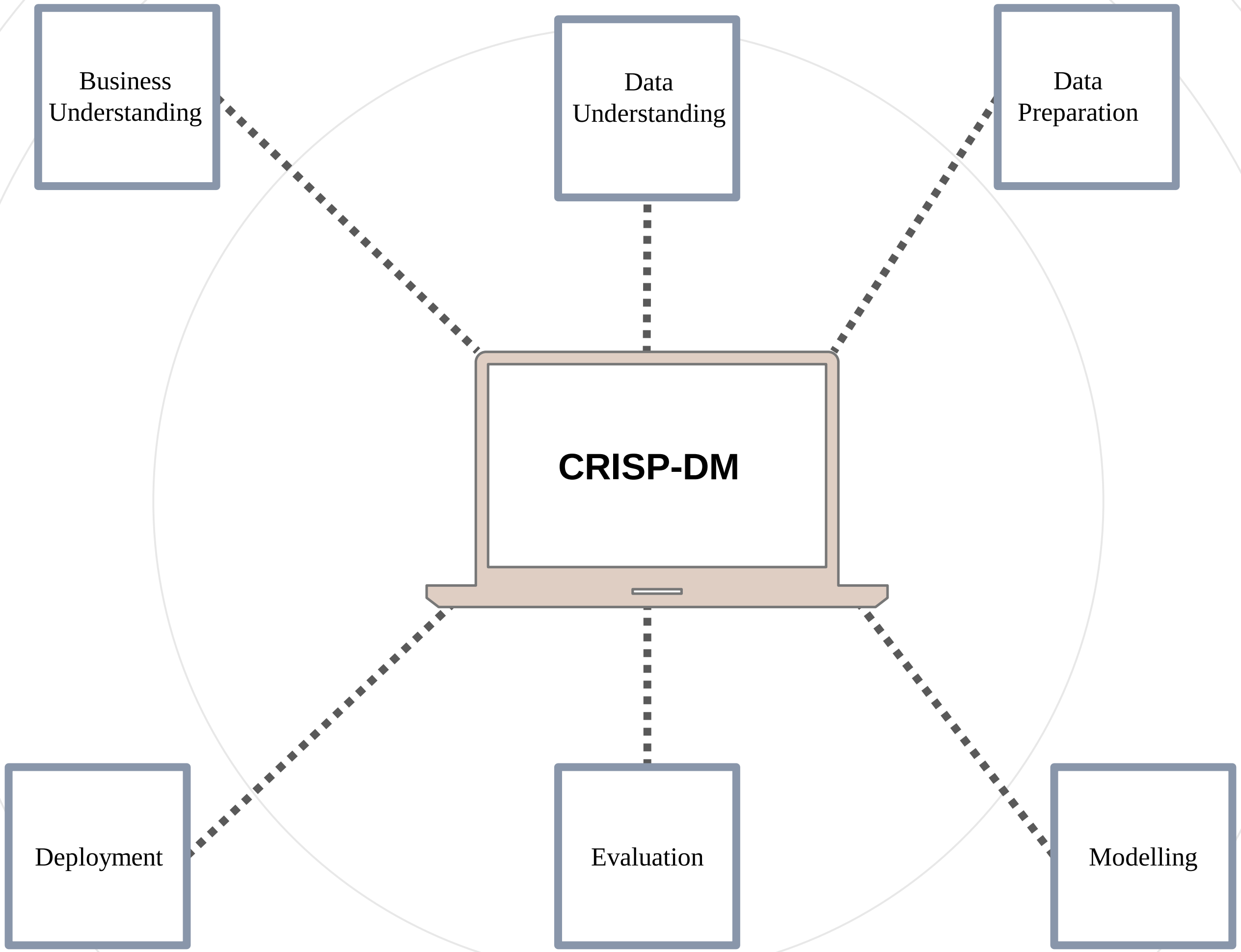
DEEP DIVE ANALYSIS OF EMPLOYEE ATTRITION

R we data mungers?


Business Understanding
Data Understanding
Data Preparation
CRISP-DM
Deployment
Evaluation
Modelling

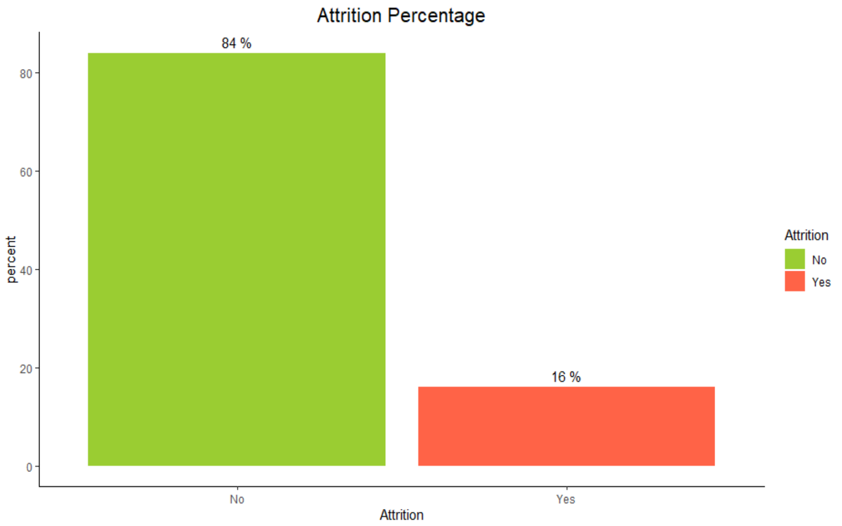
1

DATA EXPLORATION

Presenters:
1. Arjun Krishnan - 6622982
2. Prabha Krishnamoorthy - 6662827

Arjun Krishnan - 6622982

PROBLEM DEFINITION & OBJECTIVES

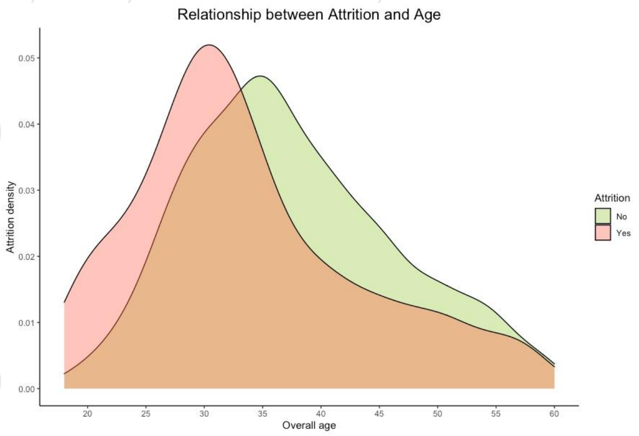


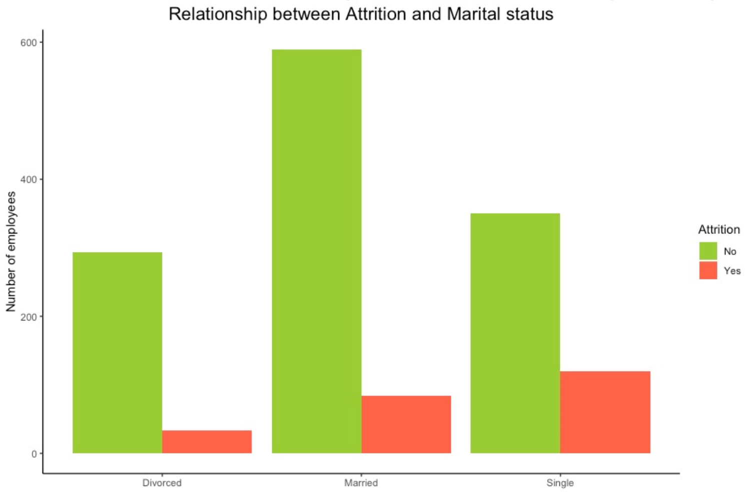
- ❑ IBM employee attrition dataset
- ❑ Significant part of the **investment** goes towards employees
- ❑ Manage attrition **within healthy threshold**
- ❑ Predicting attrition and the underlying reasons will **be helpful** in setting up reliable skilled teams, training programs and hiring processes.

