Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 45 No. 1 (2024)

Evaluation of Machine Learning Based Selected Predictive Models for Voluntary Employee Turnover

[1] Mohan Sangli, [2]Rajeshwar S Kadadevaramath, [3]Jerin Joseph, [4]B.Latha Shankar

[1]Research Scholar Industrial Engineering Department, Siddaganga institute of Technology Tumkur, Karnataka, India

[2]Professor (Rtd), Industrial Engineering Department, Siddaganga institute of Technology Tumkur, Karnataka, India

[3]Research Scholar Industrial Engineering Department, Siddaganga institute of Technology Tumkur, Karnataka, India

[4]Asso.Professor, Industrial Engineering Department, Siddaganga institute of Technology Tumkur, Karnataka, India

***Abstract*:** Employee attrition has become one of the most significant problems for any organization. Employees are important assets of the organization and the subjects who own other valuable resources that the organization need, diverse opportunity costs occur when employee attrition takes place. To prevent such unwanted loss of valuable assets, various efforts have been made to predict and prevent employee attrition. Various methods have been developed, in order to predict employee turnover, including statistical models and machine learning techniques. Statistical models, such as logistic regression and survival analysis, have been used to identify the relationships between different predictor variables and employee turnover. Machine learning algorithms, have also been applied to extract sensitive parameters and better turnover prediction by removing parameters that are not important and instead add to noise.

Keywords: Employee attrition, Machine learning, random forest, XG Boost, Artificial Neural Network

1 Introduction:

The manufacturing industry is one of the largest contributors to India's economy, accounting for nearly 16% of the country's Gross Domestic Product (GDP). The service industry is another significant contributor to India's economy, accounting for nearly 55% of the country's GDP. The sector employs over 28% of the total workforce in India and hence study plays a very critical role where we find huge attrition rate compared to manufacturing industries.

The goal here is to critically examine the variables that significantly impact an employee's likelihood of attrition. An investigation into how distance from home, net income, age, environment among others impact an employee's attrition. This paper will probe for specific answers for such queries using Machine Learning &AI algorithms.

2 Literature Survey :

2.1 Employee Turnover:

Employee attrition, also known as employee turnover, refers to the process of losing employees from an organization. A strong organizational culture can negatively associated with turnover intentions [1] Another study

found that employees who perceive their organization as having a positive culture are less likely to leave [2] . The use of machine learning methods such as Artificial Neural Networks (ANNs) and support vector machines (SVMs) have been shown to be effective in handling large and complex datasets and have been widely used in the literature. [3]

2.2 Machine Learning:

This literature study is to review the most commonly used techniques and algorithms in machine learning and their applications. One of the most popular techniques in machine learning is supervised learning. learning algorithms include linear regression, logistic regression, and Support Vector Machines (SVMs) [4]. Common unsupervised learning algorithms include k-means clustering, hierarchical clustering, and Principal Component Analysis[5] A third popular technique in machine learning is reinforcement learning. This technique is often used in robotics, gaming, and decision making [6]. Deep learning algorithms include convolutional neural networks (CNNs) and Recurrent Neural Networks (RNNs) [7]. When we use multiple methods, there is a need for rank aggregation i.e to get to the right parameters in two step process [8]. In their articles, various researches have touched upon the coefficients of trained models such as elastic net [9] , [10] , Gini impurity or variance reduction scores in decision tree(regressor) [11], extra tree(regressor) [12], random forest(regressor) [13], and Extreme Gradient Boosting (XGBoost) [14] . These are the n preference lists whose aggregate we must find. For finding the rank aggregate, we use the robust rank aggregation method proposed [15].

No literature seems to be available where comparisons are made on the results of ranking method on parameter importance and ranking method on permutation importance to arrive at the reliable parameter importance. Combination of Ranking methods, permutation importance to select parameters will further improve accuracy of models.

3.0 Research Objectives:

1. To explore methods to identify and extract the sensitive parameter (dimensionality reduction) and Comparing the parameters in dimensionally reduced sets using ranking methods and permutation on importance of parameters with parameter importance and permutation importance.

2. Arrive at a framework that provide reliable set of sensitive parameters in the data set to improve the efficiency of predictive model on employee turnover.

4.0 Methodology :

The approach section of this paper will focus on the use of dimensionality reduction techniques for employee turnover prediction models.

