Journal of Mines, Metals and Fuels, 71(6): 857-866; 2023. DOI: 10.18311/jmmf/2023/34502

Print ISSN : 0022-2755

Journal of Mines, Metals and Fuels

Contents available at: www.informaticsjournals.com/index.php/jmmf

An Approach for Sub Selecting Variables that have Higher Influence on the Outcome in Developing Predictive Model using Staff Turnover

Mohan Sangli[1*], Rajeshwar S Kadadevaramath[2], Jerin Joseph[3], Akarsha Kadadevaramath[4] and Immanuel Edinbarough[5]

[1,3]Research Scholars, Industrial Engineering Department, Siddaganga Institute of Technology Tumkur, India.
[*]E-mail: mrsangli@yahoo.com
[2]Professor and Head, Industrial Engineering Department, Siddaganga Institute of Technology Tumkur, India
[4]Engineer, Intel India Pvt. Ltd., Bangalore, Karnataka, India
[5]Professor and Head, Engg. Tech, Industrial and Manufacturing EnggDept, The University of Texas, USA

Abstract

Predictive models are built by learning the combined effects of several independent variables that directly or indirectly influence the outcome. H. Response or dependent variable. In practice, data collection has data on a large number of independent variables that are outcome-sensitive and may or may not be related to the outcome. Some independent variables have a large impact on the results, while others may have little or no impact on the results. The presence of some independent variables that are irrelevant to the outcome can affect the performance of the predictive model. In this context, it is desirable and essential to identify the independent variables that most influence the forecast model to keep it lean and efficient. In this work, we used a dataset containing employee turnover rates and explored how to identify a subset of outcome-sensitive variables, thus eliminating variables that hinder the development of effective predictive models. By partially selectively influencing the independent variables, we developed lean and efficient predictive models that enabled us to act on an actionable subset of the variables to reduce staff turnover, thereby improving corporate save effort and cost.

Keywords: *Predictive model, Sensitive parameter, Dimensionality*

1.0 Introduction

Predicting the outcome of actions pursued by an organization or project is a desirable scenario for all stakeholders and applicable to nearly all professions. When it comes to the realm of business and engineering, there are a myriad of parameters and constraints that affect results. While it may appear that we can put in all the parameters and constraints and set up a formula to derive the result, in practice this is not so easy due to the environment, the parties involved, and the variability observed in many operating parameters. It's not easy. Such. It is almost impossible to know exactly the relationship between variables and their results. It is not possible to track all possible parameters and measure their impact on results. I was always looking for which parameters to focus on out of the huge number of parameters. Business knowledge, experience, and intuition are often the means to predict output. More recently, companies have begun using machine learning and statistical techniques to identify sensitive parameters that have a greater impact on the target outcome being studied.

Machine learning is a rapidly growing field that promises to help you model your business, operations or simulations based on historical parameter values and corresponding

[*]Author for correspondence

results. Supervised machine learning attempts to find all relationships between parameters and outcomes to build a model that mimics the system. This is an improvement over guesswork and provides a means of fact-based assessment, but identifying some of these parameters is a more complex science. Different weights are assigned to each parameter because each machine learning algorithm learns data and builds a model based on the target metric differently. It is important to compare and contrast statements of weights assigned to parameters by a particular algorithm refined by iteration. Obtaining stable and reliable weights is therefore critical for the reliability of dimensionality reduction models, which is the subject of this work.

Machine learning is the technology by which computers learn how to learn data, relationships within and with other functions, and solve problems without being explicitly programmed.The dataset is divided into a training set and a test set. The model is trained on training data. This is called visible data. Trained models are tested on test data or unseen data during training. Accuracy and other metrics are typically gleaned from model performance on test data. Models are now used in many fields for simple needs such as predicting customer churn, or to use models in the design process as a cheaper and faster alternative to computationally expensive response estimation it has been.