- ❑ Identify which employees are **more likely** to leave the organization
- ❑ Identify what **factors** are decisive for employee's resignation
- ❑ Identify the likelihood of resignation from a **specific job role or department**

Arjun Krishnan - 6622982

EMPLOYEE PERSONAL DATA



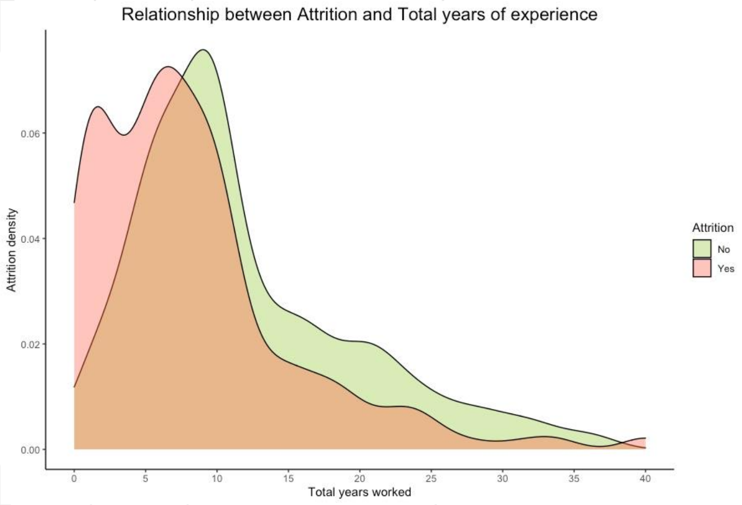


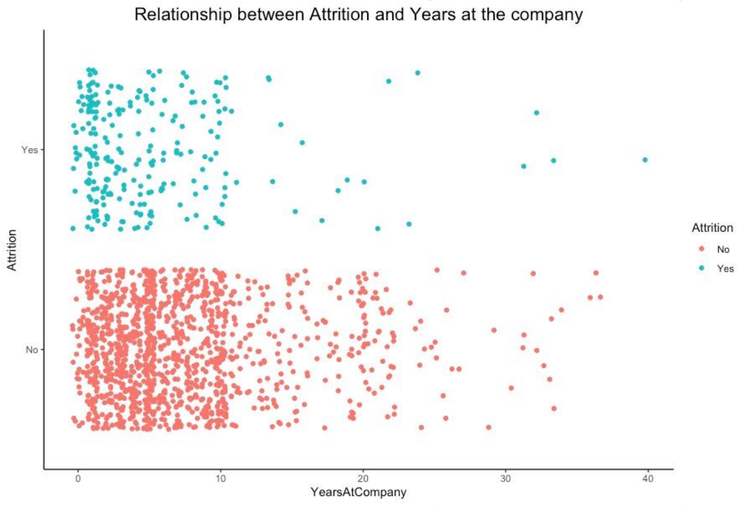
AGE: Attrition is higher in the age group between 20 and 30 and was visibly lower in age above 30.

MARITAL STATUS: Employees who are married are less likely to leave the organisation compared to single or divorced.

Arjun Krishnan - 6622982

EMPLOYEE TENURE DATA



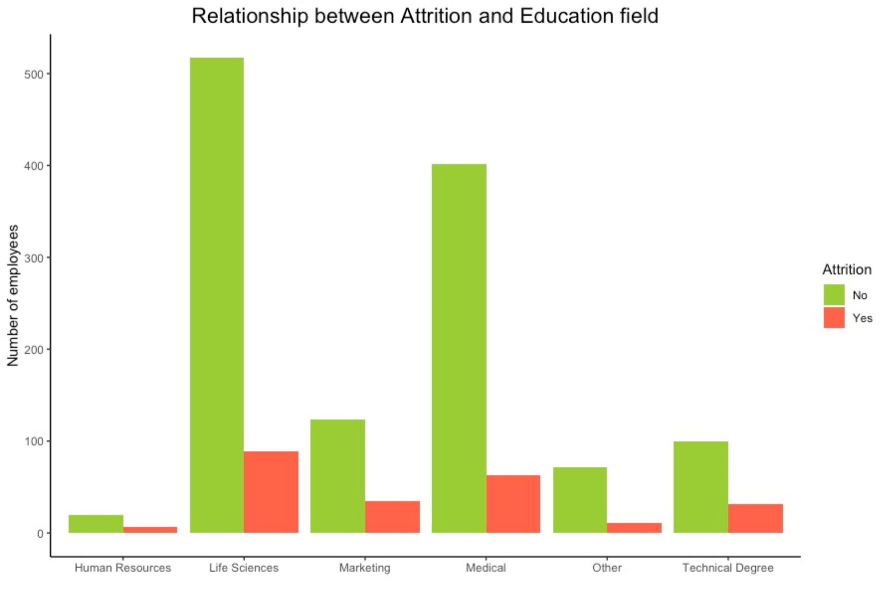


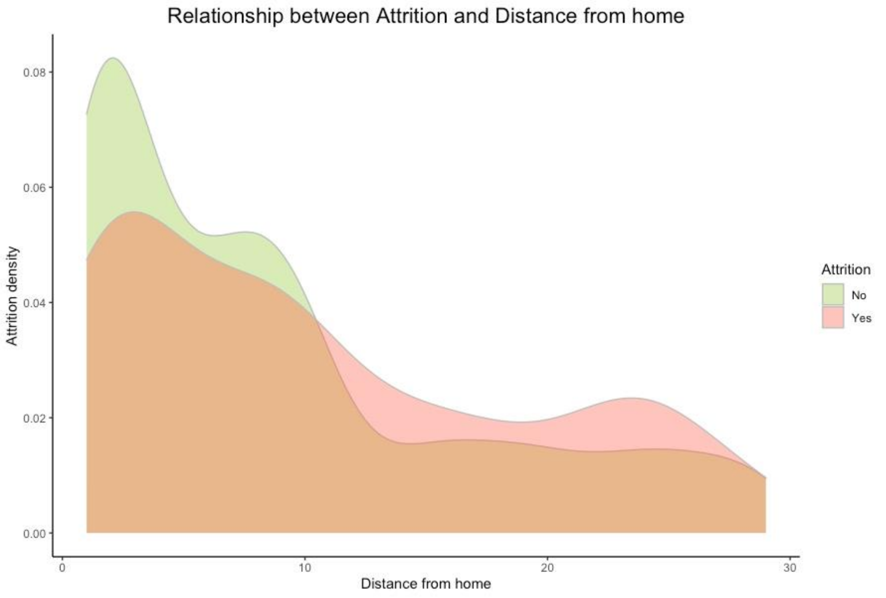
TOAL WORKING YEARS: Employees with **less than 10** years of overall experience leave the organisation.

YEARS AT COMPANY: Employees who spend less than 10 years leave the organisation. However, it is not significant enough to prove as a major reason for attrition.

Arjun Krishnan - 6622982

EMPLOYEE EDUCATION AND TRAVEL DATA



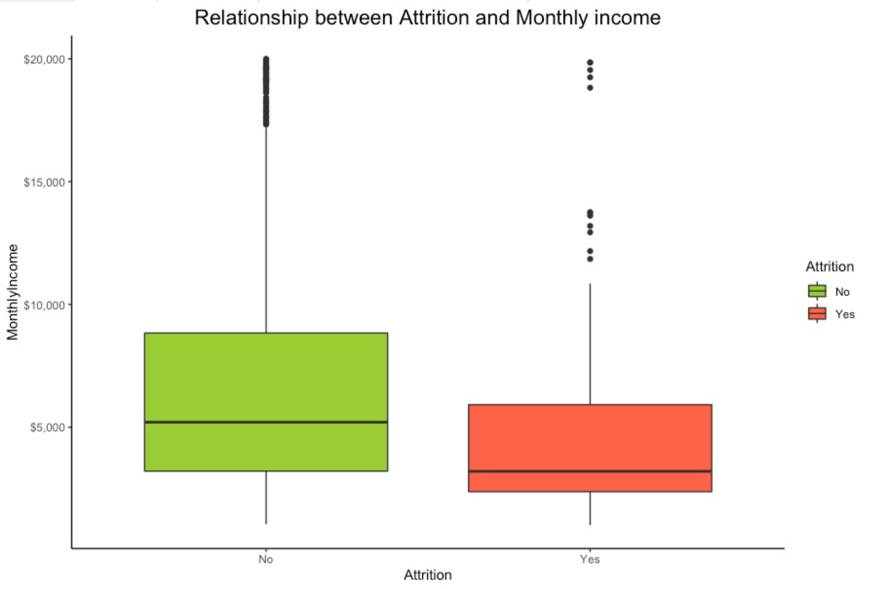


EDUCATION FIELD: Attrition seems to be **more prominent** in employees with education in human resources and technical degrees,

DISTANCE FROM HOME: Attrition level increase with the i**ncreased distance travelled** by the employees from the home to office.

