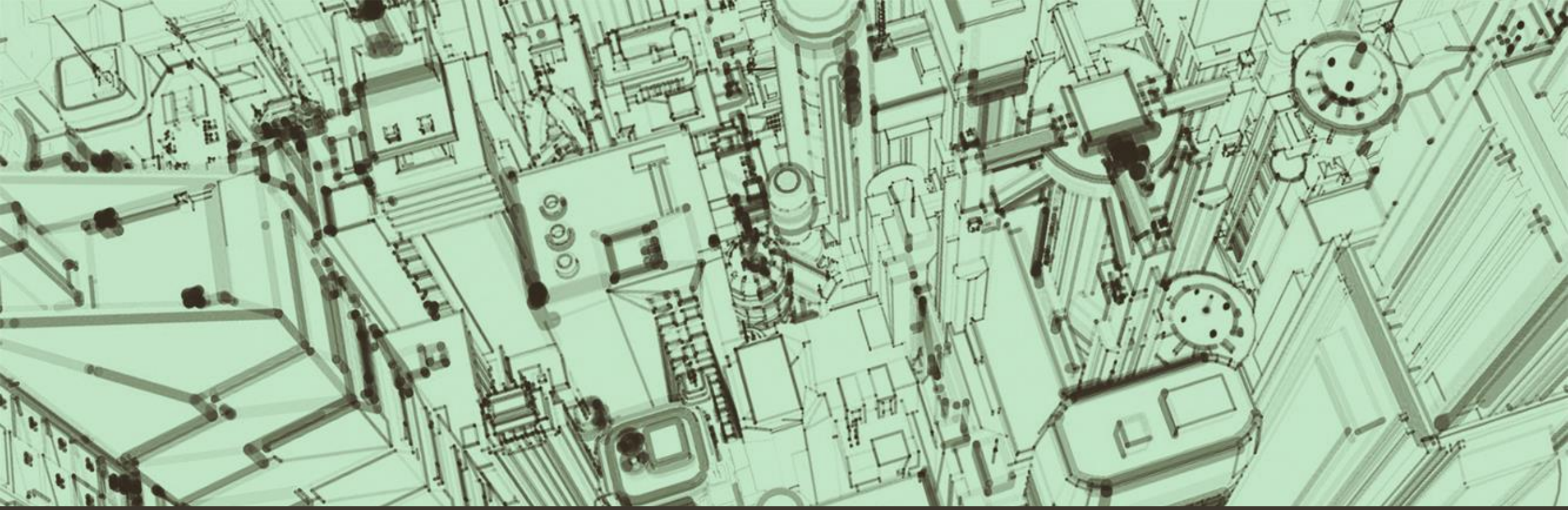


# HR Attrition Problem

For XYZ Pvt. Ltd.

