8 Ways to Improve Accuracy of Machine Learning Models (Updated 2023)

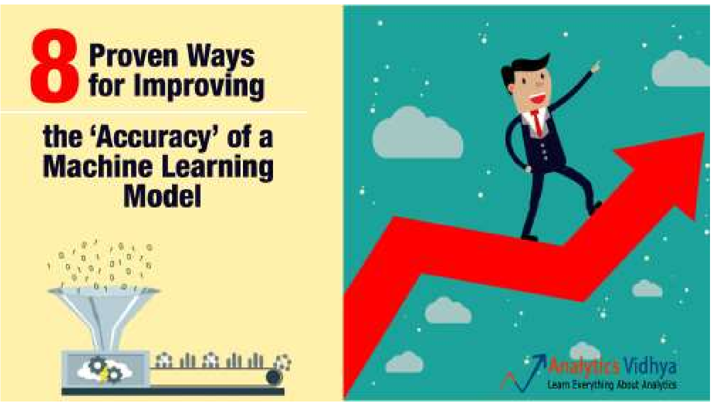
Sunil Ray — Published On December 29, 2015 and Last Modified On March 29th, 2023

Intermediate Machine Learning Technique

Introduction

Enhancing a machine learning model's performance can be challenging at times. Despite trying all the strategies and algorithms you've learned, you tend to fail at improving the accuracy of your model. You feel helpless and stuck. And this is where 90% of the data scientists give up. The remaining 10% is what differentiates a master data scientist from an average data scientist. This article covers 8 proven ways to re-structure your model approach to improve its accuracy.

A predictive model can be built in many ways. There is no 'must-follow' rule. But, if you follow my ways (shared below), you'll surely achieve high accuracy in your models (given that the data provided is sufficient to make predictions). I've learned these methods with experience. I've always preferred to know about these learning techniques practically than digging into theories. In this article, I've shared some of the best ways to create a robust python, machine-learning model. I hope my knowledge can help people achieve great heights in their careers.


8 Proven Ways for Improving
the 'Accuracy' of a Machine Learning Model
Analytics Vidhya
Learn Everything About Analytics

Learning Objectives

- The article aims to provide 8 proven methods for achieving high accuracy in Python ML models.
- It emphasizes the importance of practical learning and structured thinking for improving a data scientist's performance.
- It covers topics such as hypothesis generation, dealing with missing and outlier values, feature engineering, model selection, hyperparameter tuning, and ensemble techniques so that you can increase the performance of the model.

Table of Contents

8 Methods to Boost the Accuracy of an ML 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 underrated step of predictive modeling.

It is important that you spend time thinking about the given problem and gaining 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 dataset. This is a crucial step that usually improves a model's accuracy.

At this stage, you are expected to apply structured thinking to the problem, i.e., a thinking process that 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

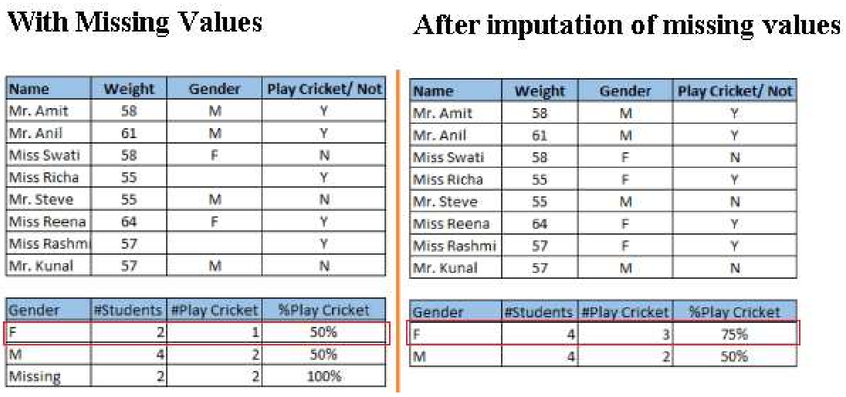
Having more data is always a good idea. It allows the "data to tell for itself" instead of relying on assumptions and weak correlations. Presence of more data results in better and more accurate machine-learning models.