4.1 Machine learning concepts

Machine learning is a subfield of artificial intelligence that involves the development of algorithms and statistical models that enable computers to learn from data and make predictions or decisions without being explicitly programmed. There are several key concepts in machine learning that are important to understand to effectively use and apply these techniques.

4.2 Machine Learning Classification Metrics:

Machine Learning Classification Metrics are Accuracy, Precision, Recall, F1 Score, Specificity. AUC-ROC Curve (Area Under the Receiver Operating Characteristic Curve): A measure of the performance of a binary classifier, it plots the true positive rate against the false positive rate at various threshold settings.

4.3 Parameters

Other nomenclature used for parameters are attributes, field, column, variable, dimension, parameters, and design variables. In an input dataset for supervised machine learning, the Parameters are input attributes used for prediction or classification and Parameter Selection methods used in this study

5.0 Employee Turnover: Data Collection Analysis and Discussions

5.1 Experimentation

For this experiment we used the HR employee attrition dataset used by IBM for their Watson demo. This dataset is a publicly available dataset used for employee turnover prediction studies. The dataset contains 1470 employee records and 35 attributes that describe various employee characteristics, such as age, gender, education, job satisfaction, performance rating, work-life balance, and salary. The dataset is a real-world dataset, Overall, the dataset is highly relevant to employee turnover prediction studies and reflects the characteristics of employees in actual organizations.

5.2 The file characteristics

Following is list of parameters, parameter type and the possible values when the parameter type is categorical shown in table 1.

Table 1 Characteristics of Data file used in experiment.

SL. No.	Parameter	Type	Description	Allowed values, if any
1	Age	Numeric	the age of the employee	
2	Attrition	Binary	whether the employee left the company	"Yes" or "No"
3	BusinessTravel	Categorical	the frequency of travel for business	"Non-Travel", "Travel_Rarely", "Travel_Frequently"
4	DailyRate	Numeric	the employee's daily pay rate	
5	Department	Categorical	the department in which the employee works	"Research & Development", "Sales", "Human Resources"
6	DistanceFromHome	Numeric	distance between the employee's home and work	
7	Education	Categorical	the level of education of the employee	"Below College", "College", "Bachelor", "Master", "Doctor"
8	EducationField	Categorical	the field of study of the employee	"Human Resources", "Life Sciences", "Marketing", "Medical", "Technical Degree", "Other"
9	EmployeeCount	Numeric	the number of employees in the company	
10	EmployeeNumber	Numeric	the employee's identification number	
11	EnvironmentSatisfaction	Categorical	the employee's level of satisfaction with the work environment	"Low", "Medium", "High", "Very High"

12	Gender	Binary	the gender of the employee	"Male" or "Female"
13	HourlyRate	Numeric	the employee's hourly pay rate	
14	JobInvolvement	Categorical	the level of involvement the employee has in their job	"Low", "Medium", "High", "Very High"
15	JobLevel	Categorical	the level of the employee's job	"1", "2", "3", "4", "5"
16	JobRole	Categorical	the role of the employee	"Sales Executive", "Research Scientist", "Laboratory Technician", "Manufacturing Director", "Healthcare Representative", "Manager", "Sales Representative", "Research Director", "Human Resources"
17	JobSatisfaction	Categorical	the employee's level of satisfaction with their job	"Low", "Medium", "High", "Very High"
18	MaritalStatus	Categorical	he marital status of the employee	"Single", "Married", "Divorced"
19	MonthlyIncome	Numeric	the employee's monthly income	
20	MonthlyRate	Numeric	the employee's monthly pay rate	
21	NumCompaniesWorked	Numeric	the number of companies the employee has worked for	
22	Over18	Binary	whether the employee is over 18 years of age	"Y" or "N"
23	OverTime	Binary	whether the employee works overtime	"Yes" or "No"
24	PercentSalaryHike	Categorical	the percentage increase in the employee's salary	
25	PerformanceRating	Categorical	the employee's performance rating	"Low", "Good", "Excellent", "Outstanding"
26	RelationshipSatisfaction	Categorical	the employee's level of satisfaction with their relationship with coworkers	"Low", "Medium", "High", "Very High"

27	StandardHours	Numeric	the number of standard hours for the employee's role	
28	StockOptionLevel	Categorical	the level of stock options given to the employee	"0", "1", "2", "3"
29	TotalWorkingYears	Numeric	the total number of years the employee has worked	
30	TrainingTimesLastYear	Numeric	the number of training sessions the employee attended last year	
31	WorkLifeBalance	Categorical	the employee's work-life balance	"Bad", "Good", "Better", "Best"
32	YearsAtCompany	Numeric	the number of years the employee has been with the company	
33	YearsInCurrentRole	Numeric	the number of years the employee has been in their current role	
34	YearsSinceLastPromotion	Numeric	the number of years since the employee's last promotion	
35	YearsWithCurrManager	Numeric	the number of years the employee has worked with their current manager	

Using an autoML tool, the exploratory analysis and the base models were created.