By Shruti Diwakar, Swapnik Chimalamarri, Tarunay Roy and Rakesh M G

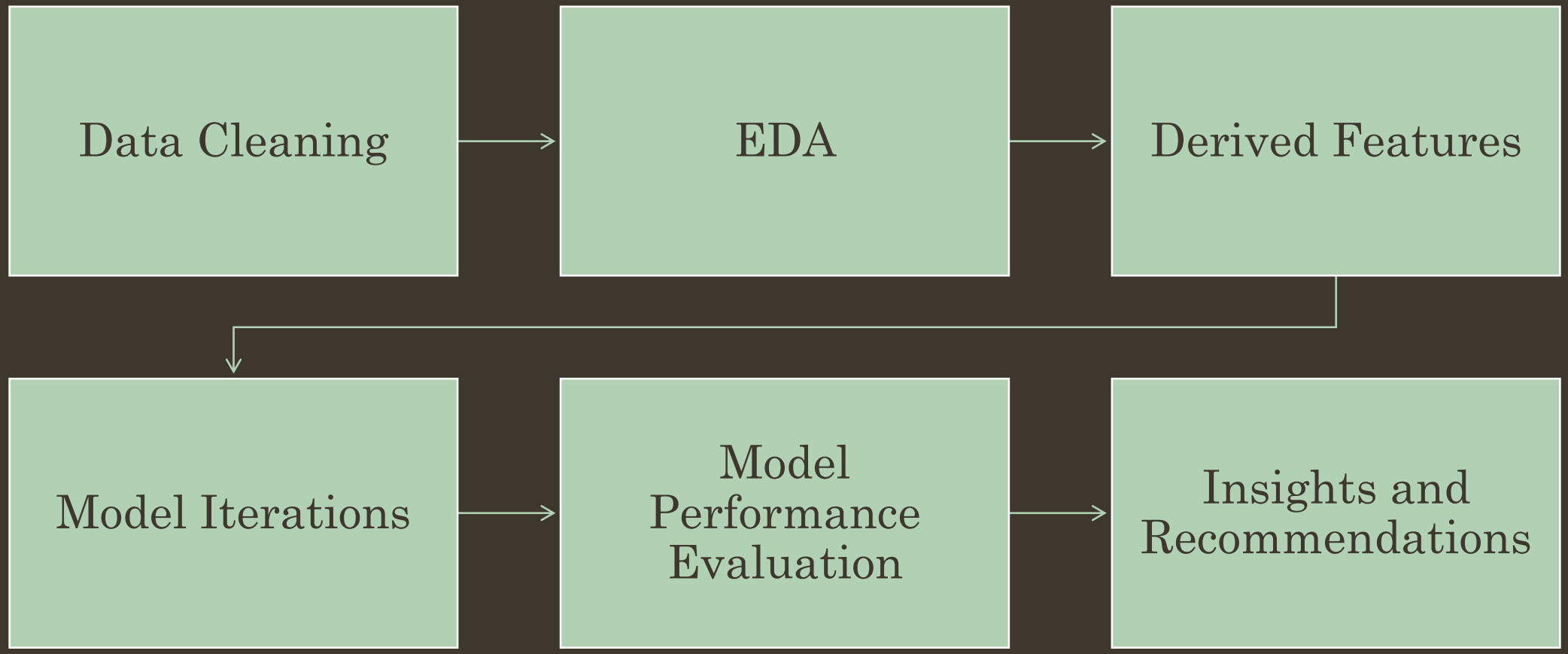
# Content Layout

- Objective of Case Study
- Problem Solving Methodology
- Data Cleaning and Preparation
- Visualizations from EDA
- Derived Features (Brainstormed and through Decision Trees)
- Model Performance Evaluation
- Key Insights and Recommendations

# Objective of Case Study

- XYZ Pvt. Ltd. is facing high attrition rates of ~15% annually
- This leads to fluctuations in delivery of current projects and time as well as money is spent in ramping up new employees
- Thus, our objective is to help identify the factors leading to attrition so that XYZ can focus on them to rein in the attrition rate
- Our model will also help in identifying employees most likely to resign with a yes/no indicator as part of the output
- Since this is a classification problem (whether an employee with resign or not) and we need to explain all the factors to stakeholders, we will go with Logistic Regression as the preferred technique

