



**WNS Analytics Wizard 2018**  
 Are you ready for the War of the Wizards?  
**Sept 14-16<sup>th</sup> 2018**  
 Participate to win cash prizes upto INR 3Lacs



[Home](#) > [Machine Learning](#) > [8 Proven Ways for improving the "Accuracy" of a Machine Learning Model](#)

MACHINE LEARNING

# 8 Proven Ways for improving the "Accuracy" of a Machine Learning Model

SUNIL RAY, DECEMBER 29, 2015

## Introduction

Enhancing a model performance can be challenging at times. I'm sure, a lot of you would agree with me if you've found yourself stuck in a similar situation. You try all the strategies and algorithms that you've learnt. Yet, you fail at improving the accuracy of your model. You feel helpless and stuck. And, this is where 90% of the data scientists give up.

But, this is where the real story begins! This is what differentiates an average data scientist from a master data scientist. Do you also dream of becoming a master data scientist ?

If yes, you need these 8 proven ways to re-structure your model approach. A predictive model can be built in many ways. There is no 'must-follow' rule. But, if you follow my ways (shared below), you'd surely achieve high accuracy in your models (given that the data provided is sufficient to make predictions).

I've learnt these methods with experience. I've always preferred to learn practically than digging theories. And, my approach has always encouraged me. In this article, I've shared the 8 proven ways using which you can create a robust machine learning model. I hope my knowledge can help people in achieving great heights in their careers.



## 8 Methods to Boost the Accuracy of a Model

The model development cycle goes through various stages, starting from data collection to model building.

But, before exploring the data to understand relationships (in variables), It's always recommended to perform **hypothesis generation**. (To know more about hypothesis generation, refer to [this link](#)). I believe this is the most under – rated step of predictive modeling.

It is important that you spend time thinking on the given problem and gaining the domain knowledge. So, how does it help?






This practice usually helps in building better features later on, which are not biased by the data available in the data-set. This is a crucial step which usually improves a model's accuracy.

At this stage, you are expected to apply structured thinking to the problem i.e. a thinking process which takes into consideration all the possible aspects of a particular problem.

Let's dig deeper now. Now we'll check out the proven way to improve the accuracy of a model:

### 1. Add more data

### TOP ANALYTICS VIDHYA USERS

Rank	Name	Points
1	 Rohan Rao	8856
2	 SRK	8817
3	 aayushmnit	7739
4	 mark12	6798
5	 sonny	5947

[More Rankings](#)



Having more data is always a good idea. It allows the “data to tell for itself,” instead of relying on assumptions and weak correlations. Presence of more data results in better and accurate models.

I understand, we don't get an option to add more data. For example: we do not get a choice to increase the size of training data in data science competitions. But while working on a company project, I suggest you to ask for more data, if possible. This will reduce your pain of working on limited data sets.

## 2. Treat missing and Outlier values

The unwanted presence of missing and outlier values in the training data often reduces the accuracy of a model or leads to a biased model. It leads to inaccurate predictions. This is because we don't analyse the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.

Look at the below snapshot carefully. It shows that, in presence of missing values, the chances of playing cricket by females is similar as males. But, if you look at the second table (after treatment of missing values based on salutation of name, “Miss” ), we can see that females have higher chances of playing cricket compared to males.

**With Missing Values**

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55		Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57		Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	2	1	50%
M	4	2	50%
Missing	2	2	100%

**After imputation of missing values**

Name	Weight	Gender	Play Cricket/ Not
Mr. Amit	58	M	Y
Mr. Anil	61	M	Y
Miss Swati	58	F	N
Miss Richa	55	F	Y
Mr. Steve	55	M	N
Miss Reena	64	F	Y
Miss Rashmi	57	F	Y
Mr. Kunal	57	M	N

Gender	#Students	#Play Cricket	%Play Cricket
F	4	3	75%
M	4	2	50%

Above, we saw the adverse effect of missing values on the accuracy of a model. Gladly, we have various methods to deal with missing and outlier values:

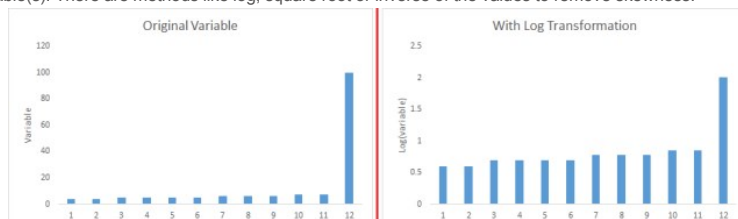
1. **Missing:** In case of continuous variables, you can impute the missing values with mean, median, mode. For categorical variables, you can treat variables as a separate class. You can also build a model to predict the missing values. KNN imputation offers a great option to deal with missing values. To know more about these methods refer article “[Methods to deal and treat missing values](#)”.
2. **Outlier:** You can delete the observations, perform transformation, binning, Imputation (Same as missing values) or you can also treat outlier values separately. You can refer article “[How to detect Outliers in your dataset and treat them?](#)” to know more about these methods.

## 3. Feature Engineering

This step helps to extract more information from existing data. New information is extracted in terms of new features. These features may have a higher ability to explain the variance in the training data. Thus, giving improved model accuracy.

Feature engineering is highly influenced by hypotheses generation. Good hypothesis result in good features. That's why, I always suggest to invest quality time in hypothesis generation. Feature engineering process can be divided into two steps:

- **Feature transformation:** There are various scenarios where feature transformation is required:
  - A) Changing the scale of a variable from original scale to scale between zero and one. This is known as data normalization. For example: If a data set has 1st variable in meter, 2nd in centi-meter and 3rd in kilo-meter, in such case, before applying any algorithm, we must normalize these variable in same scale.
  - B) Some algorithms works well with normally distributed data. Therefore, we must remove skewness of variable(s). There are methods like log, square root or inverse of the values to remove skewness.



C) Some times, creating bins of numeric data works well, since it handles the outlier values also. Numeric data can be made discrete by grouping values into bins. This is known as data discretization.

- **Feature Creation:** Deriving new variable(s ) from existing variables is known as feature creation. It helps to unleash the hidden relationship of a data set. Let's say, we want to predict the number of transactions in a store based on transaction dates. Here transaction dates may not have direct correlation with number of transaction, but if we look at the day of a week, it may have a higher correlation. In this case, the information about day of a week is hidden. We need to extract it to make the model better.

## 4. Feature Selection

### POPULAR POSTS

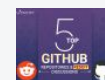
- 24 Ultimate Data Science Projects To Boost Your Knowledge and Skills (& can be accessed freely)
- A Complete Tutorial to Learn Data Science with Python from Scratch
- Essentials of Machine Learning Algorithms (with Python and R Codes)
- 7 Types of Regression Techniques you should know!
- 20 Challenging Job Interview Puzzles which every analyst should solve atleast once
- Understanding Support Vector Machine algorithm from examples (along with code)
- A comprehensive beginner's guide to create a Time Series Forecast (with Codes in Python)
- A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)

### RECENT POSTS



An End-to-End Guide to Understand the Math behind XGBoost

SEPTEMBER 6, 2018



The 5 Best Machine Learning GitHub Repositories & Reddit Threads from August 2018

SEPTEMBER 2, 2018



DataHack Radio Episode #9: Data Science at Airbnb & Lyft with Dr. Alok Gupta

SEPTEMBER 2, 2018



Build High Performance Time Series Models using Auto ARIMA in Python and R

AUGUST 30, 2018

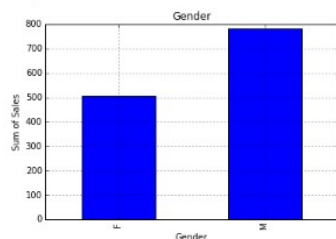


Feature Selection is a process of finding out the best subset of attributes which better explains the relationship of independent variables with target variable.