The goal of machine learning modelling is to construct an approximation $bf(x)$ of a function $f(x)$ given a set of n observations $f(x(1); y(1)). (x(2); y(2)); : : : ; (x(n); y(n))g$. where x represents the p-dimensional input vector, the ith sample point is $x(i) = (xi1; xi2; : : : ; xip)T$, and $y(i)$ is the A realization of the function f. Approximations $bf(x)$ are obtained by applying various modelling algorithms. Machine learning modelling consists of three phases: (i) data acquisition, (ii) model training and optimization, and (iii) model validation. Data collection involves a sampling procedure to select sample points $x(i)$ and running a computational model to obtain the response or function evaluation $y(i)$ at the selected sample points or the observed y. It involves using (i) for the appropriate input of sample points. The captured dataset is split into training, testing, and validation datasets, which are used in different phases of model building. The training and test datasets are used not only to obtain the feature importance of the learned model (model training), but also to identify the hyper parameters that control the learning process and surrogate complexity (model tuning). will be used. A validation dataset is used to assess the quality of the final model. In practice, the validation dataset is small compared to the training and test datasets.

Model accuracy and its evaluation, and associated computation time, depend on data allocation during model training and tuning. Two data mapping methods are typically used for training and validation: (i) simple validation methods and (ii) cross-validation methods.

2.0 Literature Review

In the field of machine learning, parameters are commonly called features. Feature selection is an important part of machine learning. Today, more than ever, the focus is on reducing the complexity and time required to process data. A straightforward and easy way to improve performance is to identify and remove the noise parameters, leaving only the features that contribute positively to the model's performance.

The presence of regressive, redundant, or noisy features can significantly affect the performance and interpretability of models built on data containing such features (Guyon et al. 2003). Feature filtering, or dimensionality reduction, is a common method in machine learning. Naturally, the fewer features, the more likely the model will be trained faster. The fewer noise-inducing features in your data set, the less likely error in model performance. The industry has routinely used algorithms, linear or tree-based or kernel-based logic, to identify features and build models. Algorithms to extract variable importance (features) such as SVM, CatBoost, Random Forest, XGBoost. These have proven to be very successful[17,6]. Some practices involve the use of coefficients in linear models (logistic regression, ridge)[19].

In life sciences, the most commonly used methods to quantify feature importance are linear models and decision trees. Linear SVM and linear logistic regression are well-studied theoretical models that can provide interpretable classification rules via model parameters[2]. The overarching problem with all these methods is that there is no one-stop shop for all types of records. Some perform better with more horizontally aligned datasets, while others perform better with vertically aligned datasets. Even within a category of records, there are many factors such as: B. Linearity of parameters that affect the unpredictable nature of the performance of these models. A more intuitive solution to this problem is to use multiple methods to find the importance of each feature and find the average importance of all methods, or aggregate them using statistical methods[10,1,15]. Rank aggregation is required when using multiple methods. This process consists of two steps (Sangli M et al. 2020). The first is choosing a ranking method and the second is aggregating the ranks provided by those ranking methods. It's also worth noting that most ranking methods rely on measures of variance in some way, so there's likely to be some degree of consensus. Various researchers have published articles on the coefficient, Gini contamination or variance. We are dealing with reduced scores. Decision tree (regressor)[4], extra tree (regressor)[11]. random forest (regressor)[16] and XGBoost (Extreme Gradient Boosting)[6]. These are n preferred lists of aggregates that should be searched. To find rank aggregations, use[15]. A proposed robust rank aggregation method[15].

Importance may not be reliable unless each model is tuned over multiple iterations and feature (weight) importance is preserved. Optimizing all models and then using ranking methods requires significant time and computational power as dimensionality increases, defeating the purpose of reducing dimensionality early in the model development process.

3.0 Research Problem

The accuracy of predictive models is compromised by the presence of variables that are unaffected by the dependent variable, known as noise in the data. The model building phase is more CPU and time consuming due to the large number of variables and is called the dimensional bane. Different models used to reduce dimensionality may have slightly different sets of independent variables used to build the models, leaving some of the otherwise important independent variables. How to address such concerns should be considered to ensure that all important variables are used during construction of the final machine learning model.

3.1 Research Objectives

This work aims to achieve the following goals.

- Ensures that the correct independent variables are selected for modelling when significant variables are selected for modelling up to a certain percentage, reducing the curse of dimensionality.
- The accuracy of the final model selected is higher for the same number of independent variables are used while building models.

3.2 Approach

We will go through the following steps:

- Pre-requisite concepts and cover algorithm, model bases and selecting few models.
- Feature elimination through feature characteristics.
- Feature elimination through feature importance from multiple models and ranking method.
- Selecting few models and tuning for RMSE or accuracy using top 90% importance contributing features for each feature importance.
- Deriving new feature importance from tuned models and ranking; developed models using Ridge, XGBoost, LightGBM and Linear SVM.
- Comparing and concluding.