# Problem Solving Methodology

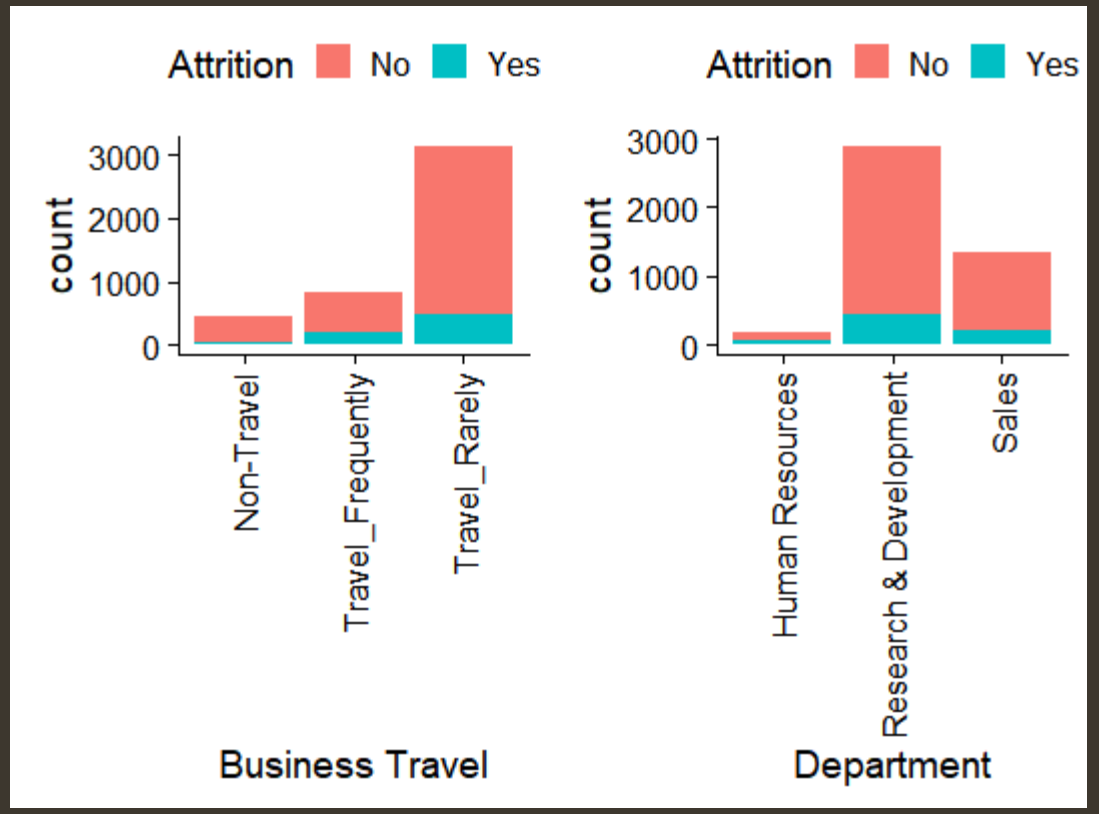


# Data Cleaning and Preparation

- Corrected column name of EmployeeID in In-Time and Out-Time files
- Replaced NAs in file with median where applicable
- Performed Outlier identification and capped them at meaningful levels
- Converted in-time and out-time values to date datatype for easier analysis
- Created Dummy variables for categorical features
- Performed scaling of continuous features

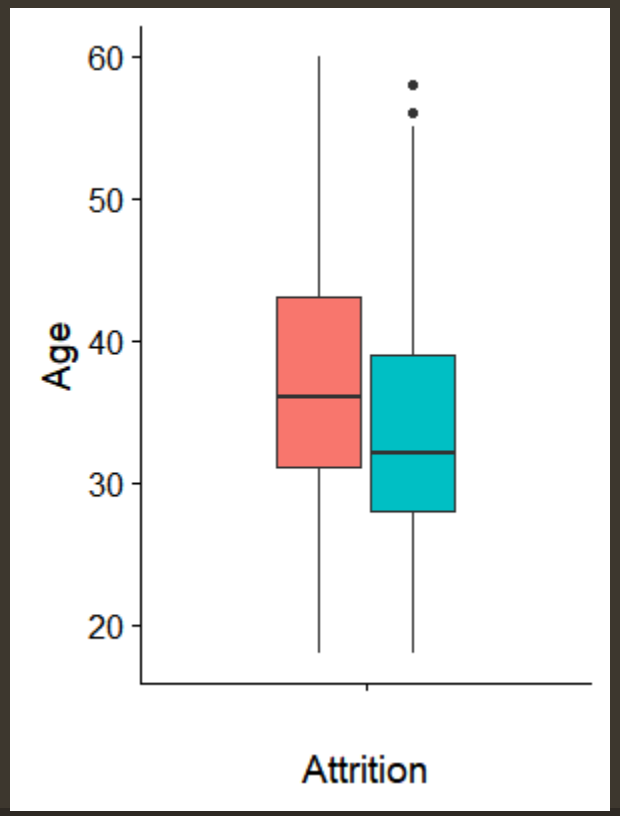
# Visualizations from EDA
## (All other EDA Charts in Appendix)

