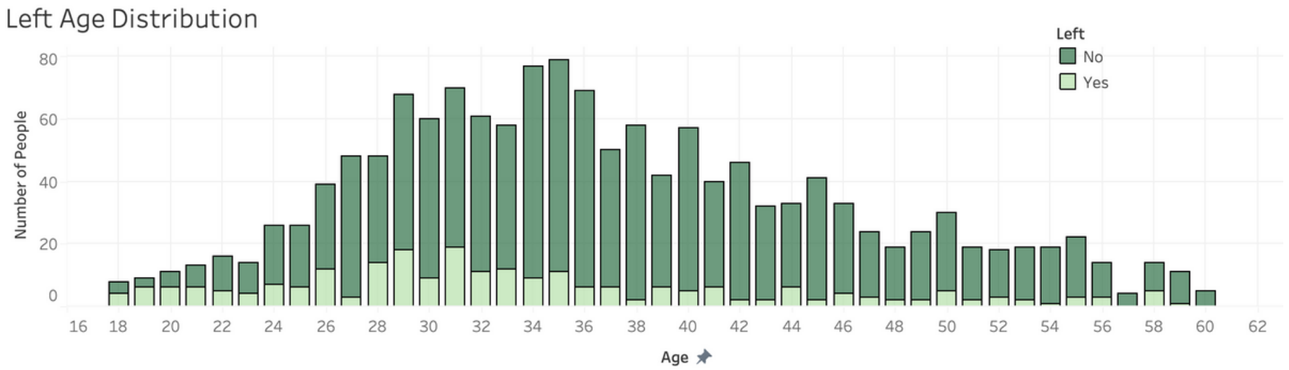
Project Outline / Operational Objectives

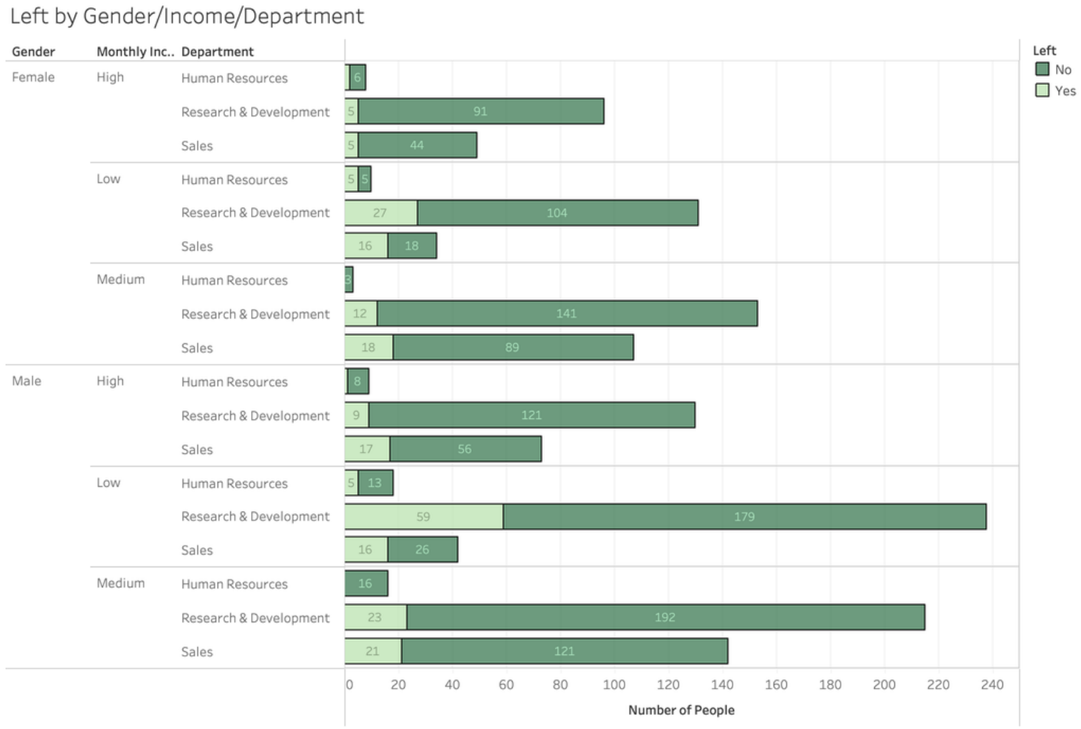


Back to Agenda

Exploratory Data Analysis - I







- Most of the employees are between 28 and 40 years old.
- The churn percentage of 18-36 years old is greater than +36 years old
- Highest leaving churn at 31 years old