An algorithm is a step-by-step procedure for solving a problem or accomplishing some end (MW dictionary 2021). Typically, an algorithm contains series of instructions knit together to achieve one or more objectives of transforming inputs to an outcome. An algorithm, for example may have steps to analyse and filter data, assign weights and predict outcome from inputs and assigned weightages in the process.

3.3 Model Types

Linear, tree-based, kernel, and deep learning are commonly used as the basis for algorithms. Linear is easiest and fastest with many variables and less noise, but is more powerful.

3.3.1 Linear Regression

The regression process assigns a weight parameter theta to each of the training traits. The predicted output (h(θ)) is a linear function of the features and the θ coefficients as given in equation (1).

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots \quad \ldots (1)$$

All theta are randomly initialized at the start of training. As training progresses, each theta corresponding to each feature is modified in a way that minimizes the deviation between expected and predicted outputs (also called loss minimization). Align the θ values in the correct direction using a gradient descent algorithm.Two features are said to be collinear if one feature can be linearly predicted from the other with reasonable accuracy. In such cases, one of the features is often removed before training the model.

Other linear algorithms include logistic regression (mainly used for classification problems).

3.3.2 Tree-based algorithms. Regression

Trees are used for continuous dependent variables and classification trees are used for discrete dependent variables. The node is a condition that determines which node to move to next. The chain predicts an output when it reaches a leaf node. Entropy/information gain is used as a criterion for selecting node conditions. A recursive greedy-based algorithm is used to derive the tree structure. The Gini index is a commonly used classification metric to calculate how well data points are blended together. Selects the attribute with the higher Gini index as the next condition (Eq 2). Other tree-based algorithms include random forests, collections of decision trees, and more robust and generalized solutions that reduce overfitting.

$$gininnindex = 1 - \Sigma P_t^2 \quad \ldots (2)$$

3.3.3 Kernel-based Algorithm

The kernel allows dot products to be computed in otherwise difficult-to-compute regions. Linear algorithms that only use inner products can be implicitly performed using kernels. H. You can construct nonlinear versions of linear algorithms very elegantly. Commonly used kernel functions include Gaussian RBF, polynomial, sigmoid and spline kernels. (websus-Robert Müller, 2001). Kernel-based

unsupervised learning involves kernel PCA, a nonlinear extension of PCA for finding projections that provide useful nonlinear descriptors of data. Also included is his SVM algorithm for a single class. The problem of outlier detection in high dimensions.

Support Vector Machines are a popular algorithm in this area and are often used when they perform better than LR due to the large number of parameters.

3.4 Features

In a supervised machine learning input dataset, features are the input attributes used for prediction or classification. Functions are easy to understand. Temperature, house size, location type, salary range and dates are some example characteristics. Interpretability of features is a big assumption. But if it's hard to understand the input function, it's even harder to understand what the model is doing. The input set is a matrix X of size x(ij). Where i is an instance. i.e., rows and j are features.

A goal is information that a machine learns and predicts. In formulas, the target is usually called y and the set is Y. y(i), where i is any instance or row.

3.5 Features and their role in modelling

In the case of real business or technical problems, records contain both categorical and numerical data. Most models work on numeric data, and algorithmic techniques require an encoding method to convert categorical data to numeric. The type of encoding employed affects model performance. Some algorithms, such as tree-based machine learning algorithms, are equipped to process categorical and numeric input data without coding requirements. Regularization methods such as Lasso and Elastic Net can help filter out variables of low importance. A commonly used random forest can also be used to rank inputs, and this information can be used to improve model accuracy.

3.6 Feature Removal by Statistical Methods

Even before arriving at feature importance, some of the features are filtered by statistical methods. Some feature filter approaches:

- Missing Values: If more than x% of features used as training data sets for machine learning are missing, it is recommended to remove the features. For example, below 10%, you can usually impute values in different ways depending on whether the value is numeric or categorical. You can also simply impute values by local predictive models.
- Low Feature Variance: Nearly constant values of features are not very important for target prediction and are discarded.
- Highly Correlated: Pairs of variables that are highly correlated increase the multicollinearity of the dataset. Importance is shared when there are correlated features. You can keep the most important ones related to the function and omit the rest by sequential filtering.
- High Cardinality: Cardinality is the number of distinct values in a variable. Even if the machine learning algorithm tolerates categorical variables, high cardinality can be taxing in terms of computational resources. Postal codes are a good example of a categorical variable with very high cardinality.
- ISOMAP: This technique is useful for feature filtering when the data is highly nonlinear.
- t-SNE: This is also used to filter features when the data is highly nonlinear. Also very suitable for visualization.
- UMAP: A better method than t-SNE for high dimensional data.