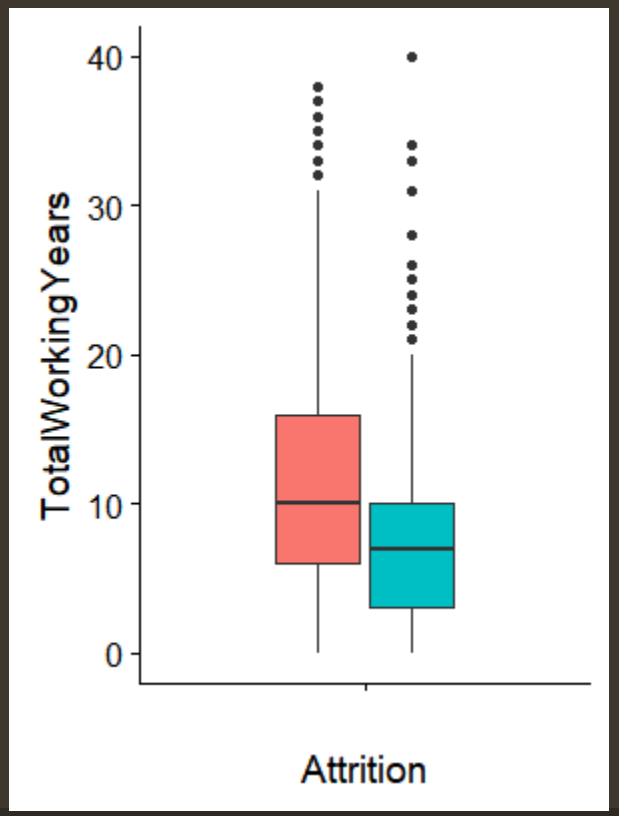
Distribution of Categorical Variables with Respect to Attrition. We observe some interesting patterns here

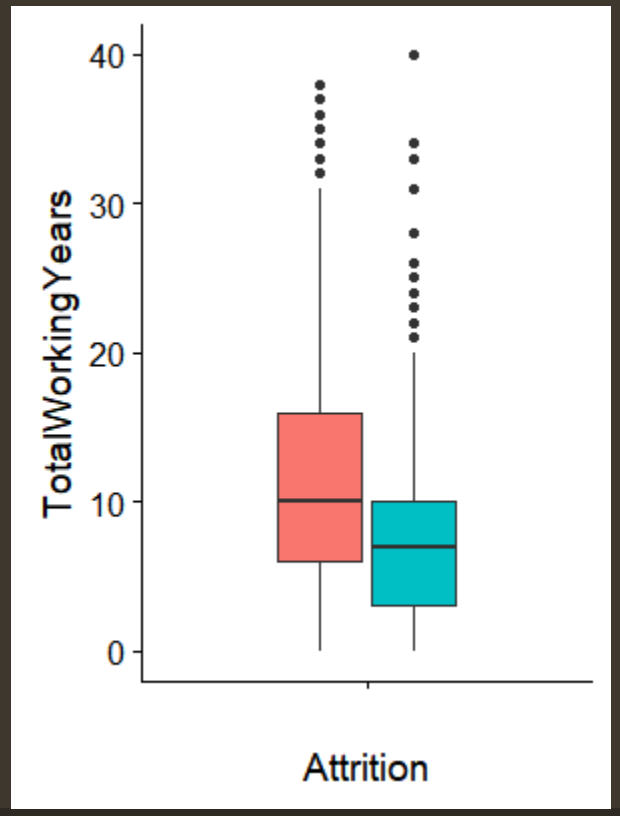



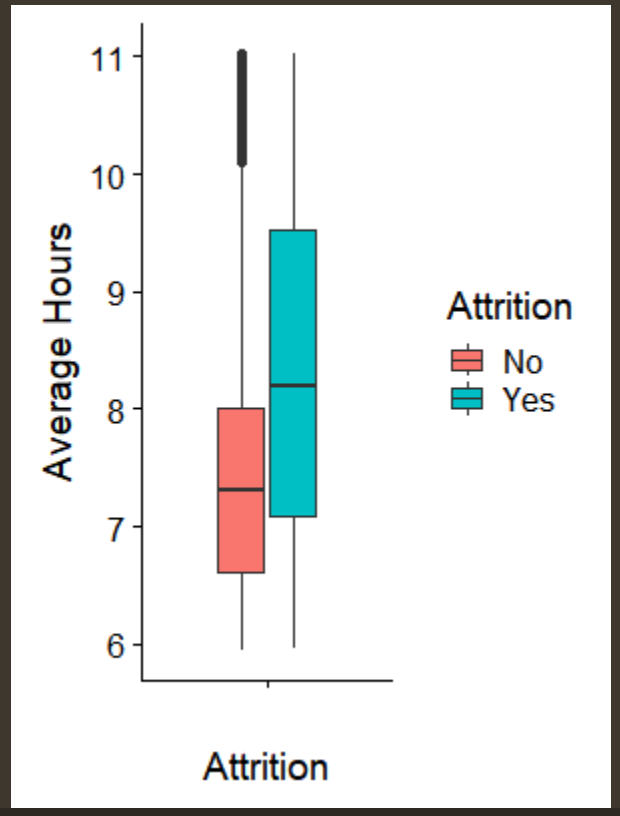
# Visualizations from EDA
## (All other EDA Charts in Appendix)

Distribution of Continuous Variables are plotted with respect to the Attrition result.
In general, younger people (in terms of age, wtotal work experience, work experience in company tend to have higher attrition. Also, please working longer hours on an average are more likely to quit.