- Average Age of Employees - 36.95
- Median of Age of Employees - 36
- Median of Age of Employees - 35

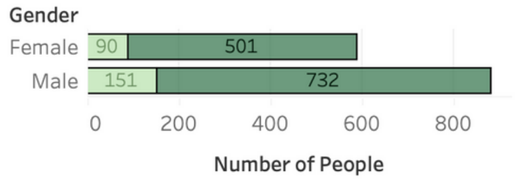
- Average Age of Leaving Employees - 33.79
- Median of Age of Leaving Employees - 32
- Median of Age of Leaving - 31

Source: Tableau

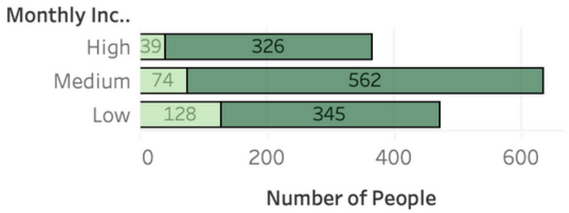
Back to Agenda

Exploratory Data Analysis - II

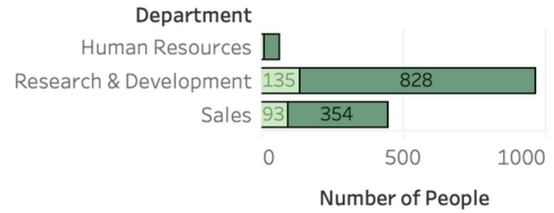
Left by Gender



Left by Income

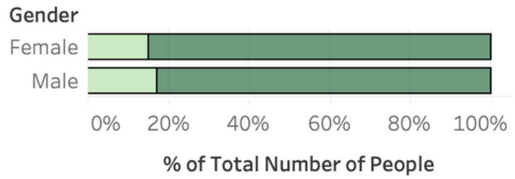


Left by Department

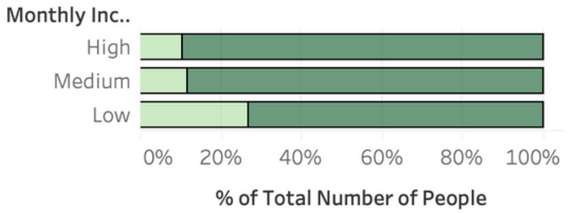


Left by Business Travel

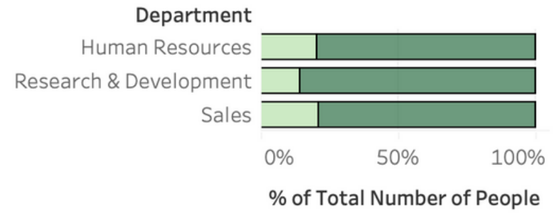
Percentage Left by Gender



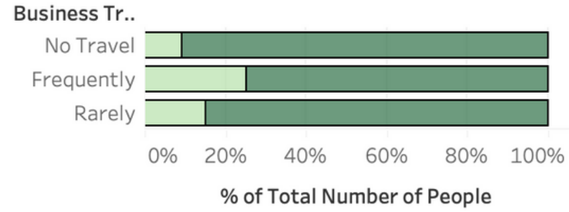
Percentage Left by Income



Percentage Left by Department



Percentage Left by Business Travel

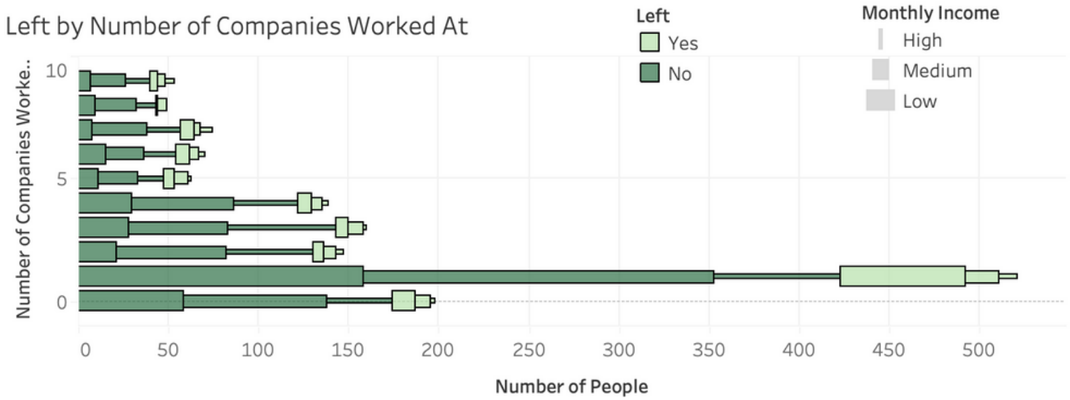


- Gender has little correlation on leaving churn (Churn difference 2%)
- Employees on a low salary double churn percentage (27%) of employees on medium and high wages
- All the departments have similar leaving percentage churn, but HR and Sales are the highest ones with a 20% churn percentage
- Travelling has a direct influence on the leaving percentage. The more travelling, the higher leaving percentage