Prabha Krishnamoorthy - 6662827

EMPLOYEE INCOME DATA

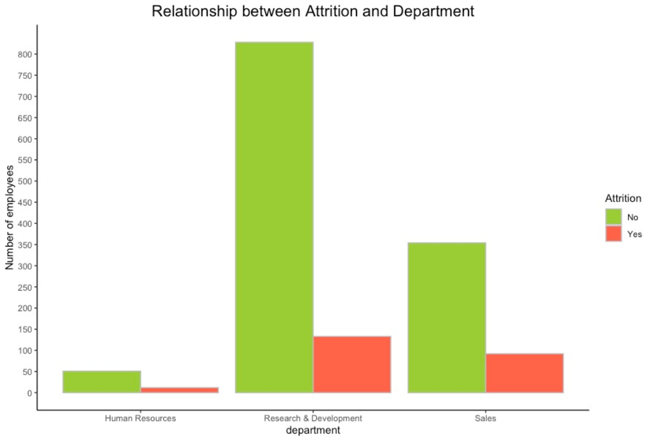


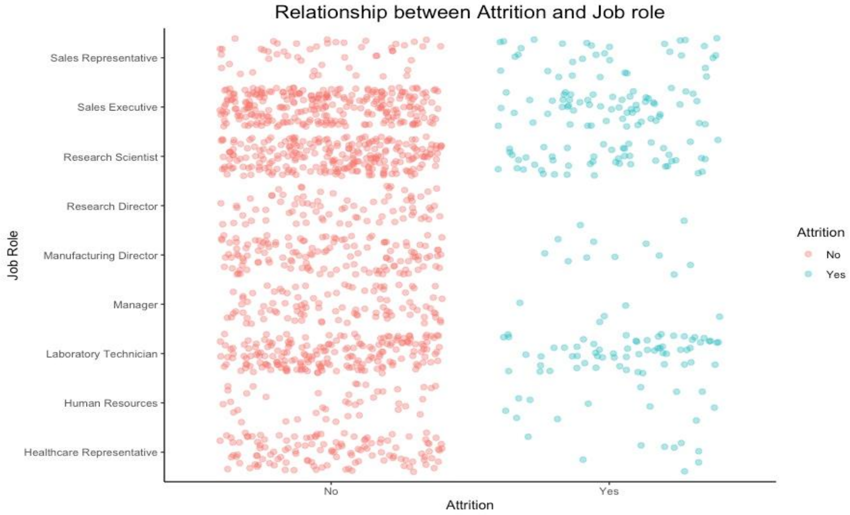
MONTHLY INCOME: Lower the monthly income (lesser than $5000), higher is the attrition.

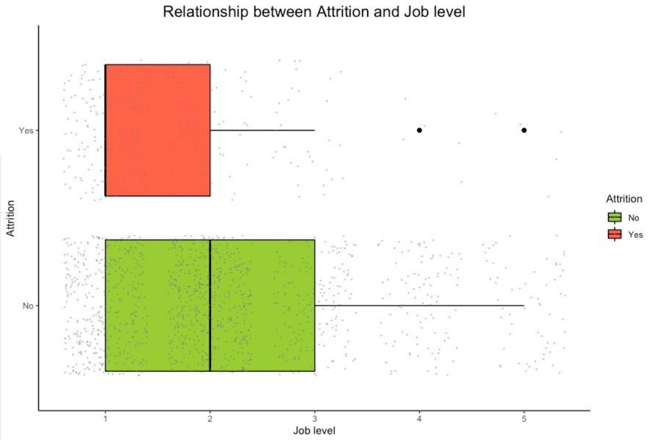
STOCK OPTIONS: No impact on employee attrition with or without stock options

Prabha Krishnamoorthy - 6662827

EMPLOYEE JOB DATA







DEPARTMENT: Attrition levels are higher in sales department given the total number of employees in sales compared to research and development.

JOB ROLES: Job roles like sales representatives, executives and healthcare representatives seem to have higher attrition levels compared to senior positions like manager, manufacturing and research directors

JOB LEVELS: Employees in lower job levels leave the organisation as compared to the ones in higher positions

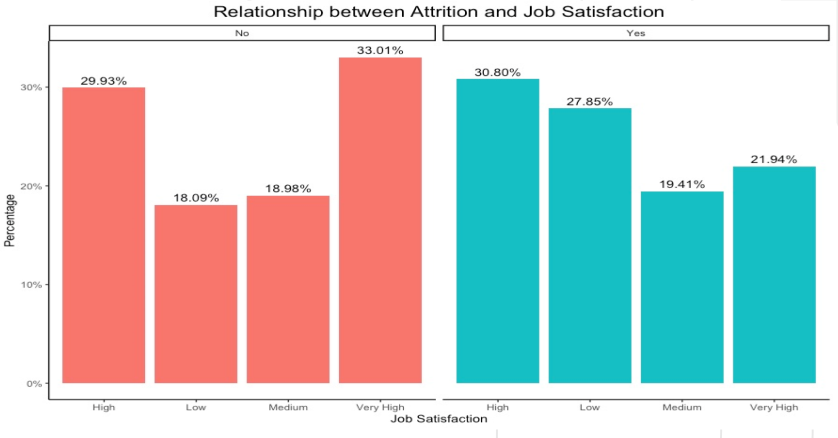
Prabha Krishnamoorthy - 6662827

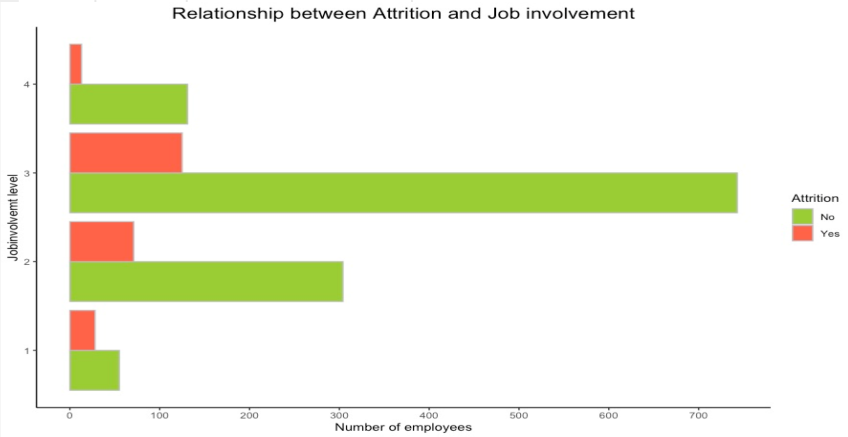
EMPLOYEE ENGAGEMENT DATA

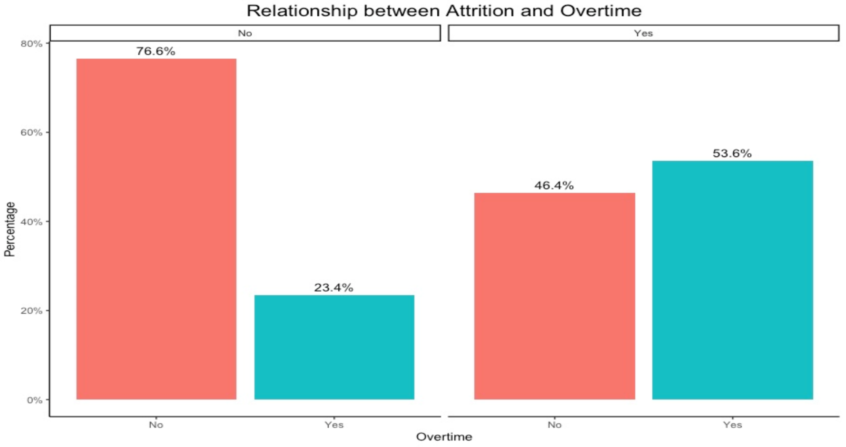
JOB SATISFACTION: Employees experiencing low or high job satisfaction have left the organisation. Job satisfaction could be one of the factor leading to attrition.

OVERTIME STATUS: Majority of the employees (54%) who worked overtime tend to leave the organization.