5.3 Exploratory Data Analysis

The table 2 below presents the data summary:

Table 2 Data file size summary

Data cleansing summary		**Records summary**	
Total Records:	1470	Records considered for analysis:	1470
Total Parameters:	35	Records ignored due to missing values in target parameter:	0
Total unique categorical levels:	31	Records removed due to duplicates:	
Percentage of missing data:	0 (0 cells)		

5.4 Parameter elimination through parameter characteristics (statistical)

The statistical based parameter elimination was done using the methods discussed and following table 3 is the summary of parameters retained and dropped.

Table 3 Statistical parameter elimination *results*

Parameter Summary	**Numeric (continuous)**	**Numerical (Ordinal)**	**Categorical**
Total parameters	5	21	9
Parameters dropped	0	3	1
Uni-Level	0	2	1
		Standard Hours	Over18
		Employee Count	
Near constant	0	1	0
		Performance rating	
Parameters used for analysis	5	18	8
	Monthly Rate	Relationship Satisfaction	Business Travel
	Hourly Rate	Years At Company	Attrition
	Daily Rate	Years With CurrManager	Job Role
	Monthly Income	Stock Option Level	Marital Status
	Employee Number	Job Satisfaction	Over Time
		Work Life Balance	Education Field
		Education	Department
		Percent Salary Hike	Gender
		Total Working Years	
		Environment Satisfaction	
		Distance From Home	
		Job Level	
		Years In Current Role	
		Job Involvement	
		Age	
		Years Since Last Promotion	
		Num Companies Worked	
		Training Times Last Year	

Missing Values: All parameters had 100% values and hence no parameters were removed due to missing values for parameters

Low Variance in a parameter: Following parameters were filtered out as they were Uni-level or near constant values in a parameter and as it will not have much importance in predicting the target.

Uni-Level drops: *Employee Count:* had constant value of 1 ,*Over18:* had constant value of "yes" as employees will be above this age anyway, *Standard Hours*: this was set to 80 for all employees; normally full time employee

wrk 80 hours in every pay cycle of 2 weeks,Near-Constant drops, *Performance rating:* 226 out of 1470 records only had a rating of 4. Removed as this distribution is 15.5% to 84.5%, a case of class imbalance

High Correlation: Table 4 shows the correlation between the parameters **.** JobLevel and MonthlyIncome has highest correlation of 0.95 and still has a merit to not filter out. No other pair is with significantly high correlation to filter out.

Table 4 Higher Correlations

Parameters	JobLevel	MonthlyIncome	PercentSalaryHike	PerformanceRating	TotalWorkingYears	YearsAtCompany	YearsInCurrentRole	YearsWithCurrManage
JobLevel		**0.950**	-0.035	-0.021	**0.782**	0.535	0.389	0.375
MonthlyIncome	**0.950**		-0.027	-0.017	**0.773**	0.514	0.364	0.344
PercentSalaryHike	-0.035	-0.027		**0.774**	-0.021	-0.036	-0.002	-0.012
PerformanceRating	-0.021	-0.017	**0.774**		0.007	0.003	0.035	0.023
TotalWorkingYears	**0.782**	**0.773**	-0.021	0.007		0.628	0.460	0.459
YearsAtCompany	0.535	0.514	-0.036	0.003	0.628		**0.759**	**0.769**
YearsInCurrentRole	0.389	0.364	-0.002	0.035	0.460	**0.759**		**0.714**
YearsWithCurrManager	0.375	0.344	-0.012	0.023	0.459	**0.769**	**0.714**	

High Cardinality: Employee Number is unique and hence but numerical. It does assist in predicting and hence retained. There were no categorical parameters with high cardinality.

5.5 Initial parameter importance score

With Attrition as target parameter, the parameter extraction through random forest resulted in following parameter importance table 5.