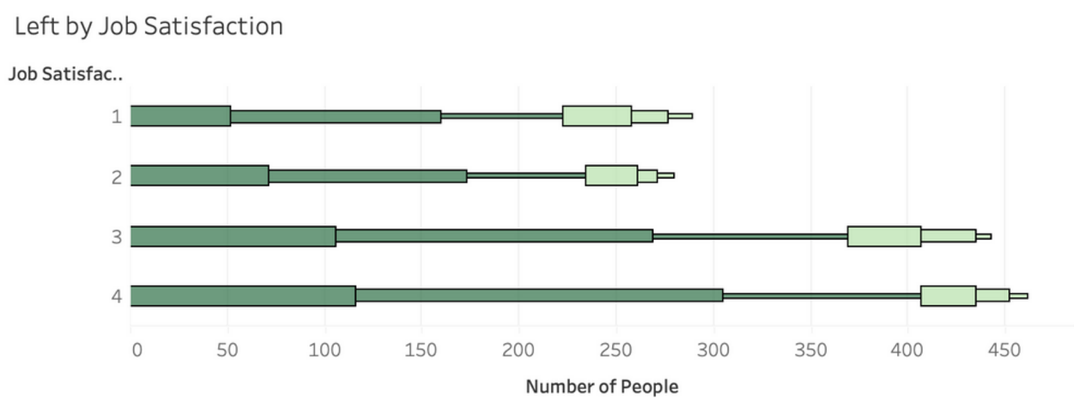
Source: Tableau

Back to Agenda

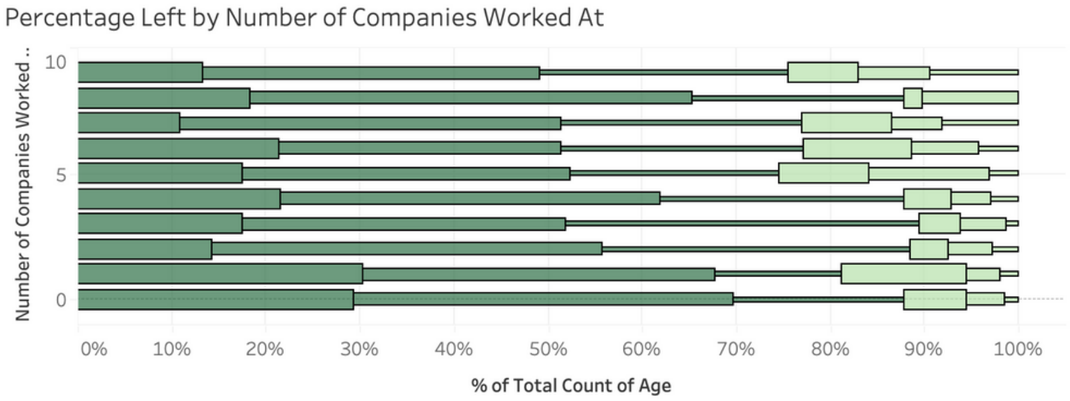
Exploratory Data Analysis - III



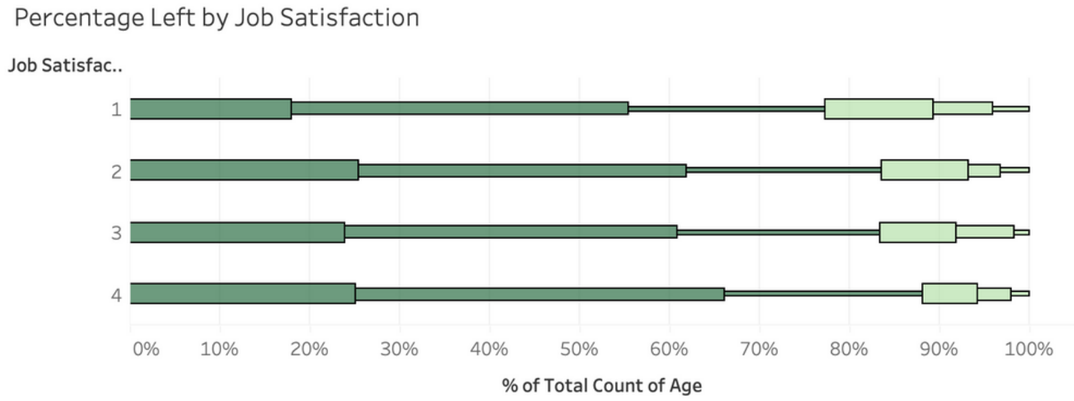
- The highest leaving churn is for people on a low salary that have worked at 2 companies



- The highest leaving churn is for people on a low salary with a job satisfaction level 3



- The percentage of leaving is similar for all employees despite the number of previous companies they worked at. However, the churn is 8.53% higher for employees that have worked at five or more companies, especially if they have low wages.