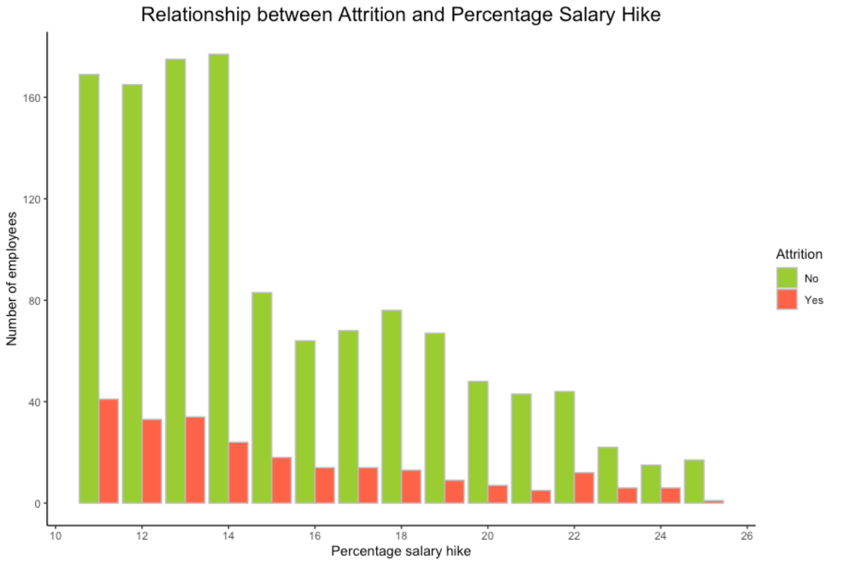
JOB INVOLVEMENT: People who have high job involvement have resigned from the organisation, when compared to medium level of involvement.

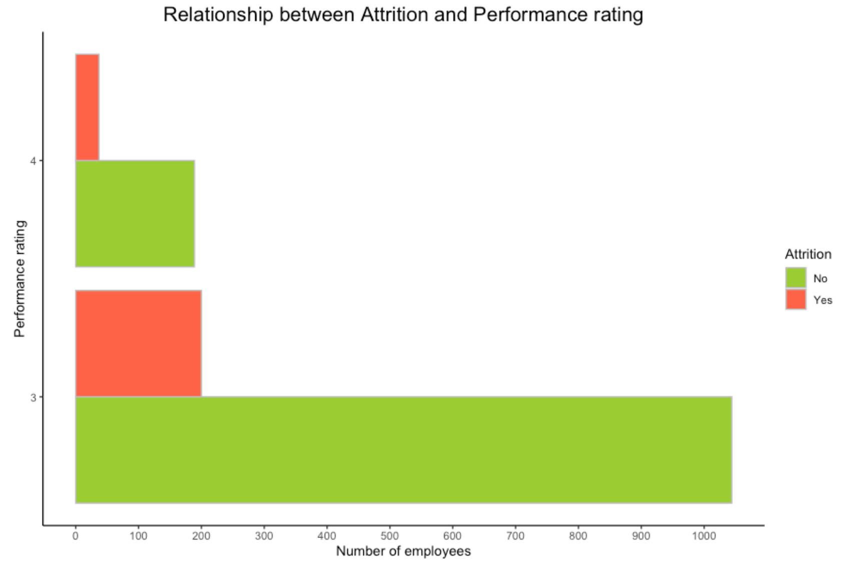






EMPLOYEE PERFORMANCE DATA





PERCENTAGE SALARY HIKE: Percentage salary hike data alone is not a key contributing factor to attrition since employees predominantly fall under the 15% salary hike and continue to be with the organisation. The percentage salary has a direct correlation with performance rating.

PERFORMANCE RATING: Dataset clearly demonstrated no impact on attrition, since the ratings cannot be exceptional for all the employees (As per HR standards, organisations will tend to have a bell curve distribution).

Prabha Krishnamoorthy - 6662827

DATA EXPLORATION - INFERENCES

Attributes leading to attrition

- **Personal Data** - Age, Marital Status
- **Tenure Data** -Years of experience, Years in the current role
- **Education Data** - Education field
- **Travel Data** - Distance from home
- **Income Data** - Monthly income
- **Job Data** - Department, Job roles, Job levels
- **Engagement Data** - Overtime Status

2

DATA PREPROCESSING

Presenters:
1. Rohini Raghukumar - 6659049
2. Sharath Kumar Muthu Anand Kumar - 6657482

DATA UNDERSTANDING

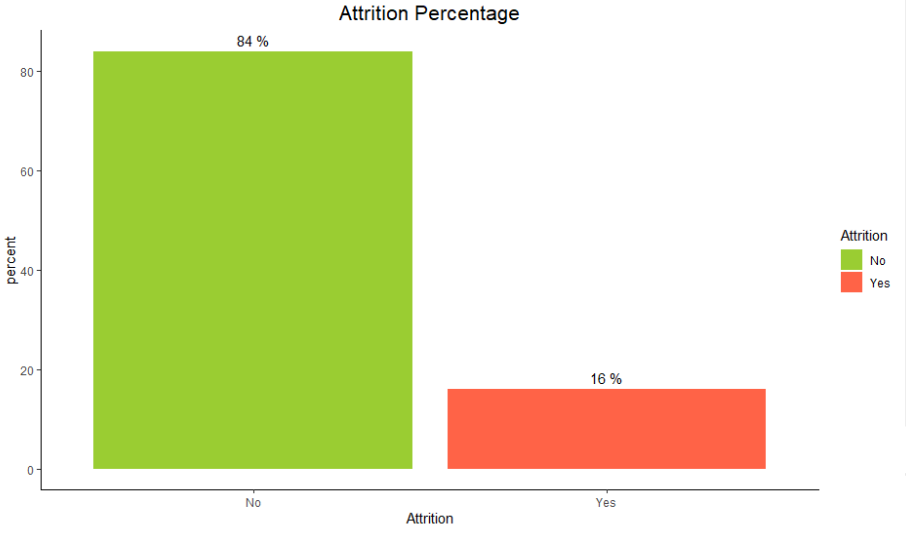
- Observing the number of features and records in the dataset
- Determining the independent and dependent variables for feature selection
- The dataset consists of 1470 observations with 35 variables
- The dataset does not have any NULL or missing values.

Rohini Raghukumar - 6659049

DATA UNDERSTANDING

Imbalance Check

- The column consists of 1233 No and 167 Yes, which clearly shows the imbalance of the target variable.
- The percentage of attrition identified from the attrition dataset was 16%



DATA UNDERSTANDING

Variability Check

The following variables do not have any variability in them.

- **EmployeeCount**: This consists of the number of employees and the value it takes is always 1.
- **Over18**: We observed that this variable denotes if an employee is 18 years of age and the value that it takes is ‘Yes’ is all cases.
- **StandardHours**: This depicts the number hours an employee has worked in a week. This has a constant value of 80.

Rohini Raghukumar - 6659049

DATA PREPARATION

Identifying and Handling outliers

- The outliers are identified using the percentile value of the field. They are handled using the percentile capping method. The fields with consisting of values greater than 95 percentile , the values which are greater than 95 percentile values are replaced by the 95-percentile value and values that are lesser than 5 percentile are replaced by the 5-percentile value.

- Using scatter plot we have identified the outliers and capped them using percentile capping method.

FIELD ENCODING

One Hot Encoding

- One-hot (dummy) encoding is applied to categorical features in this dataset, generating a binary column for each category
- Support vector machine is used for predictive analysis, One hot encoding is SVM friendly as it deals with numerical values.
- Even though this has a disadvantage of producing more fields it was manageable for this particular dataset

FEATURE SCALING

Min Max Normalization

- Continuous variables in different scales do not contribute equally to model fitting and the classification model might end up creating a bias.
- So, the continuous variables are scaled using Min Max Normalization in which the minimum value of feature gets transformed into a 0, the maximum value gets transformed to 1.

3

MODELLING

Presenters:
1. Prabesh B K - 6661057
2. Tsang Fan Yam - 6656440

DATA MINING GOALS

Which

h **employees are going to leave ?**

y

MODEL SELECTION

Decision Tree

Decision Tree is easy to understand and interpretable.

Random Forest

Performs better in predicting unseen data than Decision tree.

Support Vector machine

SVM is effective in dealing with dataset with considerable number of features (high dimensionality)

MEASURES

Confusion Matrix

		Actual	
		Attrition	No Attrition
Predicted	Attrition	TP	FP
	No Attrition	FN	TN

TP (True Positive)	Correct Attrition Predictions
FP(False Positive)	Misclassified Attrition
FN(False Negative)	Misclassified Non-Attrition
TN(True Negative)	Correct Non-Attrition predictions

MEASURES

Advanced Measures

Recall/Sensitivity

The ratio of number of predicted positive attrition to the actual number of attrition.

Recall= TP/(TP+FN)

Precision

The ratio of number of predicted attrition to the total predicted number of attrition.