You can select the useful features based on various metrics like:

- **Domain Knowledge:** Based on domain experience, we select feature(s) which may have higher impact on target variable.
- **Visualization:** As the name suggests, it helps to visualize the relationship between variables, which makes your variable selection process easier.

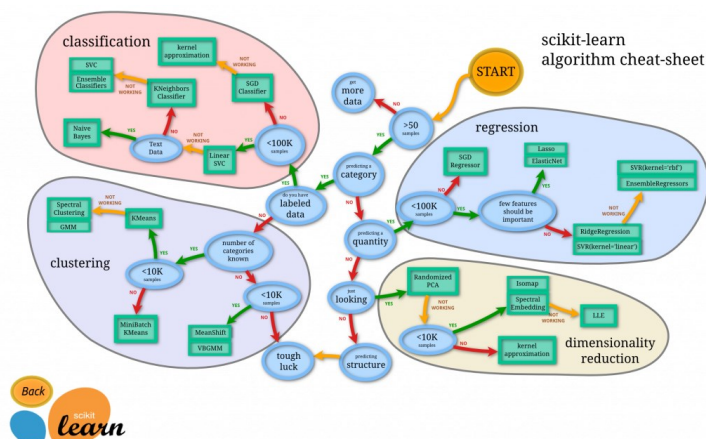


- **Statistical Parameters:** We also consider the p-values, information values and other statistical metrics to select right features.
  - **PCA:** It helps to represent training data into lower dimensional spaces, but still characterize the inherent relationships in the data. It is a type of dimensionality reduction technique. There are various methods to reduce the dimensions (features) of training data like factor analysis, low variance, higher correlation, backward/ forward feature selection and others.

## 5. Multiple algorithms

Hitting at the right machine learning algorithm is the ideal approach to achieve higher accuracy. But, it is easier said than done.

This intuition comes with experience and incessant practice. Some algorithms are better suited to a particular type of data sets than others. Hence, we should apply all relevant models and check the performance.



Source: [Scikit-Learn cheat sheet](#)

## 6. Algorithm Tuning

We know that machine learning algorithms are driven by parameters. These parameters majorly influence the outcome of learning process.

The objective of parameter tuning is to find the optimum value for each parameter to improve the accuracy of the model. To tune these parameters, you must have a good understanding of these meaning and their individual impact on model. You can repeat this process with a number of well performing models.

For example: In random forest, we have various parameters like max\_features, number\_trees, random\_state, oob\_score and others. Intuitive optimization of these parameter values will result in better and more accurate models.

You can refer article "[Tuning the parameters of your Random Forest model](#)" to learn the impact of parameter tuning in detail. Below is random forest scikit learn algorithm with list of all parameters:-



```
RandomForestClassifier(n_estimators=10, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, bootstrap=True, oob_score=False, n_jobs=1, random_state=None, verbose=0, warm_start=False, class_weight=None)
```



## 7. Ensemble methods

This is the most common approach found majorly in winning solutions of Data science competitions. This technique simply combines the result of multiple weak models and produce better results. This can be achieved through many ways:

- **Bagging** (Bootstrap Aggregating)
- **Boosting**

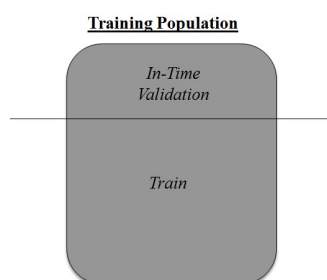
To know more about these methods, you can refer article "[Introduction to ensemble learning](#)".

It is always a better idea to apply ensemble methods to improve the accuracy of your model. There are two good reasons for this: a ) They are generally more complex than traditional methods. b) The traditional methods give you a good base level from which you can improve and draw from to create your ensembles.

## Caution!

Till here, we have seen methods which can improve the accuracy of a model. But, it is not necessary that higher accuracy models always perform better (for unseen data points). Sometimes, the improvement in model's accuracy can be due to over-fitting too.