Therefore, the execution time is shorter than t-SNE. Now that we have excluded features from our dataset using the various methods described above, the next step is to assign importance to the features using the subject machine learning algorithms. Remove noise, reduce dimensionality, and improve machine learning performance in terms of accuracy and speed. In the experimental section, we filter features with missing values, low variance, high correlation, and high cardinality.

3.7 Feature Importance and Dimension Reduction

Feature importance is a highly condensed global insight into model behaviour that can be derived from the importance that algorithms attach to features during modelling. (Christoph Molnar, 2021).

Function importance allows one to,

- Train machine learning algorithms faster.
- Reducing model complexity and making it easier to interpret
- Improve model accuracy by selecting appropriate subsets.
- Reduce over fitting.

Few other widely used methods, such as principal component analysis (PCA) and independent component analysis (ICA), split the data into several components to better explain the variance and All used functions are replaced. in the resulting function. I do not recommend using this, especially if you want to explain your predictions. For the same reason, we refrain from factor analysis when sets of variables are strongly correlated. This divides the variables into different groups based on their correlation and presents each group with a factor that removes the original feature. Omits some of the commonly used but controversial machine learning.

3.8 Backward feature removal and forward feature selection methods

This is because they eliminate features on a feature-by-feature basis and eliminate the impact of omitted features on prediction accuracy, which takes a lot of computational resources and time. It is important. It works well for small datasets. Recently, a three-class feature selection method was proposed[3].

3.9 Filter methods

Rank features by calculating a score for each feature independent of a model. For many filter methods, the score calculation can be done in parallel. Feature selection is independent of any machine learning algorithms. Instead, features are selected based on their scores from various statistical tests.

3.10 Wrapper methods

Consider subsets of the set of all features. The subsets are evaluated by a performance measure calculated on the resulting model (e.g., classification accuracy). Wrapper methods include simple approaches like greedy sequential searches, but also more elaborate algorithms like recursive feature elimination as well as evolutionary and swarm intelligence algorithms for feature selection. In Wrapper methods, we decide to add or remove features from your subset based on the inferences that we draw from the model. The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

3.11 Embedded Method

Include feature selection in the model fitting process. Examples of prediction methods that perform embedded feature selection are lasso regression, tree-based methods such as classification and regression trees, or random forests and gradient boosting.

This study uses the embedding method.

In this study, we plan to use the following models for the initial importance of features and then use a ranking scheme to rank features based on the importance given by the various models.

Linear Models: Elastic Net, Ridge, Lasso for Regression. Logistic Regression and Ridge

Tree Models for Classification: Decision Tree, Random Forest, XGBoost, CatBoost, LightGBM

3.12 Kernel Models: Linear SVM

Linear Regression: with coefficients that minimize the residual sum of squares between observed targets Fit a linear model. Dataset and target predicted by linear fitting. The coefficients of this model are used as effects. Indicate a linear relationship, if any, between feature-response pairs.

- Ridge: This model addresses some of the linear regression problems by penalizing the size of the coefficients. The complexity parameter alpha (a > 0) controls the amount of shrinkage. The higher the value of alpha, the greater the amount of shrinkage, and the more robust the coefficients are to collinearity.
- Lasso: This model is a linear model that estimates sparse coefficients. It is useful in some contexts as it tends to favor solutions with fewer non-zero coefficients, effectively reducing the number of features that a particular solution depends on.
- Elastic Net: An Elastic Net is a linear regression model trained with both l1 and l2 norm regularization of the coefficients. This combination allows you to learn sparse models with few nonzero weights, such as Lasso, while preserving the ridge regularization property.
- Logistic Regression (Classification): Logistic Ridge Regression is logistic regression combined with a ridge penalty (Izenman, 2013) . The ridge parameter ë balances the goodness of fit (log-likelihood) and the size of the regression parameters. For ë = 0, this is ordinary logistic regression without ridge penalties. ë = 0 is not feasible if the dataset contains more features than instances, or if the feature space has a hyperplane that completely separates the two classes. Large values of ë shrink all regression coefficients toward 0.
- Decision Trees: Decision trees are constructed by greedily searching the given data top-down, testing each attribute of each node. The importance of features in decision trees is based on information gain, a mathematical method of obtaining the amount of information by choosing certain attributes.
- Random Forest: This is one of the most commonly used techniques to indicate the importance of each feature present in a data set. The dimensionality is reduced because you can determine the importance of each feature and keep the top features. Random Forest creates multiple decision trees and merges them to make more accurate and robust predictions.
- XGBoost: Gradient boosting is an approach that creates a new model that predicts the residuals or errors of previous models and sums them to give the final prediction. Build multiple decision trees using pre-sorted and histogram-based algorithms to compute optimal splits.
- Cat Boost: Provides a new technique called Minimum Variance Sampling (MVS). This is a weighted sampling version of stochastic gradient boosting. In this technique, weighted sampling is done at the tree level instead of the split level. Each boosting tree observation is sampled to maximize the accuracy of the split scoring.