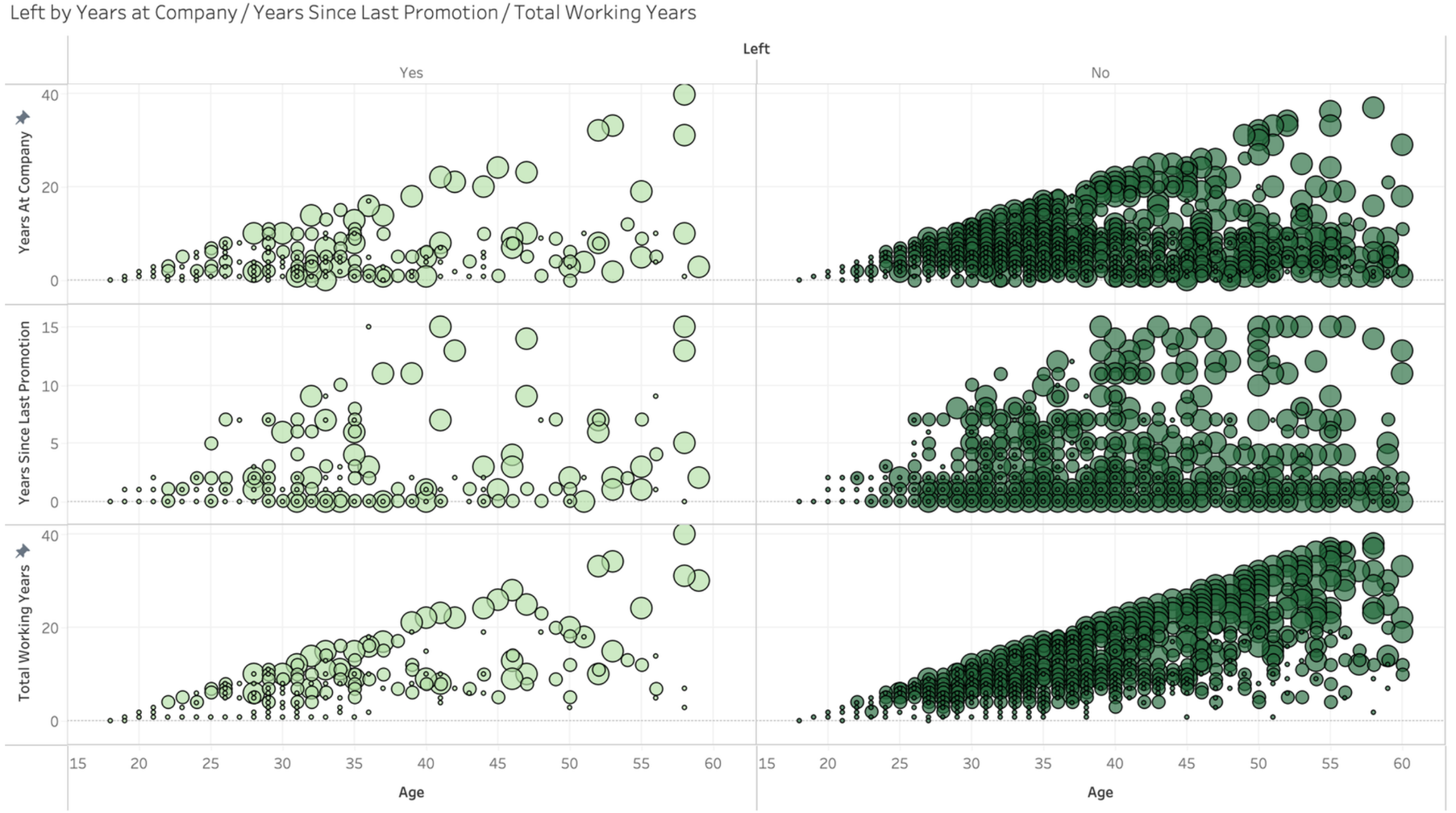


- There is a direct link between churn and job satisfaction. The churn percentage at job satisfaction rating 1 is 12% higher than at job satisfaction rating 4.