Table 5 Parameter importance

#	Parameter Name	Importance Score (%)		#	Parameter Name	Importance Score (%)	
		Individual	Cumulative			Individual	Cumulative
1	Employee Number	8.35	8.35	18	Education Field	2.13	80.95
2	Monthly Rate	7.94	16.29	19	Years Since Last Promotion	2.13	83.08
3	Monthly Income	7.78	24.07	20	Job Involvement	1.97	85.05
4	Daily Rate	7.12	31.19	21	Years In Current Role	1.91	86.96
5	Age	6	37.19	22	Job Level	1.81	88.77
6	Hourly Rate	5.11	42.3	23	Stock Option Level	1.78	90.55

7	Distance From Home	4.67	46.97	24	Gender	1.72	92.27
8	Job Role	3.97	50.94	25	Training Times Last Year	1.65	93.92
9	Environment Satisfaction	3.84	54.78	26	Marital Status	1.59	95.51
10	Percent Salary Hike	3.68	58.46	27	Work Life Balance	1.46	96.97
11	Over Time	3.59	62.05	28	Education	1.43	98.4
12	Job Satisfaction	3.02	65.07	29	Business Travel	1.4	99.8
13	Years At Company	2.92	67.99	30	Department	0.19	99.99
14	Relationship Satisfaction	2.83	70.82	31	Employee Count	Uni-Level - dropped	
15	Num Companies Worked	2.76	73.58	32	Over 18	Uni-Level - dropped	
16	Years With CurrManager	2.67	76.25	33	Standard Hours	Uni-Level - dropped	
17	Total Working Years	2.57	78.82	34	Performance Rating	Near Const - dropped	

Discussion: The set target of 90% cumulative importance of parameters reached at top 23 parameters i.e. 77% of all parameters and even the remaining 7 parameters had decent important score unlike in many datasets the top 90% importance is covered by 40-60% parameters and last few will have near zero importance. This suggests the dataset was already prepared well for predictive modelling.

Using AutoML tool several models were generated on this all parameter dataset with parameter importance of 100%. Following table 6 shows various models and their performance and the winning model is highlighted.

Table 6 Base models performance when all parameters are used

Name	Accuracy	Precision	Recall	F1 statistic	Specificity	Mis classification rate	Area Under Curve(AUC)
Hyp.Xtreme Gradient Boosting	**87.03**	**89.76**	**95.92**	**92.73**	**38.25**	**12.97**	**0.791**
Xtreme Gradient Boosting	86.14	89.28	95.13	92.09	35.41	13.86	0.784
AdaBoost	86.55	89.14	95.84	92.36	33.79	13.45	0.744
Logistic Regression	87.03	89.08	97.44	93.04	32.59	12.97	0.801
Hyp.Random Forest	86.35	88.00	97.20	92.35	25.43	13.66	0.772
Decision Tree	76.90	87.61	84.80	86.17	32.77	23.10	0.588
Ensemble Model	86.69	87.23	98.80	92.64	18.47	13.32	0.687
Random Forest	86.42	86.43	99.68	92.57	11.57	13.59	0.790
SGD Classifier	84.99	84.99	100.00	91.88	-	15.01	0.500

The average confusion matrix with K-Means of 4 showed the result as below in the table 7:

Table 7 Confusion matrix on test set of base model

All K		Predicted	
		No - (P)	Yes - (N)
Actual	No - (P)	298	15
	Yes - (N)	32	24
	P = Positive N = Negative		

Discussion: Though the Hyp.Xtreame Gradient Boosting and Logistic Regression have slight edge over other algorithms, all the models have performed with decent accuracies, i.e. most of them above 80% accuracy or within 10% of top performing algorithm.

5.6 Parameter elimination through parameter importance from multiple models and ranking method

Table 8 shows the generated parameter importance from multiple algorithms and then applied a ranking algorithms to arrive at importance rank of all parameters and arranged in sorted order. Then we cut a dataset with ranked parameters with importance cumulating to 90%. Figure 1 shows selected correlations and sensitive parameters from another method.