- Light GBM: Uses a new technique of gradient-based one-sided sampling (GOSS) to filter out data instances and find split values to build trees. GOSS strikes a balance between improving speed by reducing the number of data instances and maintaining the accuracy of learned decision trees.
- Linear SVM: Support Vector Machines use hyperplanes in the feature space as decision boundaries. This is optimal with respect to the maximum margin principle. A kernel function is used to change the shape of the hyperplane to a nonlinear one (Izenman, 2013, pp. 369ff.). Use support vector machines with RBF kernels. It has two hyperparameters. The regularization parameter C and the core width parameter ó. For linear svm we use a linear kernel where the model tries to predict the best line within a threshold.

3.13 Feature Importance

For linear and kernel-based models, obtain the feature importance from the coefficients of the model for each variable. For tree-based models, the importance of a single decision tree is calculated based on how much each attribute split point improves the performance index, weighted by the number of observations a node is responsible for. A measure of performance is the purity (Gini index) used to select the split point, or some other more specific error function. Feature importance is averaged over all decision trees in the model.feature selection: feature selection typically uses a combination of filter and wrapper methods. It can be implemented using the XG Boost package which has its own built-in function selection method.

4.0 Experiment

Datasets Considered

1. Personnel Turnover Dataset: Uncover the factors that lead to employee turnover and answer questions such as “Please show me a breakdown of distance from home by job and turnover” or “Monthly average Examine key questions such as “compare income from education and wear and tear”. This is a fictitious dataset created by an IBM data scientist.
2. Employee turnover: This employee turnover data set is the actual data set from Edward Babushkin’s blog and is used to predict employee layoff risk (survival analysis model use). Edward Babushkin says: Unlike other transport services such as buses and subways, these systems explicitly record travel times, departures, and arrivals, so bike-sharing systems can become virtual sensor networks that can be used to record city movements. Become. Therefore, we expect to be able to detect most of the important events in the city by monitoring this data.

4.1 Record Characteristics and Results Summary

Table 1, detailing the number of samples, attributes, filtered attributes, and cumulative importance (FI) of characteristics by rank.we can observe from the table that, About 60% attributes will suffice to fairly predict the turnover (i.e the attributes “event” and “attrition” in two data sets).

Once the importance of the attributes are ranked and as you start adding attributes in that order, the first few will have a good impact of response prediction and the incremental importance reduces after that.Thus we can deduce that the ranking of importance helps us to select the important attributes while reducing the dimensionality and hence could providebetter dimensionality reduced models.

4.2 Model performance

We used the features with importance cumulating to up to 90% after ranking from the feature importance output of 8 models and tuned Ridge, XGBoost, LightGBM and Linear SVM models with default parameters over 25 iterations. The model was matched using the same four algorithms and 25 iterations with the same default parameters for all functions. Table 2 are the results regarding the accuracy of the classification model.