Precision= TP/(TP+FN)

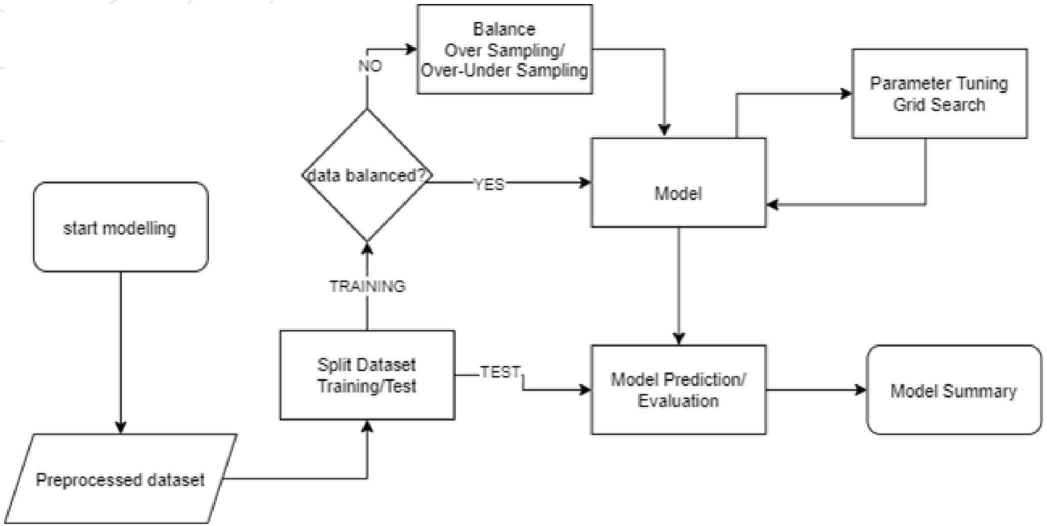
F1-Score

The harmonic mean of precision and recall.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

DECISION TREE AND RANDOM FOREST

Model Flow



DECISION TREE

Balancing Method

Random Oversampling of Minority Class was chosen

Over

Random Oversampling of minority and Random Under sampling of Majority Class due to good overall F1 score

model	TP	FN	TN	FP	↓ F1
DT- TRUE , 15	48	31	303	59	0.516
DT- TRUE , 30	52	27	288	74	0.507
DT- TRUE , 40	56	23	272	90	0.498

model	TP	FN	TN	FP	↓ F1
DT- FALSE , 70	56	23	267	95	0.487
DT- FALSE , 40	52	27	278	84	0.484
DT- FALSE , 60	55	24	268	94	0.482
DT- FALSE , 15	49	30	286	76	0.48
DT- FALSE , 20	51	28	273	89	0.466

DECISION TREE

Hypermeter Optimization

Hyperparameters	Range	Function
Winnow	TRUE / FALSE	feature selection of attributes
Boosting Iterations	Positive Integers	improves the result by converting weak learners into strong learners

Decision Tree Model Comparison

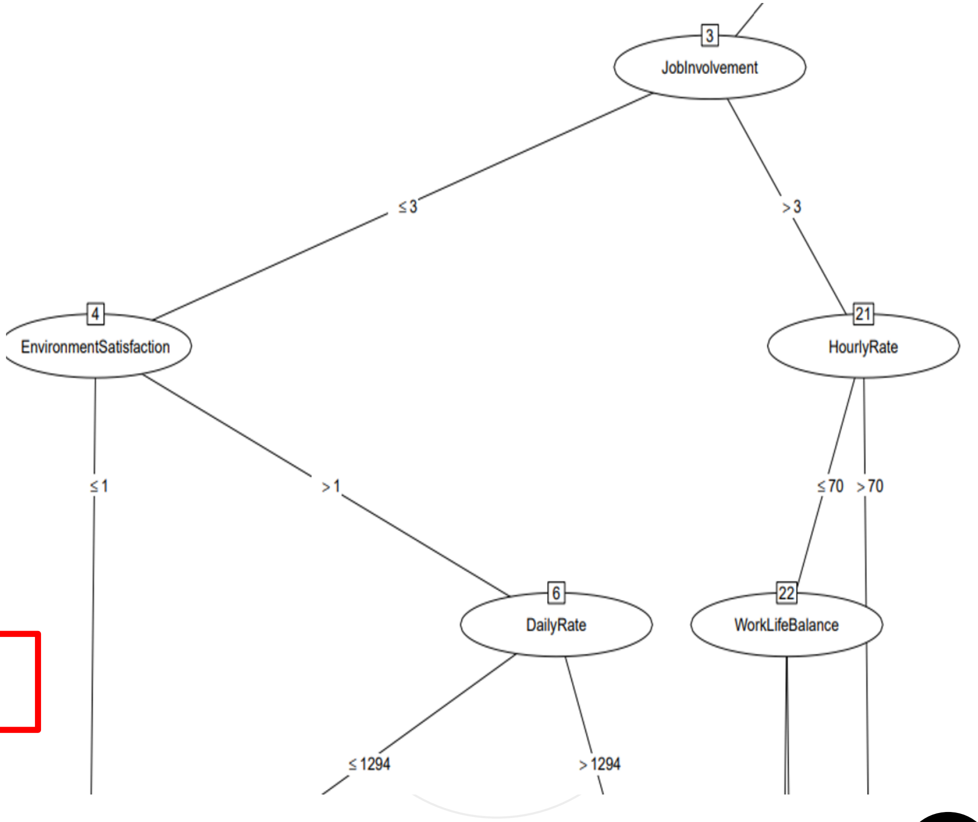
	model	TP	FN	TN	FP	Precision	Recall	F1
1	DecisionTree: T- , 1	48	27	288	78	38.0952	64	0.4776
2	DecisionTree: T- , 10	51	24	276	90	36.1702	68	0.4722
3	DecisionTree: T- , 22	54	21	272	94	36.4865	72	0.4843
4	DecisionTree: T- , 30	56	19	271	95	37.0861	74.6667	0.4956
5	DecisionTree: F- , 1	47	28	258	108	30.3226	62.6667	0.4087
6	DecisionTree: F- , 10	50	25	282	84	37.3134	66.6667	0.4785
7	DecisionTree: F- , 22	56	19	291	75	42.7481	74.6667	0.5437
8	DecisionTree: F- , 30	52	23	292	74	41.2698	69.3333	0.5174

Prabesh B K - 6661057

DECISION TREE

Rules and DT-Diagram

OverTime = Yes AND HourlyRate > 84 AND MonthlyIncome > 2973 AND StockOptionLevel <= 0 AND TrainingTimesLastYear > 2	NoAttrition
JobInvolvement > 3 AND MonthlyIncome <= 2973 AND WorkLifeBalance <= 2	NoAttrition
JobRole in {Sales Representative, Sales Executive,* Laboratory Technician} AND OverTime = Yes AND StockOptionLevel <= 0 AND TrainingTimesLastYear <= 2 AND YearsWithCurrManager <= 7	Attrition
BusinessTravel = Travel_Rarely AND OverTime = No AND DailyRate > 337 AND DailyRate <= 1062 AND EnvironmentSatisfaction <= 3 AND MonthlyIncome <= 2973 AND MonthlyRate <= 22670 AND PerformanceRating <= 3 AND StockOptionLevel <= 0 AND YearsInCurrentRole > 1	Attrition
OverTime = Yes AND DailyRate <= 1294 AND Education > 1 AND JobInvolvement <= 3 AND MonthlyIncome <= 2973 AND RelationshipSatisfaction > 2 AND RelationshipSatisfaction <= 3	Attrition
JobRole in {Manager, Manufacturing Director, Research Scientist} AND EnvironmentSatisfaction <= 2 AND JobInvolvement <= 1 AND StockOptionLevel <= 0	Attrition