**8. Cross Validation:** To find the right answer of this question, we must use **cross validation** technique. Cross Validation is one of the most important concepts in data modeling. It says, try to leave a sample on which you do not train the model and test the model on this sample before finalizing the model.



This method helps us to achieve more generalized relationships. To know more about this cross validation method, you should refer article "[Improve model performance using cross validation](#)".

## End Notes

The process of predictive modeling is tiresome. But, if you can think smart, you can outrun your fellow competition easily. Simply, think of these 8 steps. Once you get the data set, follow these proven ways and you'll surely get a robust machine learning model. But, these 8 steps can only help you, after you've mastered these steps individually. For example, you must know of multiple machine learning algorithms such that you can build an ensemble.

In this article, I've shared 8 proven ways which can improve the accuracy of a predictive model. These methods are widely known but not used in sequence as defined above.

Did you find this tutorial useful? If you need any more help with machine learning models, please feel free to ask your questions in the comments below.

**If you like what you just read & want to continue your analytics learning, [subscribe to our emails](#), [follow us on twitter](#) or like**



our [facebook page](#).

You can also read this article on Analytics Vidhya's Android APP



TAGS : [CROSS-VALIDATION](#), [DIMENSIONALITY REDUCTION](#), [ENSEMBLE MODEL](#), [FEATURE ENGINEERING](#), [FEATURE SELECTION](#),  
[MISSING VALUE TREATMENT](#), [OUTLIER REMOVAL](#), [PCA](#)

PREVIOUS ARTICLE

< [Year in Review: Best of Analytics Vidhya from 2015](#)

...

NEXT ARTICLE

[New Year Resolutions for a Data Scientist](#) >



### Sunil Ray

I am a Business Analytics and Intelligence professional with deep experience in the Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years.

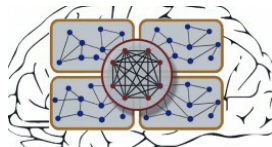
## RELATED ARTICLES

[PRANAV DAR](#), JULY 2, 2018



[The Top GitHub Repositories & Reddit Threads Every Data Scientist should know \(June 2018\)](#)

[AARSHAY JAIN](#), MARCH 16, 2016



[Fundamentals of Deep Learning – Starting with Artificial Neural Network](#)

[PRANAV DAR](#), APRIL 9, 2018



[AVBytes: AI & ML Developments this week – Comet.ml for ML Models, TensorFlow.js, a Python ANN Visualizer, etc.](#)

[FAIZAN SHAIKH](#), DECEMBER 5, 2016



[45 questions to test Data Scientists on Tree Based Algorithms \(Decision tree, Random Forests, XGBoost\)](#)

[AARSHAY JAIN](#), MAY 16, 2018



[19 Data Science and Machine Learning Tools for people who Don't Know Programming](#)

[ANALYTICS VIDHYA CONTENT TEAM](#), ...



[40 Interview Questions asked at Startups in Machine Learning / Data Science](#)

This article is quite old and you might not get a prompt response from the author. We request you to post this comment on Analytics Vidhya's [Discussion portal](#) to get your queries resolved

## 2 COMMENTS



RAGHAVA.R4U

December 31, 2015 at 12:08 pm

[Reply](#)

Superb Writing !!Great



ANJALI

March 15, 2016 at 3:22 am

[Reply](#)

Hello Sir,

I have a question for you. Right now I am a Fresher & soon I am going to work as a junior data scientist in a startup. I would like to know how much it will be beneficial for my career and what can be the growth opportunities in the future since I am working at a startup which is just a few months old.

ANALYTICS VIDHYA

About Us  
Our Team  
Career  
Contact Us  
Write for us

DATA SCIENTISTS

Blog  
Hackathon  
Discussions  
Apply Jobs  
Leaderboard

COMPANIES

Post Jobs  
Trainings  
Hiring Hackathons  
Advertising  
Reach Us

JOIN OUR  
COMMUNITY :

 46336 Followers  
 17840 Followers  
 2987 Followers  
 7513 Followers

Subscribe to emailer

>