As we can observe, if we need to consider all attributes to choose the best model, or only 60% of the attributes with 90% importance, the accuracy of the model is lower for both cases in the HRdata dataset. is the same, with only a 3% decrease. sales record. We found that the same variable can have different importance using different methods. In actual

Table 1 : Number of samples, attributes, filtered attributes, and cumulative importance

Dataset	Target	sample size	Number of attributes	Dropped attributes	Cumulative importance FI towards response (Attrition)		
					<50%	<70%	<90%
Turnover	event	1129	15	1	2	5	9
HR data	attrition	1470	35	4	9	16	23

model development, multiple iterations are used, so the weights for each parameter are adjusted across iterations. This gives you a different weight than the first quick insight to figure out which model is better. New rankings after model optimization are skewed on some datasets as Light GBM's performance focuses on only a few attributes. Apart from that, the feature importance before model optimization and the feature importance after model development are

Table 2: Results regarding the accuracy of the classification model

Model (Accuracy)	Turnover		HR data	
	All features	Top 90% features	All features	Top 90% features
Ridge	48%	53%	55%	47%
XGBoost	57%	49%	84%	84%
LightGBM	53%	50%	84%	84%
SVM	47%	52%	52%	46%

Table 3 : Feature importance

Rank	Feature	LR	Ridge	Decision tree	Random forest	XG Boost	Light GBM	Cat Boost	SVM	average_ scores	cumsum
1	Industry	29.51	33.07	16.82	15.41	27.24	11.98	13.88	32.34	22.531	22.531
2	Traffic	12.82	14.1	10.9	10.34	14.25	9.43	10.87	13.89	12.075	34.606
3	Profession	32.19	25.32	5.74	7.53	32.51	5.38	7.26	26.5	17.804	52.41
4	Age	3.87	4.17	14.81	11.77	1.37	16.45	13.81	4.09	8.793	61.203
5	Way	8.25	8.59	3.75	5.57	7.1	4.11	9.61	8.56	6.943	68.145
6	Coach	3.89	4.31	3.58	5.44	7.1	4.09	5.37	4.26	4.755	72.9
7	Novator	0.7	0.76	10.74	7.65	1.35	8.67	6.4	0.69	4.62	77.52
8	Selfcontrol	0.75	0.74	5.07	7.99	1.28	9.23	7.09	0.9	4.131	81.651
9	Anxiety	1.72	1.86	11.63	7.48	1.26	8.67	6.96	1.91	5.186	86.838
10	Head-Gender	3.78	4.27	1.85	3.48	2.45	2.6	4.67	4.16	3.408	90.245
11	In depend	0.39	0.48	7.97	7.52	1.42	8.55	6.96	0.42	4.214	94.459
12	Gender	1.82	2	1.65	2.56	1.51	1.37	1.61	1.99	1.814	96.273
13	Extraversion	0.3	0.32	5.49	7.27	1.15	9.48	5.53	0.29	3.729	100.001

Table 4 : Feature importance based Dimensionality reduced tuned model

Old rank	Rank	Feature	Ridge	FI-Ridge	XGBoost	FI-XGB	LightGBM	FI-LGBM	SVM	FI-SVM
1	1	Industry	34.54	33.07	33.78	27.24	20.44	11.98	35.88	32.34
2	2	Traffic	16.22	14.1	15.23	14.25	15.91	9.43	11.05	13.89
5	3	Way	9.46	8.59	7.03	7.1	11.15	4.11	9.05	8.56
3	4	Profession	27.19	25.32	29.19	32.51	8.89	5.38	32.13	26.5
6	5	Coach	4.96	4.31	5.47	7.1	7.3	4.09	4.56	4.26
7	6	Novator	0.69	0.76	2.27	1.35	7.18	8.67	0.56	0.69
4	7	Age	3.83	4.17	2.75	1.37	13.86	16.45	2.94	4.09
8	8	Self control	0.03	0.74	2.16	1.28	5.91	9.23	0.18	0.9
9	9	Anxiety	3.1	1.86	2.11	1.26	9.38	8.67	3.66	1.91
			100	92.92	100	93.46	100	78.01	100	93.14