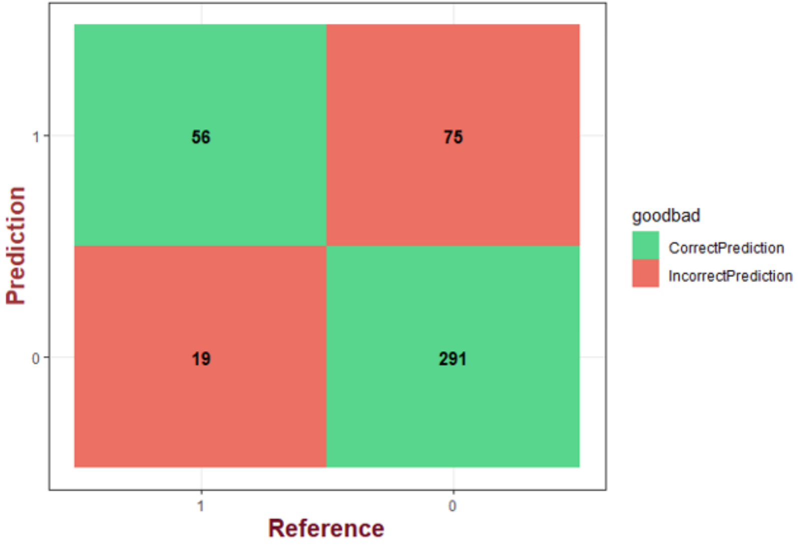


Prabesh B K - 6661057

DECISION TREE

Results

Decision Tree boost= 22



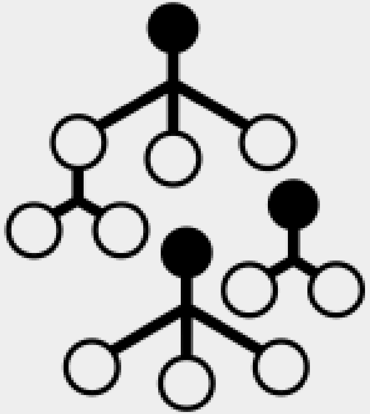
Attribute Usages

	Strength		
JobRole	100.00	YearsWithCurrManager	100.00
MaritalStatus	100.00	DailyRate	99.94
OverTime	100.00	BusinessTravel	99.71
Age	100.00	MonthlyRate	99.71
EnvironmentSatisfaction	100.00	EducationField	99.35
HourlyRate	100.00	JobLevel	97.94
JobInvolvement	100.00	DistanceFromHome	97.35
MonthlyIncome	100.00	TrainingTimesLastYear	97.23
StockOptionLevel	100.00	PercentSalaryHike	97.06
TotalWorkingYears	100.00	WorkLifeBalance	96.70
YearsAtCompany	100.00	NumCompaniesWorked	95.94
YearsInCurrentRole	100.00	RelationshipSatisfaction	95.88

RANDOM FOREST

Balancing Method

Random Oversampling of Minority Class was chosen **Over** Random Oversampling of minority and Random Under sampling of Majority Class due to good overall F1 score



Hypermeter Optimization

Hyperparameters	Range	Function
Number of Trees	Positive Integers	Could improve the ability to handle data with high dimensionality and large size

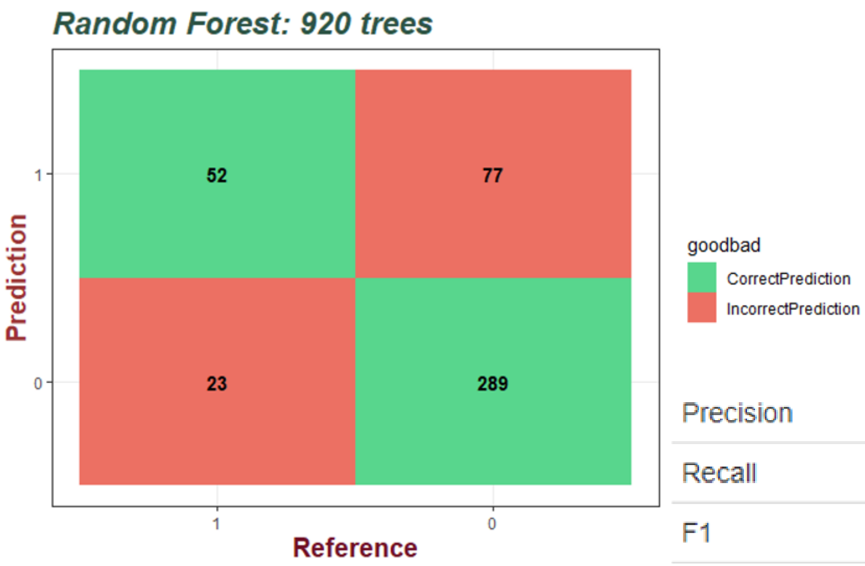
Random Forest Model Comparison

	model	TP	FN	TN	FP	Precision	Recall	F1
1	RF 100	52	23	282	84	38.2353	69.3333	0.4929
2	RF 400	53	22	277	89	37.3239	70.6667	0.4885
3	RF 750	53	22	286	80	39.8496	70.6667	0.5096
4	RF 920	52	23	289	77	40.3101	69.3333	0.5098
5	RF 1000	60	15	252	114	34.4828	80	0.4819
6	RF 1400	53	22	286	80	39.8496	70.6667	0.5096

Prabesh B K - 6661057

RANDOM FOREST

Results



Precision	40.3101
Recall	69.3333
F1	0.5098

Attribute Usages	Strength
MonthlyRate	58.46189
DailyRate	57.00605
HourlyRate	56.60596
MonthlyIncome	56.29213
DistanceFromHome	52.94924
OverTime	52.90263
PercentSalaryHike	52.41634
Age	48.12519
RelationshipSatisfaction	45.20093
JobSatisfaction	44.33758
NumCompaniesWorked	44.18953
TrainingTimesLastYear	43.20816

ATTRIBUTE USAGES

Most closely related attributes are Monthly Income, Hourly Rate, Overtime and Age which are common features from both the model.

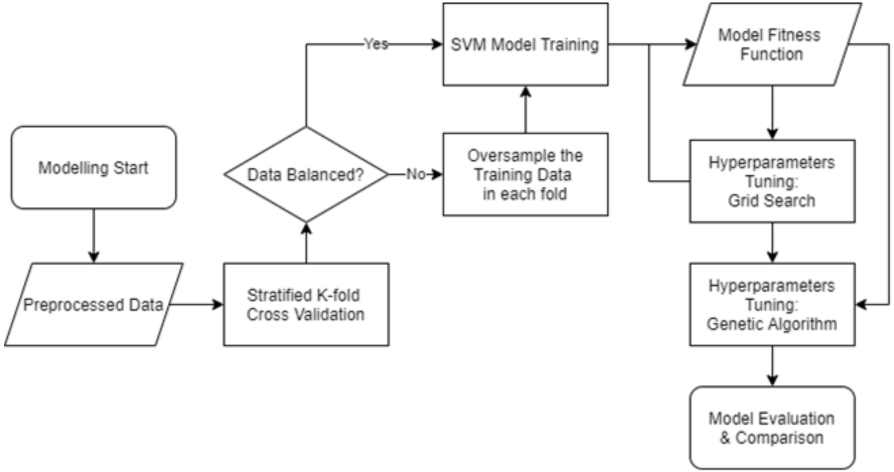
From Random Forest:
Monthly Rate, Daily Rate, Distance From Home, Percent Salary Hike are some other strong attributes.

Since DT attribute usages was high where13 attributes were used completely, it is unclear to determine attributes leading to attrition.

Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

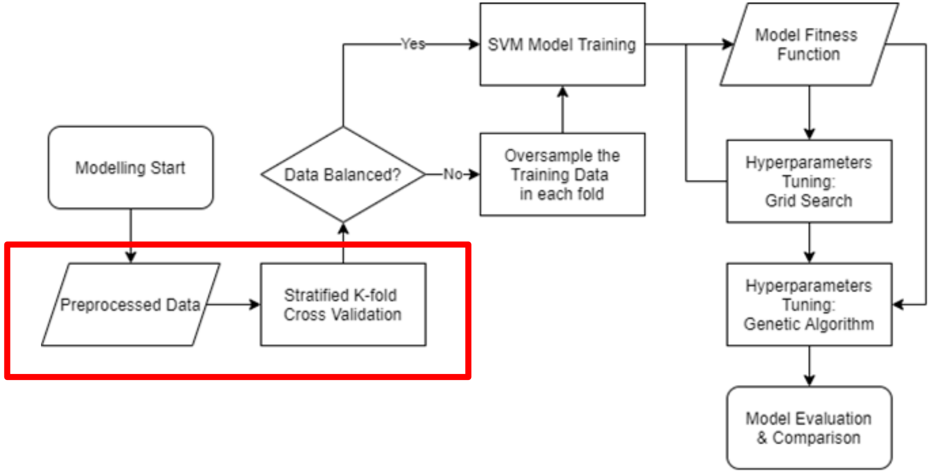
- ❑ How to validate the training model?
- ❑ Data balanced?
- ❑ Hyperparameters Tuning?
- ❑ Is SVM good comparing to other models?



Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

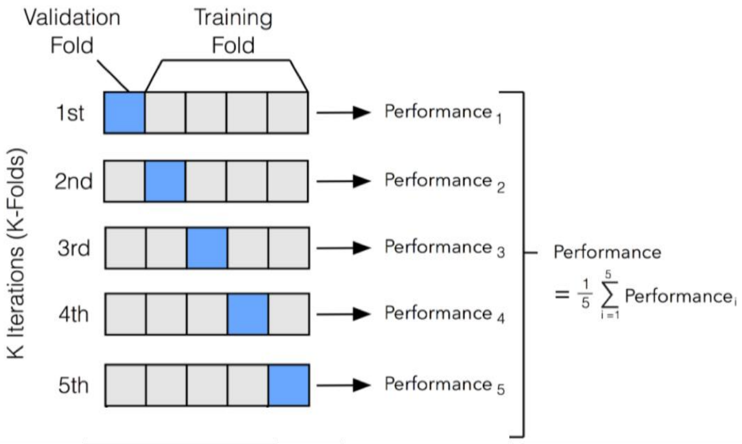
- ❑ How to validate the training model?
- ❑ Data balanced?
- ❑ Hyperparameters Tuning?
- ❑ Is SVM good comparing to other models?



SUPPORT VECTOR MACHINE

Stratified K-Fold Cross Validation

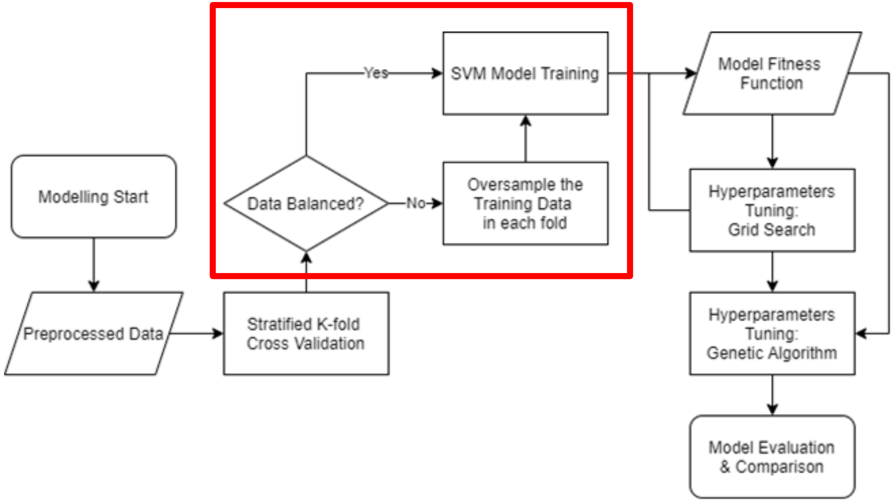
- 5 Fold in total
- 4 Fold: Training (1176 cases)
- 1 Fold: Testing (294 cases)
- Repeating 5 times to obtain a better generalised model evaluation



Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

- ❑ How to validate the training model?
- ❑ Data balanced?
- ❑ Hyperparameters Tuning?
- ❑ Is SVM good comparing to other models?



Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

Oversampling Techniques

- ❑ Random Oversampling
- ❑ Synthetic Minority Oversampling Technique (SMOTE)

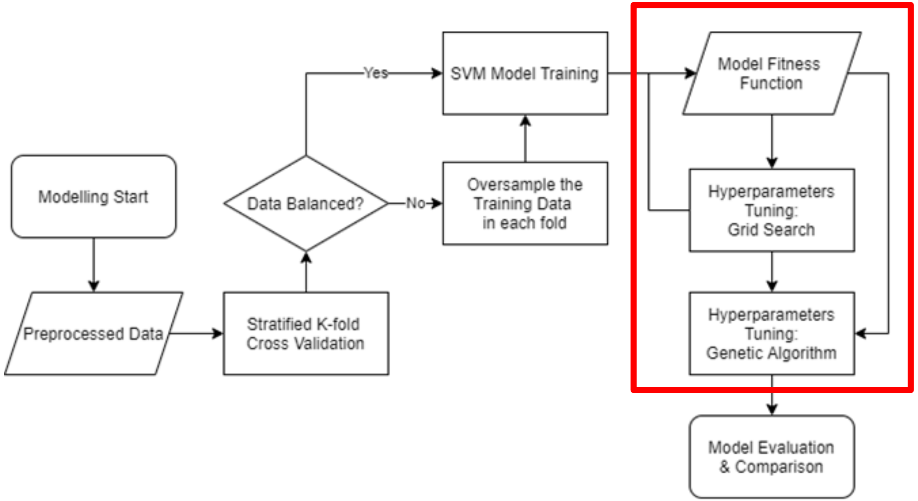
	TP	FN	TN	FP	Precision	Recall	F1
SVM with Random oversampling	56	19	296	70	44.44	74.67	0.5572
SVM with SMOTE oversampling	56	19	293	73	43.41	74.67	0.549

- ❑ Same as Decision Tree and Random Forest, **Random Oversampling** is chosen due to better F1-Score

Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

- ❑ How to validate the training model?
- ❑ Data balanced?
- ❑ Hyperparameters Tuning?
- ❑ Is SVM good comparing to other models?

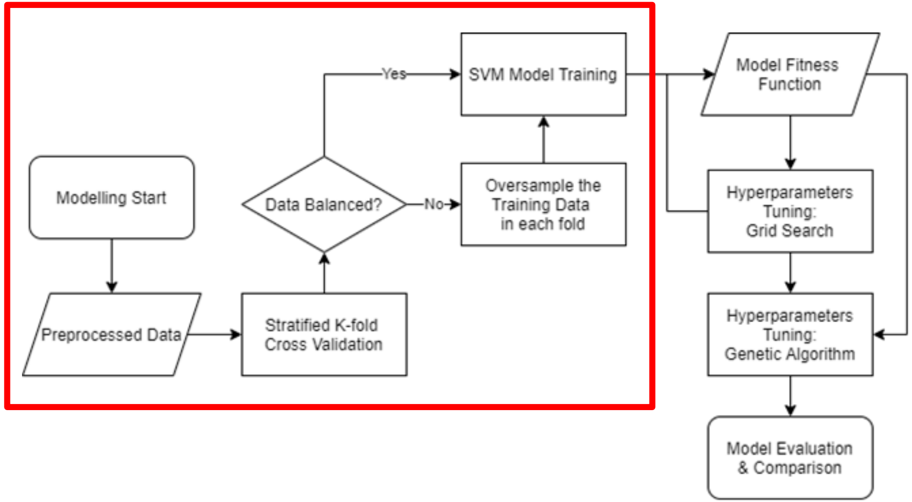


Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

Hyperparameter Optimisation: Fitness Function

- ❑ Hyperparameters: Cost & Gamma
- ❑ Setting up a **fitness function** to evaluate the model
- ❑ To make sure the parameters chosen are **not overfitting** to the training data
- ❑ **Return the mean F1-score** of 5-Fold Stratified Cross Validation



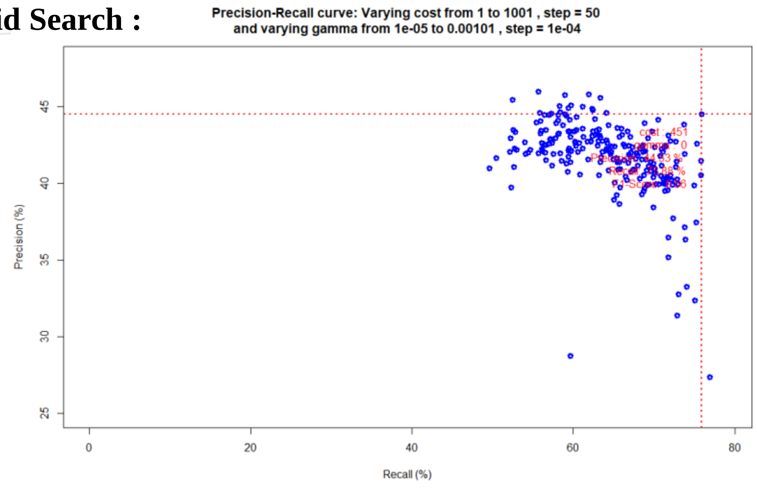
SUPPORT VECTOR MACHINE

Hyperparameter Optimisation: Grid Search + GA

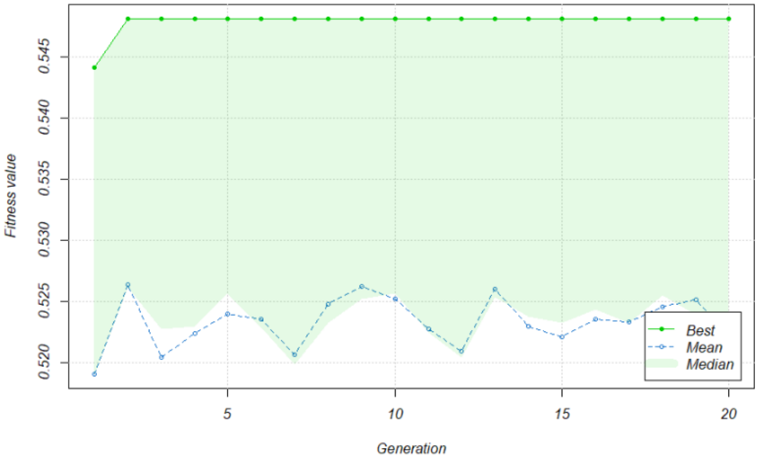
- Grid search: Computationally expensive, to narrow down the search space (cost: 451, gamma: 0.00001)
- Genetic Algorithm(GA): Search for more precise parameters
- Best cost(c): 441.6821, Best gamma: 0.0000227

```
-- Genetic Algorithm -------------------

GA settings:
Type                  =  real-valued
Population size       =  50
Number of generations =  20
Elitism               =  2
Crossover probability =  0.8
Mutation probability  =  0.1
Search domain =
      gamma   c
lower 1e-05 400
upper 1e-04 500

GA results:
Iterations             = 20
Fitness function value = 0.5481185
Solution =
        gamma        c
[1,] 2.27884e-05 441.6821
SVM Genetic algorithm ended.
```