# Derived Features through Brainstorming

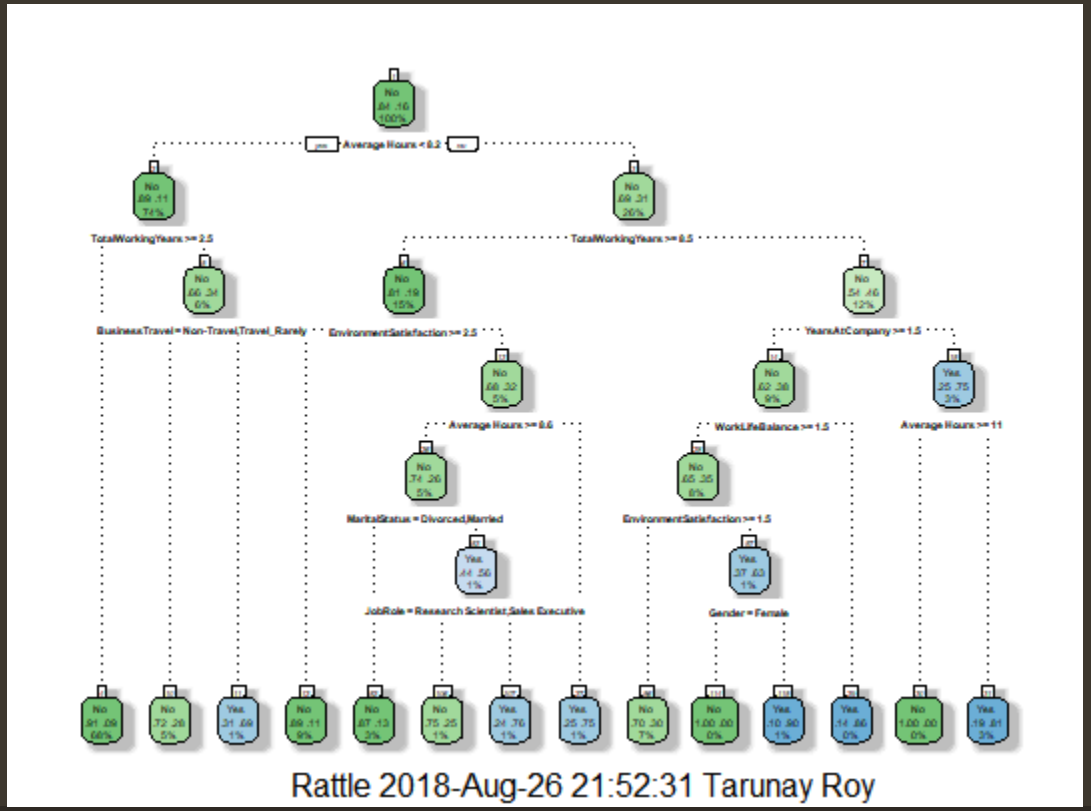
Below set of Derived Features were created through brainstorming –

- Average number of hours worked each day by every employee using in-time and out-time data
- Overtime flag to categorically mark employees working more than 8 hours on an average
- Inadequate time flag marking employees working less than 7 hours per day on an average
- Number of days taken as leave by the employees using in-time, out-time data
- Segmented total number of working years as per experience buckets

# Interactive Features through Decision Tree

Since Logistic Regression does not account for interaction between variables, we ran a Decision Tree on the data to observe variable interaction. We used this information to create derived interaction variables so that the model could identify patterns better.

- Flag for users who's average working hours are low (<8 hrs/day) and have a total work experience > 2.5 years
- That is, users working less hours and relatively inexperienced
- Flag for users having a high total work experience (>= 8.5 years) and a long time since the last promotion (> 5.5 years)
- That is, experienced employees being overlooked for promotion



# Model Performance Evaluation - I

In order to evaluate the performance of our model, we will use the below metrics. They are plotted graphically in the next slide alongside different cut-offs.