Table 5: Feature Importance

Attribute	Rank	Feature	LR	Ridge	Decision tree	Random forest	XG Boost	Light GBM	Cat Boost	SVM	average _scores	Cumsum
13	1	Job role	16.59	13.62	5.66	4.15	21.96	3.63	4.89	19.89	11.3	11.3
16	2	Monthly income	4.57	6.01	8.07	6.58	1.06	7.21	5.88	4.56	5.49	16.79
22	3	Stock option level	5.64	5.56	4.17	3.57	10.42	1.9	6.01	4.85	5.27	22.06
25	4	Work life balance	4.37	4.26	4.26	3.55	7.25	2.33	4	3.78	4.23	26.28
19	5	Over time	5.65	5.71	4.78	4.79	2.09	0.99	7.25	4.92	4.52	30.8
18	6	Num companies worked	3.88	4.03	4.76	3.14	1.73	3.43	4.95	3.58	3.69	34.49
23	7	Total working years	3.51	3.84	11.68	5.14	2.94	3.34	2.66	3.28	4.55	39.04
12	8	Job level	8.92	11.64	2.02	3.01	3.01	1.05	2.62	9.47	5.22	44.26
8	9	Environment satisfaction	4.65	4.67	3.25	3.18	4.66	2.95	4.29	4.12	3.97	48.23
0	10	Age	1.41	2.21	7.04	5.33	1.25	6.31	6.09	1.24	3.86	52.09
21	11	Relationship satisfaction	4.12	3.54	0.66	3.05	9.56	2.87	4.78	3.5	4.01	56.1
11	12	Job involvement	3.49	4.49	1.39	2.82	4.96	1.85	2.4	3.32	3.09	59.19
29	13	Years with curr manager	2.76	1.72	0.67	3.48	1.14	2.69	3.88	2.61	2.37	61.56
26	14	Years at company	1.51	1.74	2.78	3.44	0.94	2.87	2.22	1.8	2.16	63.72
14	15	Job satisfaction	3.68	3.7	0.96	2.84	5.36	2.54	3.82	3.22	3.27	66.99
4	16	Distance from home	2.74	1.62	3.68	3.47	1.1	4.98	3.17	2.29	2.88	69.87
3	17	Department	4.14	3.42	3.53	1.76	2.03	0.88	2.78	7.3	3.23	73.1
6	18	Education field	3.46	4.8	1.98	3.28	3.82	3.22	1.56	3.44	3.2	76.29
17	19	Monthly rate	0.64	0.57	3.83	3.84	0.5	7.56	2.39	0.72	2.51	78.8
27	20	Years in current role	1.55	1.7	2.89	2.6	1.3	1.66	1.71	1.56	1.87	80.67
28	21	Years since last promotion	2.34	1.53	0.85	2.12	0.91	2.36	2.23	2	1.79	82.46
1	22	Business travel	4.54	3.78	0.43	1.98	4.03	0.98	2.51	3.78	2.75	85.22
2	23	Daily rate	1.25	0.76	4.62	4.28	0.76	7.54	3.25	1.06	2.94	88.16
10	24	Hourly rate	0.3	0.67	5.01	3.81	0.64	6.22	3.41	0.26	2.54	90.7
5	25	Education	0.84	1.53	1.11	2.56	1.85	2.59	1.67	0.55	1.59	92.28
24	26	Training times last year	1.33	1.05	1	1.95	1.12	2.36	1.8	1.2	1.48	93.76
7	27	Employee number	0.04	0.47	2.11	3.55	0.48	7.32	2.36	0.05	2.05	95.81
15	28	Marital status	0.59	0.44	1.03	2.71	1.7	1.63	1.88	0.68	1.33	97.14
20	29	Percent salary hike	0.4	0.19	3.8	2.71	0.65	3.91	2.75	0.09	1.81	98.95
9	30	Gender	1.08	0.75	1.98	1.31	0.79	0.83	0.8	0.87	1.05	100

comparable within the expected deviation of the top features.

Ranked models and individual model feature importance, permutation importance and matched models, feature re-evaluation from both feature importance and permutation importance for each dataset are shown in the following Tables 3 to 6.

5.0 Conclusions

Some features of each datasets were filtered by statistical methods. The feature importance was then determined from the eight models and the features were ranked using a ranking algorithm. We used the cumulative average