Table 8 Parameter importance from multiple algorithms and ranks

Rank	Feature	LR	Ridge	Decision Tree	Random Forest	XGBoost	CatBoost	SVM	average_scores	cumsum
1	JobRole	16.59	13.62	5.19	4.07	21.96	4.59	19.89	12.27	12.27
2	OverTime	5.65	5.71	4.78	4.63	2.09	7.46	4.92	5.03	17.31
3	WorkLifeBalance	4.37	4.26	3.57	3.49	7.25	3.83	3.78	4.36	21.67
4	MonthlyIncome	4.57	6.01	6.22	6.43	1.06	5.39	4.56	4.89	26.56
5	StockOptionLevel	5.64	5.56	4.17	3.45	10.42	6.18	4.85	5.75	32.32
6	NumCompaniesWorked	3.88	4.03	5.96	3.13	1.73	4.67	3.58	3.85	36.17
7	TotalWorkingYears	3.51	3.84	13.71	5.44	2.94	2.62	3.28	5.05	41.22
8	JobLevel	8.92	11.64	0.86	3.14	3.01	2.78	9.47	5.69	46.91
9	EnvironmentSatisfaction	4.65	4.67	3.55	2.85	4.66	4.56	4.12	4.15	51.06
10	Age	1.41	2.21	7.58	5.18	1.25	5.78	1.24	3.52	54.58

11	**JobSatisfaction**	**3.68**	**3.7**	**0.77**	**3.34**	**5.36**	**3.92**	**3.22**	**3.43**	**58.01**
12	**YearsWithCurrManager**	**2.76**	**1.72**	**1.32**	**3.84**	**1.14**	**3.82**	**2.61**	**2.46**	**60.47**
13	**RelationshipSatisfaction**	**4.12**	**3.54**	**0.66**	**3.12**	**9.56**	**4.75**	**3.5**	**4.18**	**64.64**
14	**JobInvolvement**	**3.49**	**4.49**	**1.47**	**2.8**	**4.96**	**2.26**	**3.32**	**3.26**	**67.9**
15	**YearsAtCompany**	**1.51**	**1.74**	**2.59**	**3.69**	**0.94**	**2.5**	**1.8**	**2.11**	**70.01**
16	**Department**	**4.14**	**3.42**	**4.23**	**1.84**	**2.03**	**2.87**	**7.3**	**3.69**	**73.7**
17	**DistanceFromHome**	**2.74**	**1.62**	**3.08**	**3.12**	**1.1**	**3.43**	**2.29**	**2.48**	**76.18**
18	**EducationField**	**3.46**	**4.8**	**0.54**	**3.15**	**3.82**	**2.04**	**3.44**	**3.04**	**79.22**
19	**BusinessTravel**	**4.54**	**3.78**	**0.91**	**1.96**	**4.03**	**2.39**	**3.78**	**3.06**	**82.27**
20	**YearsSinceLastPromotion**	**2.34**	**1.53**	**0.85**	**2.25**	**0.91**	**2.56**	**2**	**1.78**	**84.05**
21	**Education**	**0.84**	**1.53**	**1.11**	**2.62**	**1.85**	**1.95**	**0.55**	**1.49**	**85.54**
22	**YearsInCurrentRole**	**1.55**	**1.7**	**3.06**	**2.37**	**1.3**	**1.61**	**1.56**	**1.88**	**87.42**
23	**DailyRate**	**1.25**	**0.76**	**4.78**	**4.1**	**0.76**	**3.08**	**1.06**	**2.26**	**89.68**
24	TrainingTimesLastYear	1.33	1.05	0.93	2.04	1.12	1.77	1.2	1.35	91.03
25	HourlyRate	0.3	0.67	5.57	3.87	0.64	3.7	0.26	2.14	93.17
26	MonthlyRate	0.64	0.57	4.41	3.62	0.5	2.31	0.72	1.82	95
27	MaritalStatus	0.59	0.44	1.05	2.72	1.7	1.93	0.68	1.3	96.3
28	Gender	1.08	0.75	1.59	1.29	0.79	0.84	0.87	1.03	97.33
29	EmployeeNumber	0.04	0.47	2.73	3.83	0.48	2.31	0.05	1.42	98.74
30	PercentSalaryHike	0.4	0.19	2.78	2.59	0.65	2.1	0.09	1.26	100