- **Sensitivity** (Rate of true positives. That is, number of correctly predicted positives out of all the actual positives)
  - Obtained value – 73%
- **Specificity** (Rate of true negatives. That is, number of correctly negatives out of all the actual negatives)
  - Obtained value – 73%
- **Accuracy** – Rate of correctly predicted outcomes. That is, number of positives and negatives correctly predicted out of the entire population.

# Model Performance Evaluation - II

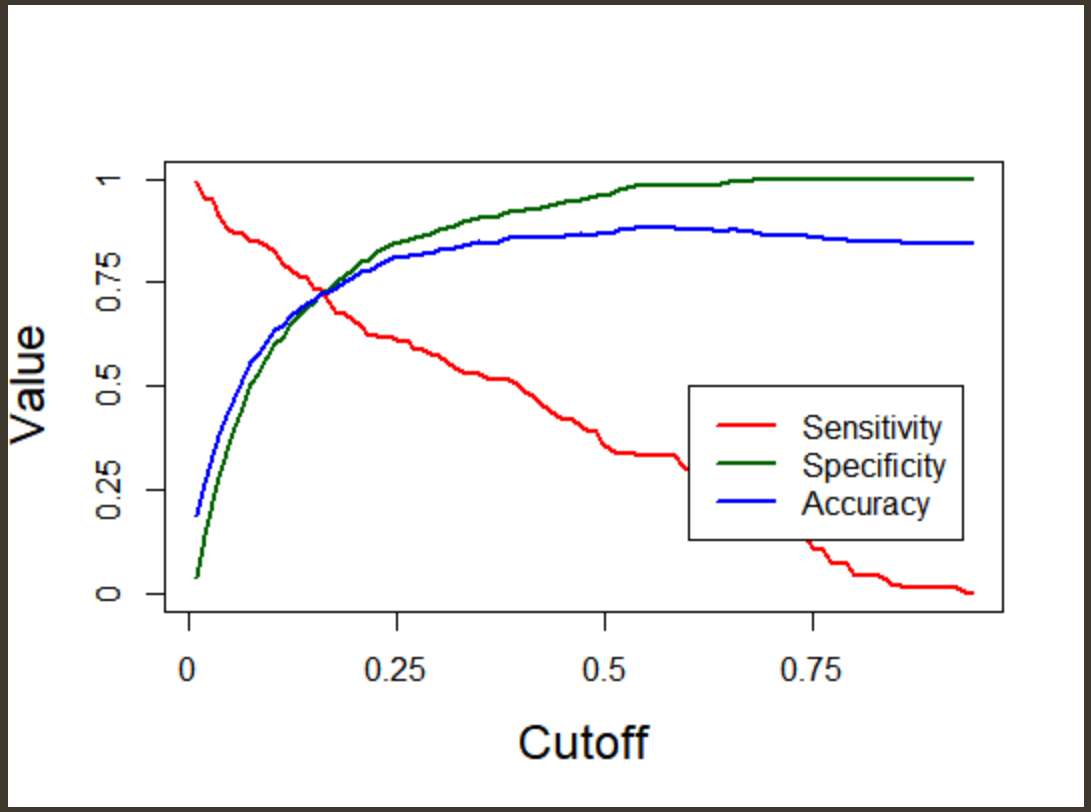
The Logistic Regression model gives us a probability as an output for each employee. The probability tells us how likely the employee is to resign.

In order to decide whether a certain probability means a yes/no for resignation, we have to decide an optimal cut-off to classify them.

Different cut-offs would lead to different predictions and consequently, different levels of performance of the model.

To balance the needs to of a good sensitivity, specificity and accuracy as per business needs, we have selected the intersection point to be the cut-off.