Source: Tableau

Back to Agenda

Exploratory Data Analysis - IV

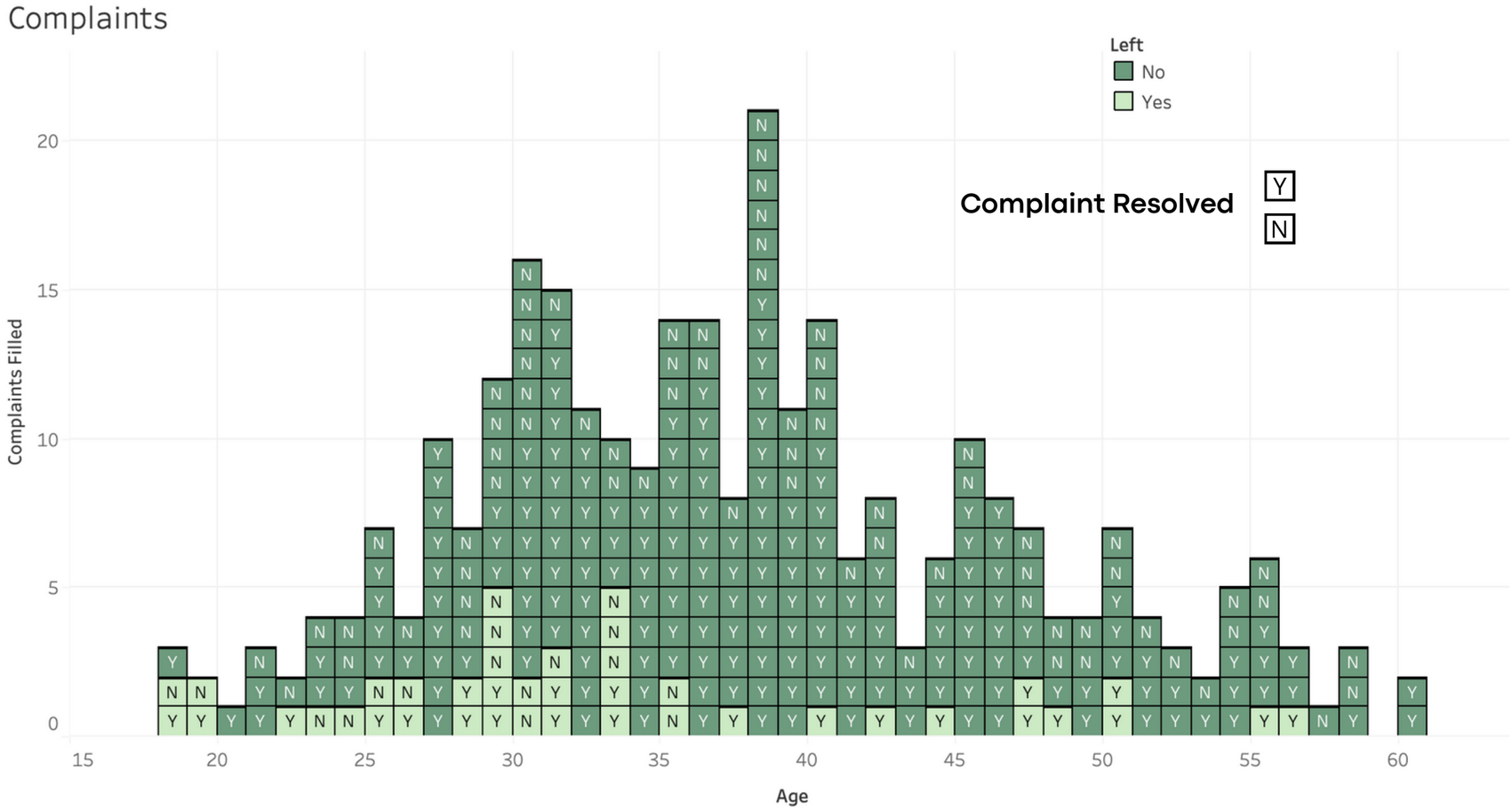


- Employees with higher salaries and +36 years old usually stay.
- There is correspondence between years since the last promotion, years at the company, total working years, income and leaving churn.
- The total working years against the age chart shows how employees leave the company if they are on a low salary after 5-10 years.
- Similarly, in the Years Since Last Promotion graph, it can be seen how low-income employees leave after 1-3 years, and medium salaries either leave at the beginning or after approximately five years (probably because they get better offers).

Source: Tableau

Back to Agenda

Exploratory Data Analysis - V



- Complaints filed are more frequent from the ages of 27 - 40, but don't seem to have a strong connection with the leaving churn.
- Out of the people who complained and left, more of those complaints were resolved.
- 41% of people who left did not get their complaints resolved.
- 28% of people who stayed did not get their complaints resolved.
- A 100% of people with more than 36 years old who left got their complaints resolved.

Source: Tableau

Back to Agenda

Recommendations

EDA I	EDA II	EDA III	EDA IV	EDA V
Implement a strategy to keep talent (ages 18 - 30) through mentorship schemes. Increase promotions from low to mid or high salaries. Reduce travel requirements. Implement an anonymous feedback system for leavers to detect flaws in each department.		