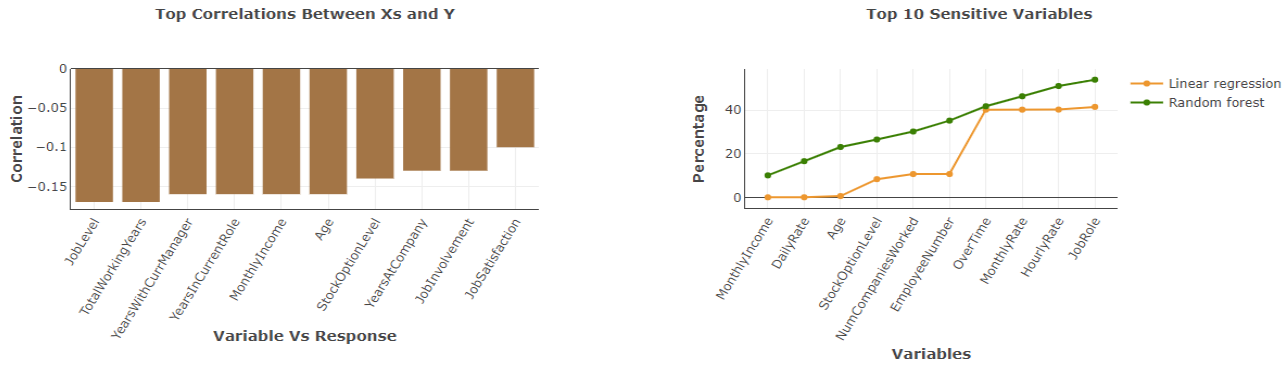

Figure 1 selected correlations and sensitive parameters from another method

Using AutoML tool several models were generated on this 90% cumulative parameter importance. Following were various models and their performance and the winning model is highlighted. The model performance and confusion matrix for this dataset is depicted below in table 9 and 10 respectively.

Table 9 Ranked parameters based model performance post dimensionality reduction

Name	Accuracy	Precision	Recall	F1 statistic	Specificity	Mis classification rate	Area Under Curve(AUC)
Hyp.Xtreme Gradient Boosting	**86.89**	**90.12**	**94.94**	**92.46**	**41.53**	**13.11**	**0.802**
Xtreme Gradient Boosting	85.40	89.60	93.67	91.58	39.21	14.61	0.801
AdaBoost	87.09	89.27	96.39	92.68	34.50	12.91	0.755
Logistic Regression	87.02	89.25	96.29	92.62	34.30	12.98	0.802
Decision Tree	78.94	88.73	86.16	87.42	38.37	21.06	0.623
Hyp.Random Forest	86.89	88.33	97.43	92.65	27.70	13.11	0.789
Ensemble Model	87.23	88.00	98.39	92.89	24.11	12.77	0.707
Random Forest	86.75	86.69	99.76	92.75	13.85	13.25	0.801
SGD Classifier	85.06	85.05	100.00	91.91	0.54	14.95	0.503

The confusion Matrix on test file is as shown below:

Table 10 Confusion matrix from ranked parameter importance model

All K		Predicted	
		No - (P)	Yes - (N)
Actual	No - (P)	297	16
	Yes - (N)	33	23
	P = Positive N = Negative		

Discussion: As expected, there is no significant improvement in the performance of models but with 90% cumulative importance from the 23 i.e. 77% of parameters (68% of original dataset) the model still performed with similar or better accuracy when compared with models using all the 30 i.e 100% parameters.

5.7 Parameter elimination through permutation importance from multiple models and ranking method

Table 11 depicts the generated permutation importance from multiple algorithms and then applied a ranking algorithms to arrive at importance rank of all parameters and arranged in sorted order. Then we cut a dataset with ranked parameters with importance cumulating to 90%.

Table 11 Permuted parameter importance from multiple algorithms and ranks

Ra	Feature							SVM		
1	OverTime								16.00	
2	MonthlyIncome								9.08	
3	StockOptionLev								5.63	
4	NumCompanies								6.15	
5	WorkLifeBalan								7.54	
6	TotalWorkingY								4.81	
7	YearsWithCurr								4.22	
8	Age								6.14	
9	RelationshipSati								3.18	
10	JobRole								2.96	
11	EnvironmentSat								3.18	
12	DistanceFromH								2.27	
13	JobInvolvement								2.39	
14	JobLevel								2.91	
15	EducationField								2.17	
16	YearsInCurrent								1.37	
17	BusinessTravel								2.17	
18	YearsAtCompa								1.90	
19	YearsSinceLast								1.49	
20	TrainingTimesL								1.13	
21	JobSatisfaction								2.04	
22	MonthlyRate								1.18	
23	DailyRate								2.03	
24	HourlyRate								1.78	
25	Department								2.43	
26	EmployeeNumbe								0.93	
27	PercentSalaryHik								0.86	
28	Education								0.78	
29	MaritalStatus								0.70	
30	Gender								0.59	