Table 6: Feature importance based Dimensionality reduced tuned model

old Rank	Rank	Feature	Ridge	XGBoost	LightGBM	SVM
1	1	Job role	14.54	11.61	4.62	10.56
5	2	Over time	5.97	7.87	4.05	6.24
3	3	Stock option level	6.09	7.68	3.29	5.36
8	4	Job level	12.64	7.07	3.28	14.79
7	5	Total working years	4.1	3.33	4.56	4.28
4	6	Work life balance	4.52	6.49	3.64	5.25
11	7	Relationship satisfaction	3.7	5.5	3.87	3.33
15	8	Job satisfaction	3.65	5.54	3.41	4.31
2	9	Monthly income	6.33	2.41	8.8	3.16
9	10	Environment satisfaction	4.97	6.37	4.01	4.28
18	11	Education field	4.93	6.12	2.23	8.92
13	12	Years with curr manager	1.8	2.63	3.72	1.52
12	13	Job involvement	4.56	5.9	2.85	4.35
6	14	Num companies worked	4.28	2.08	4.18	3.48
23	15	Daily rate	0.86	1.3	7.98	1.87
21	16	Years since last promotion	1.6	1.34	2.93	1.69
17	17	Department	3.4	4.26	2.85	4.08
10	18	Age	2.25	2.15	7.44	2.21
20	19	Years in current role	1.81	1.38	2.76	2.19
22	20	Business travel	3.83	4.09	1.13	5.69
14	21	Years at company	1.89	2.34	4.37	1.05
19	22	Monthly rate	0.55	1.16	7.28	0.68
16	23	Distance from home	1.75	1.38	6.74	0.72

importance and selected parameters that cumulatively contribute up to 90% importance. From the ordered features, we developed four different models representing linear, tree, and kernel-based models and again obtained the feature importance from the matched models. We observed that models developed using features with reduced dimensions perform better with comparable accuracy when using smaller dimensions. Opportunities to observe the same across species and sizes or datasets to further validate results.

6.0 References

1. Aerts, Stein, et al. (2006): "Gene prioritization through genomic data fusion." Nature biotechnology 24.5: 537.
2. André Altmann", †, Laura Tolo ¸si",†, Oliver Sander‡ and Thomas Lengauer. (2010): "Permutation importance: a corrected feature importance measure" Vol.26 no.10, pages 1340–1347 doi:10.1093/bioinformatics/btq134.
3. Andrea Bommert, Xudong Sun, Bernd Bischl, JörgRahnenführer, Michel Lang (2020): "Benchmark for filter methods forfeature selection in high-dimensional classification data. Computational Statistics & Data Analysis Volume 143, March 2020, 106839.
4. Breiman, Leo, et al. (1984): Book "Classification and regression trees. Belmont, CA: Wadsworth." International Group: 432.
5. Chehata, Nesrine, Li Guo, and Clément Mallet. (2009): "Airborne lidar feature selection for urban classification using random forests." International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 38. Part 3: W8.
6. Chen, Tianqi, and Carlos Guestrin. (2016): "Xgboost: A scalable tree boosting system." Proceedings of the

22nd acmsigkdd international conference on knowledge discovery and data mining. ACM.
7. Definition of Algorithm. https://www.merriam-webster.com/dictionary/algorithm.
8. Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7.1 (2006): 3. (2021).
9. Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. (2010): "Regularization paths for generalized linear models via coordinate descent." *Journal of statistical software* 33.1: 1.
10. Griffith, Obi L. Melck, Adrienne, Steven JM Wiseman, Sam M. Jones, and S. M. Wiseman. (2006): "Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers." *Journal of Clinical Oncology* 24.31: 5043-5051.
11. Geurts, Pierre, Damien Ernst, and Louis Wehenkel. (2006): "Extremely randomized trees." *Machine learning* 63.1: 3-42.
12. Guyon, Isabelle, and André Elisseeff. (2003): "An introduction to variable and feature selection." *Journal of machine learning research* 3. Mar (2003): 1157-1182.
13. Hans, Chris.(2009): "Bayesian lasso regression." Biometrika 96.4 : 835-845.
14. Hoerl, Arthur E., and Robert W. Kennard. (1970): "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics* 12.1: 55-67.
15. Kolde, Raivo, et al. (2012): "Robust rank aggregation for gene list integration and meta-analysis." *Bioinformatics* 28.4: 573-580.
16. Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2.3 (2002): 18-22. Predrag Radivojac1, Zoran Obradovic2, A. Keith Dunker1, and Slobodan Vucetic2; J.-F. Boulicaut et al "Feature Selection Filters Based on the Permutation Test". (Eds.): ECML 2004, LNAI 3201, pp. 334–346, 2004. © Springer-Verlag Berlin Heidelberg 2004
17. Menze, Bjoern H., et al. (2009): "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data." *BMC bioinformatics* 10.1: 213.
18. Molnar, Christoph. 2019: "Interpretable machine learning. A Guide for Making Black Box Models Explainable", https://christophm.github.io/interpretable-ml-book/.
19. Xing, Eric P., Michael I. Jordan, and Richard M. Karp. (2001): "Feature selection for high-dimensional genomic microarray data." *ICML*. Vol.1.