Try to recruit new personnel that have not been in more than 5 companies. Implement an anonymous feedback system on job satisfaction to ensure possible action to increase job satisfaction levels	Implementing a monitoring strategy to determine when an employee is likely to receive or start looking for better offers. The predictions should be based on Working Years/Years at the Company/ Years since the Last Promotion. Then, offer promotions to key employees if considered necessary. Retain high-performing employees with new offers by meeting their requirements	Analyse all previous complaints to ensure no previous complaints are repeated. Reinvent the complaint form to gather more data about the complaint (category of the concerning matter, gravity, etc.). This information can be taken into account to improve the working situation of the employees and decrease churn.

Back to Agenda

Predictive Model - I

Data Preparation

- Format & Filter Data
- Split Data into training, validation and testing datasets
- Standardise Data
- Label Encoding / One Hot Encoding

Sci-Kit Learn ML Models

- K-Neighbors Classifier
- Decision Tree Regressor
- SVC
- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier
- **Gradient Boosting Classifier**

Github Repo: Quimbolos/QuantSpark_Assessment

Gradient Boosting Classifier

Confusion Matrix

228 TP	11 FP
46 FN	10 TN

Metrics

F1 score - 0.57
Precision - 0.65
Recall - 0.57
Accuracy - 0.81

GridSearchCV Scoring Method

F1 Score - harmonic mean(average) of the precision and recall, which answer how many of whom we labelled leavers are leavers? And of all the leavers, how many of those we correctly predict?