This resulted in 22 parameters, i.e. 1 lesser than ranked parameter importance in creating dimensionally reduced dataset. With Ranking on Permutation importance, permutation importance post ranking of selected correlations and sensitive parameters from another method

Is shown in figure 2

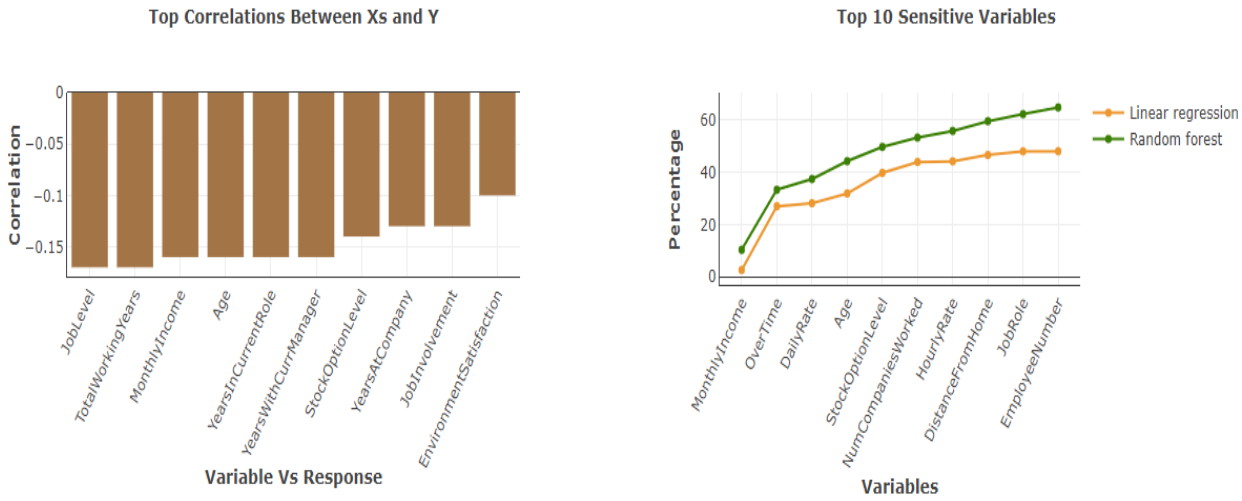

Figure 2 permutation importance post ranking of selected correlations and sensitive parameters from another method

Using AutoML tool several models were generated on this 90% cumulative parameter importance. Following were various models and their performance and the winning model is highlighted.

The performance of various models on this dataset and the confusion matrix is shown below in table 12 :

Table 12 Ranked permuted parameters based model performance post dimensionality reduction

Name	Accuracy	Precision	Recall	F1 statistic	Specificity	Mis classification rate	Area Under Curve(AUC)
Hyp.Xtreme Gradient Boosting	**87.43**	**90.37**	**95.36**	**92.79**	**42.97**	**12.57**	**0.808**
Xtreme Gradient Boosting	86.35	90.05	94.30	92.12	41.61	13.66	0.799
AdaBoost	87.37	89.76	96.08	92.80	38.14	12.64	0.758
Decision Tree	78.80	88.41	86.43	87.38	36.08	21.20	0.613
Ensemble Model	87.84	88.24	98.88	93.24	25.77	12.16	0.702
Hyp.Random Forest	87.09	88.20	97.92	92.79	26.31	12.91	0.781
Logistic Regression	87.23	87.70	98.98	92.95	22.51	12.77	0.791
SGD Classifier	81.61	87.67	79.82	77.00	23.69	18.40	0.518
Random Forest	87.10	87.04	99.68	92.91	16.58	12.91	0.792

The confusion Matrix on K-mean of 4 is presented below in table 13 :

Table 13 Confusion matrix from ranked permuted parameter importance model

All K		Predicted	
		No - (P)	Yes - (N)
Actual	No - (P)	298	15
	Yes - (N)	32	24
	P = Positive N = Negative		

Discussion: There is a slight improvement in the performance of models even though only 22 of 30 parameters, i.e. 73% of parameters used in the base models without ranking. Following table 14 shows the model performance summary of all three experiments:

Table 14 comparative model performances from three methods used

Metric of winning models	Base Model	Feature importance + Rank	Permutation Importance + Rank
Accuracy	0.870	0.869	0.874
Precision	0.898	0.901	0.904
Recall	0.959	0.949	0.954
F1 statistics	0.927	0.925	0.928
Specificity	0.383	0.415	0.430
Misclassification	0.130	0.131	0.126
AUC	0.791	0.802	0.808

The metrics have not deteriorated in the dimensionally reduced models. While parameter importance or ranked parameter importance did not show improvement, the dimensionally reduced dataset based on Permutation importance showed slightly better performance. As this dataset was created by IBM and used for research in many studies, the dataset is already sharpened for the machine learning studies. Having this slightest improvement with the reduced number of parameters can be attributed to the permutation importance and the ranking of parameters with ranking algorithm.

5.8 Consistency of parameters as top predictors:

We then did another side study on consistency of the top parameters in number of parameters within the top 3, top 5 and top 8 before and after dimensionality reduction by both methods and shown in table15. We used original ranking of parameters before model tuning for comparing with the parameters after we tuned few selected algorithms such Ridge, XGBoost, and Linear SVM.

Table 15 Consistency of parameter importance before and after dimensionality reduced tuned models

Consistency of Parameters at top	Top3	Top5	Top 8
Permutation importance	2	4	6
Parameter importance	0	3	7

We do see that a greater number of parameters in the tuned models from dimensionally reduced permutation importance rank set map to the top parameters of the top ranked permutation importance set before dimensionally reducing.

The performance these models is shown below in the table16 .:

Table 16 Accuracy metrics of tuned model performances

Model	**Base Model**	**Dimensionally reduced models**	
		Parameter Imp	**Permutation Imp**
Ridge	47.0%	50.0%	54.0%
XGBoost	85.0%	86.0%	86.0%
SVM	47.0%	50.0%	52.0%

Even the performance of the models developed from permutation importance-based reduction performed slightly better than the models tuned on all parameters-based model or the ranked parameter importance set.

6.0 Comparing and concluding

Our study indicates that dimensionally reduced models will provide equal or better performance compared to models built using all parameters after the essential cleansing of datasets that is normally carried out before building predictive models. The parameters importance needs to be obtained from multiple algorithms and then a ranking algorithm works well to rank the parameter importance across the importance assigned to each parameter. There is strong case that the permutation importance attaches the right and consistent importance to parameters.

References:

[1] Den Hartog, D. N., Koopman, P. L., Thierry, H., & Kompier, M. A. J. (1996). The relationship between organizational culture and turnover intentions. Journal of Applied Psychology, 81(6), 889.

[2] Den Hartog, D. N., Koopman, P. L., Thierry, H., & Boer, M. (1999). The influence of organizational culture on turnover intentions: The mediating role of satisfaction with the person-organization fit. Journal of Applied Psychology, 84(4), 627.

[3] Wang, X., Li, Y., & Wang, X. (2015). A review of employee turnover and retention research: A Chinese perspective. Journal of Chinese human resource management, 6(1), 6-26.

[4] Alpaydin, E. (2010). Introduction to machine learning. Cambridge, MA: MIT Press.

[5] Murphy, K. P. (2012). Machine learning: a probabilistic perspective. Cambridge, MA: MIT Press.

[6] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: an introduction. Cambridge, MA: MIT Press.

[7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. Cambridge, MA: MIT Press.

[8] .Kolde, Raivo, et al. "Robust rank aggregation for gene list integration and meta-analysis." Bioinformatics 28.4 (2012): 573-580.

[9] Sangli M., Ravishankar A. (2020). “Rank Aggregation Approach to Feature Selection for Improved Model Performance.” In: Salagame R., Ramu P., Narayanaswamy I., Saxena D. (eds) Advances in Multidisciplinary Analysis and Optimization. Lecture Notes in Mechanical Engineering. Springer, Singapore. https://doi.org/10.1007/978-981-15-5432-2_27

[10] Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. "Regularization paths for generalized linear models via coordinate descent." Journal of statistical software 33.1 (2010): 1.

[11] Hoerl, Arthur E., and Robert W. Kennard. "Ridge regression: Biased estimation for nonorthogonal problems." Technometrics 12.1 (1970): 55-67.

[12] Breiman, Leo, et al. Book "Classification and regression trees. Belmont, CA: Wadsworth." International Group (1984): 432.

[13] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." Machine learning 63.1 (2006): 3-42.

[14] Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22.

[15] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, (2016).