Grid Search :



Genetic Algorithm :



Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

Model Assessment

- The model with optimised hyperparameters perform the best
- Not possible to visualise the model due to high dimensionality of data (A lot of features)

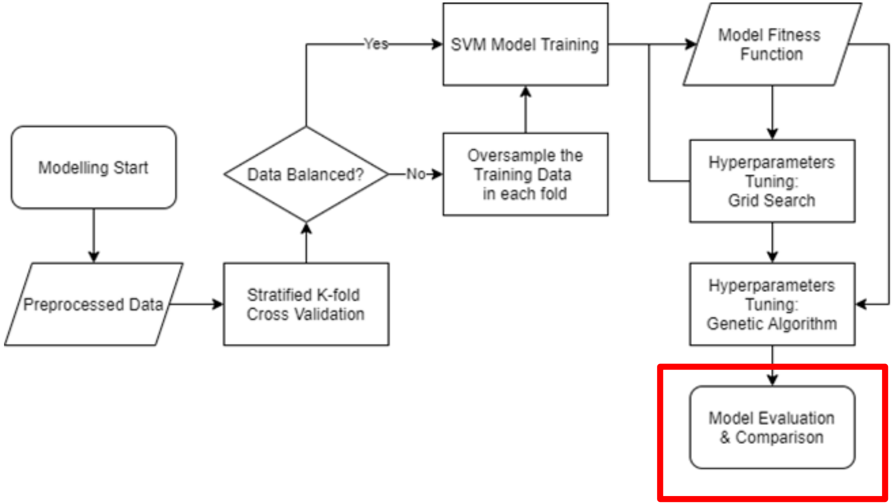
SVM Model Comparison

	model	TP	FN	TN	FP	Precision	Recall	F1
Model1	SVM: cost- 441.6 , gamma- 2.7e-05	58	17	292	74	43.9394	77.3333	0.5604
Model2	SVM: cost- 2000 , gamma- 0.01	51	24	291	75	40.4762	68	0.5075
Model3	SVM: cost- 700 , gamma- 0.05	54	21	290	76	41.5385	72	0.5268
Model4	SVM: cost- 100 , gamma- 0.1	51	24	310	56	47.6636	68	0.5604

Tsang Fan Yam - 6656440

SUPPORT VECTOR MACHINE

- ❑ How to validate the training model?
- ❑ Data balanced?
- ❑ Hyperparameters Tuning?
- ❑ Is SVM good comparing to other models?



Tsang Fan Yam - 6656440

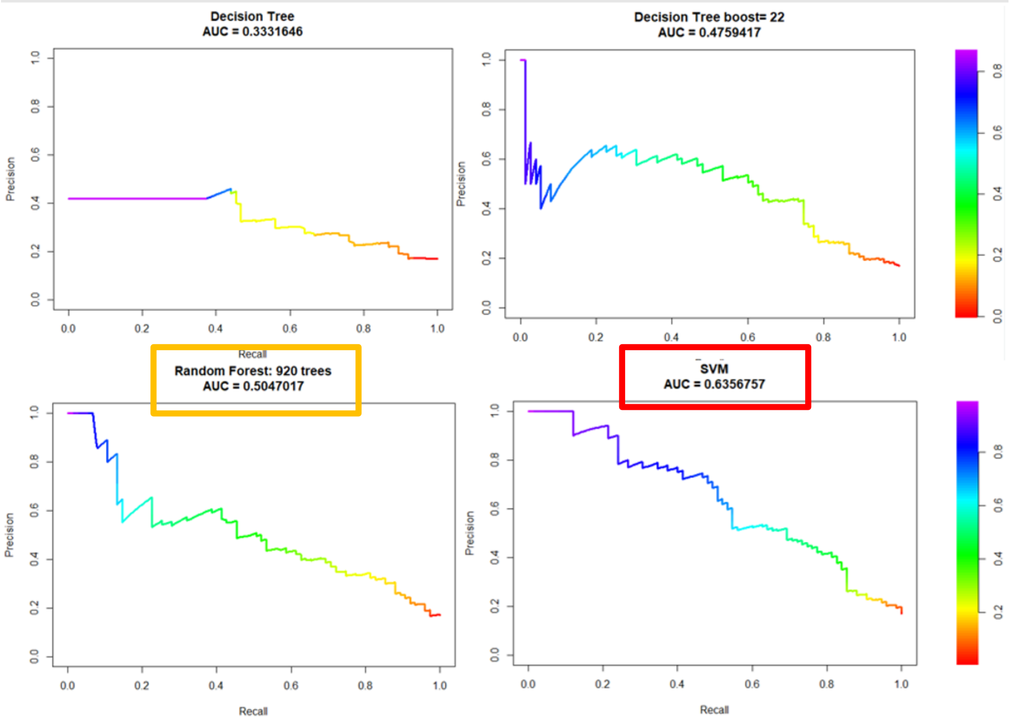
MODEL COMPARISON

- Basic Measures: TP, FN, TN, FP
- Advanced Measures: Precision, Recall, F1-score
- Precision-Recall curve
- Good **tradeoff** between Precision and Recall: F1-score

Tsang Fan Yam - 6656440

MODEL COMPARISON

- ❑ Precision-recall curve of SVM is the best
- ❑ Best Area Under the Curve of 0.63



Tsang Fan Yam - 6656440

MODEL COMPARISON - SUMMARY

- ✓ Satisfactory in predicting which employees are more likely to leave (Recall)
- ✓ SVM: Best in prediction
- ✓ Boosted DT and RF: Generate business rules/closely-related attributes

- ✗ Low precision in the prediction
- ✗ SVM is a black box
- ✗ Recall of Boosted DT & RF is much lower than SVM

	TP	FN	TN	FP	Precision (%)	Recall (%)	F1-Score	AUC (ROC)	AUC (PR Curve)
DT	47	28	258	108	30.32	62.67	0.409	0.706	0.330
Boosted DT	56	19	291	75	42.7481	74.6667	0.5437	0.7912	0.4759
RF	52	23	289	77	40.3101	69.3333	0.5098	0.802	0.504
SVM	61	14	282	84	42.069	81.3333	0.5545	0.844	0.636

4

EVALUATION

Presenters:
Sharath Kumar Muthu Anand Kumar - 6657482

EVALUATION

SVM

- Support Vector Machine performs the best in predicting “Attrition”

Random Forest

Random Forest was successful in identifying the most decisive attributes leading to attrition. Below few are notable factors :

- Monthly Income
- Hourly Rate
- Overtime
- Age

EVALUATION

Potential profit each TP:
Recruitment cost avoided × Chance of staying – Training cost
= ($4043*7.5) * 30% - $1094 = $8002
= Potential profit from each TP × 61
= $8002 × 61
= **$488,167**

Potential cost from each FP:
Extra training cost of around $1094 is spent on each misclassified employee.
= Potential cost from each FP × 84
= £1094 × 84
= **$91,896**

Therefore, the potential net profit is equal to **£(488167– 91896)**, which is **$396,271**

FUTURE DEPLOYMENT

- ❑ The deployment of SVM and Random Forest models is recommended
- ❑ The SVM model could potentially save the organisation $396,271 in operational costs during the year, which would prove to be a huge benefit in due course of time
- ❑ Random Forest is used to identify factors leading to attrition
- ❑ More data collection is recommended, to further improve the performance of the current models and to help deploy any other model in the future, for any new business requirements that might arise