Best Hyperparameters

'ccp_alpha': 0.0,
'criterion': 'squared_error',
'init': None,
'learning_rate': 0.1,
'loss': 'log_loss',
'max_depth': 3,
'max_features': 'log2',
'max_leaf_nodes': None,
'min_impurity_decrease': 0.0,
'min_samples_leaf': 1,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 100,
'n_iter_no_change': None,
'random_state': None,
'subsample': 1.0,
'tol': 0.0001,
'validation_fraction': 0.1,
'verbose': 0,
'warm_start': True,

Back to Agenda

Predictive Model - II

Data Preparation

- Format & Filter Data
- Split Data into training, validation and testing datasets
- One Hot Encoding

Pytorch NN Models

Fixed Values

- batch size = 30
- Loss Function = CrossEntropyLoss

Optimised Values

- Optimisers - ['SGD', 'Adam', 'Adagrad']
- Learning Rate - [0.01, 0.001, 0.0001, 0.00001]
- hidden layer width - [32, 64, 128, 256]
- depth - [50,100,150]

Github Repo: Quimbolos/QuantSpark_Assessment

NN Classification Model

Confusion Matrix

202 TP	23 FP
0 FN	0 TN

Metrics

F1 score - 0
Precision - None
Recall - 0
Accuracy - 0.89

Best Hyperparameters

'optimiser': 'Adagrad'
'learning_rate': 1e-05
'hidden_layer_width': 256
'depth' : 150

Scoring Method

F1 Score - harmonic mean(average) of the precision and recall, which answer how many of whom we labelled leavers are leavers? And of all the leavers, how many of those we correctly predict?

Back to Agenda

Findings & Recommendations

ML MODEL - SKLEARN	ML MODEL - NN
The best Model Predicts True Positives (Employees that stay) relatively well, but when it comes to determining leavers (Negatives) tends to label non-leavers as leavers (False Negatives). Additionally, it has a lower chance of determining one leaver as a leaver (True Negative) than a leaver as a non-leaver (False Positive). Consequently, it could be said it can only be used as an orientation measure to predict leavers.	Due to the Unbalanced Dataset, the NN does not predict any leaving churn (Negatives). Consequently, only some of the metrics can be computed. Further research should be done in implementing UpSampling to obtain models that can predict leavers.

OVERALL
The DataSet is small and imbalanced (16,35% Positive Churn), so the neural network won't probably perform better than the gradient booster classifier despite the upsampling. In addition, another reason to choose the SKLearn Model is the interpretability of the algorithm.

Back to Agenda

Challenges

EDA

- Getting the most insights possible in a brief period of time.

- When working with a small dataset, there are higher chances of encountering outliers/exceptions, which can significantly influence the overall analysis. Additionally, the data sampling could be biased due to various reasons like a small sample size, selection bias, or non-random sampling. These issues can affect the accuracy and generalizability of the results obtained from the analysis.

ML MODELS

1. Format & Filter Data
2. One hot encoding
3. Upsampling (NN)
   - Scaling continuous features
4. Hard to develop an accurate model for employees <25 yds or >45 years old (fewer data for those age groups)

Back to Agenda

Further Steps

OVERALL

Explore the segmentation of groups to only focus on high-performing employees (Eg. Focus on employees with high-Performance Ratings, high Working Years, High Working Years at the Company, etc.)

Increase the amount of available data.

Reduce imbalance in the DataSet from newly collected data

Explore UpSampling application to further reduce imbalance (SKlearn models)

Explore feature importance (SKlearn models).

Explore decreasing the number of estimators (SKlearn Model - Gradient Boosting Classifier).

Improve Sampling techniques to increase data for groups ages under 25 and over 45.

Explore the possibility of reducing the NN complexity (Model complexity is too high - reduce the number of hidden layers and width of the layers).

Back to Agenda

QuantSpark

Get In Touch

Back to Agenda

Email

bolosfernandez@hotmail.es

GitHub

Quimbolos/QuantSpark_Assessment