Cut-off selected is a probability of 0.16. That is, any probability above 0.16 is marked as a 'yes' in the output that they will resign



# Model Performance Evaluation - II

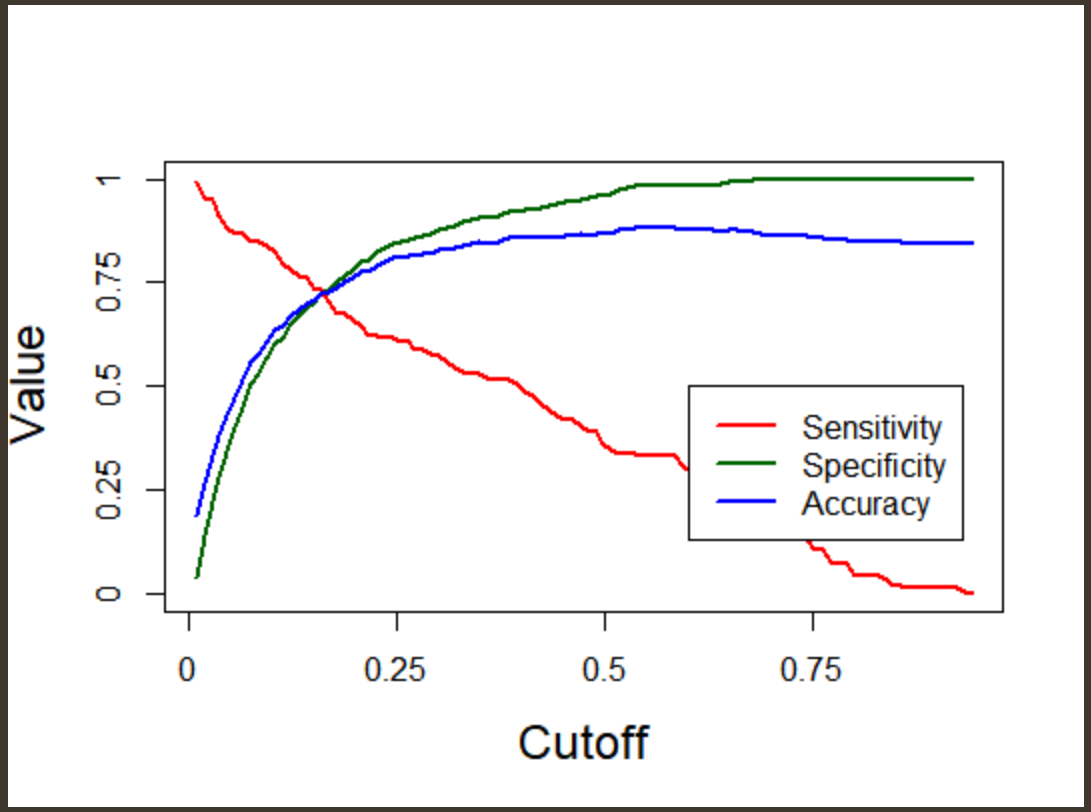
The Logistic Regression model gives us a probability as an output for each employee. The probability tells us how likely the employee is to resign.

In order to decide whether a certain probability means a yes/no for resignation, we have to decide an optimal cut-off to classify them.

Different cut-offs would lead to different predictions and consequently, different levels of performance of the model.

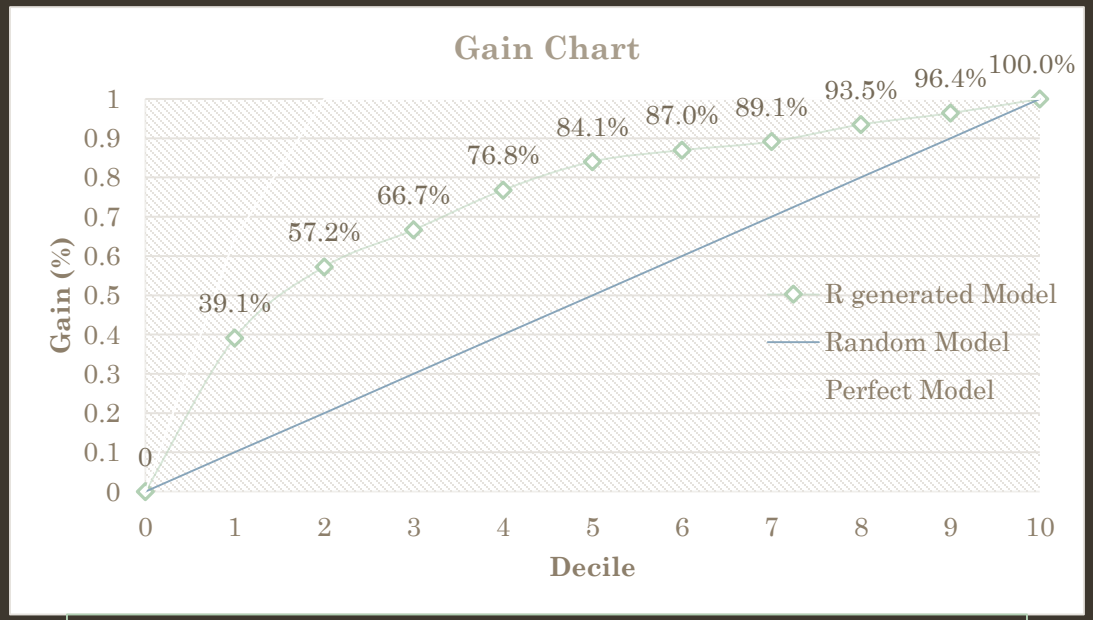
To balance the needs to of a good sensitivity, specificity and accuracy as per business needs, we have selected the intersection point to be the cut-off.

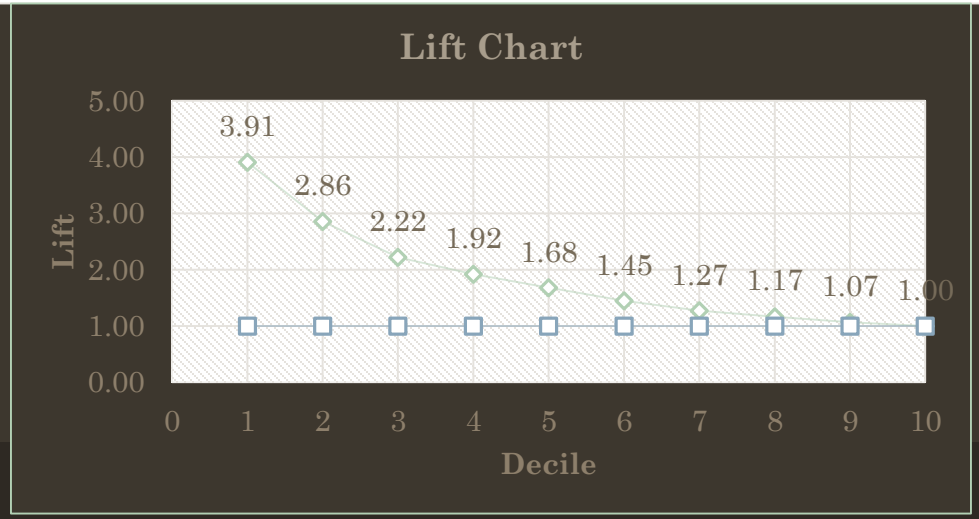
Cut-off selected is a probability of 0.16. That is, any probability above 0.16 is marked as a 'yes' in the output that they will resign



# Model Performance Evaluation - III

- We calculate the Kolmogorov–Smirnov Statistic for the model to observe how well the model is able to differentiate between yes and no. A value above 0.4 is considered to be good and we obtained ~0.45. This means the model is performing well
- Lastly, we look at the **Gain and Lift** charts of the model. Even here, the model results look good as the model is able to categorize 77% of the yes's in the top 40% of the data
- Here, it is compared with the results of a random model (straight line) and it is clear that the model is able to categorize/arrange the users well



# Key Insights and Recommendations

Using the Coefficients obtained in the model, we can assess the relative importance of the variables

Thus, the more important variables are the ones that need to be focused on by XYX Pvt. Ltd. first, to control its attrition problem. Below is a list of variables in order of their importance, along with recommendations –

- Overtime – Employees working overtime are more likely to leave. Hence, put in policies and practices to reduce this. Avoiding under-staffing and improving strength of trained employees can help with this
- Experienced employees not being promoted – This is harming the motivation of experienced employees as they seem to think they deserve a promotion, having worked for so long. Tweaking the promotion criteria and promotion process can help address this. More transparency around promotions would help as well
- Freshers – They seem to be more likely to quit. Hence, XYZ can focus more on lateral hires having > 2 years of experience
- Marital status – Single people seem more likely to quit. Hence, XYZ can make special note of these people and provide them better incentives to stay
- Frequent Business Travelers – People made to travel frequently for business seem to be more likely to quit. Hence, it would be better to distribute traveling duties among multiple employees and try to rely on videoconferencing and remote tools to get work done instead of traveling as far as possible