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 real-world company project, I suggest you ask for more data, if possible. This will reduce the 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 trained model or leads to a biased model. It leads to inaccurate predictions. This is because we don't analyze the behavior and relationship with other variables correctly. So, it is important to treat missing and outlier values well.

Look at the below test data snapshot carefully. It shows that, in the presence of missing values, the chances of playing cricket by females are similar to males. But, if you look at the second table (after treatment of missing values based on the salutation "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 trained model. Gladly, we have various methods to deal with missing and outlier values:

1. **Missing:** In the case of continuous variables, you can impute the missing values with mean, median, or mode. For categorical variables, you can treat variables as a separate class. You can also build a model on the training dataset to

predict the missing values. KNN imputation offers a great option to deal with missing values. To know more about these methods, refer to the article “Methods to deal and treat missing values“.

2. **Outlier:** You can delete the observations and perform transformations, binning, or imputation (same as missing values). Alternatively, you can also treat outlier values separately. You can refer article “How to detect Outliers in your dataset and treat them?” to learn more about these methods.

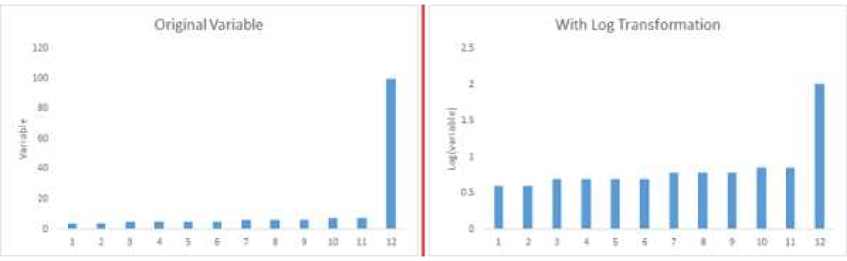
3. Feature Engineering

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 hypothesis generation. Good hypotheses result in good features. That’s why I always suggest investing quality time in hypothesis generation. The feature engineering process can be divided into two steps:

- **Feature transformation:** There are various scenarios where feature transformation is required:
    - Changing the scale of a variable from the original scale to a scale between zero and one. This is known as data normalization.

      For example, Suppose a data set has 1st variable in meter, 2nd in centimeter, and 3rd in kilo-meter, in such case, before applying any algorithm. In that case, we must normalize these variables on the same scale.
    - Some algorithms work well with normally distributed data. Therefore, we must remove the skewness of variable(s). There are methods like a log, square root, or inverse of the values to remove skewness.
    - Sometimes, 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.



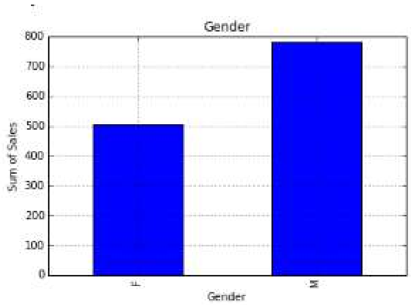
- **Feature Creation:** Deriving new variable(s) from existing variables is known as feature creation. It helps to unleash the hidden relationship of a data set. Let’s say we want to predict the number of transactions in a store based on transaction dates. Here transaction dates may not have a direct correlation with the number of transactions, but if we look at the day of the week, it may have a higher correlation. In this case, the information about the day of the week is hidden. We need to extract it to make the model accuracy better.Note that this might not be the case every time you create new features. This can also lead to a decrease in the accuracy or performance of the trained model. So every time creating a new feature, you must check the feature importance to see how that feature will affect the training process

4. Feature Selection

Feature Selection is a process of finding out the best subset of attributes that better explains the relationship of independent variables with the 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 a higher impact on the target variable.
- **Visualization:** As the name suggests, it helps to visualize the relationship between variables, which makes your variable selection process easier.

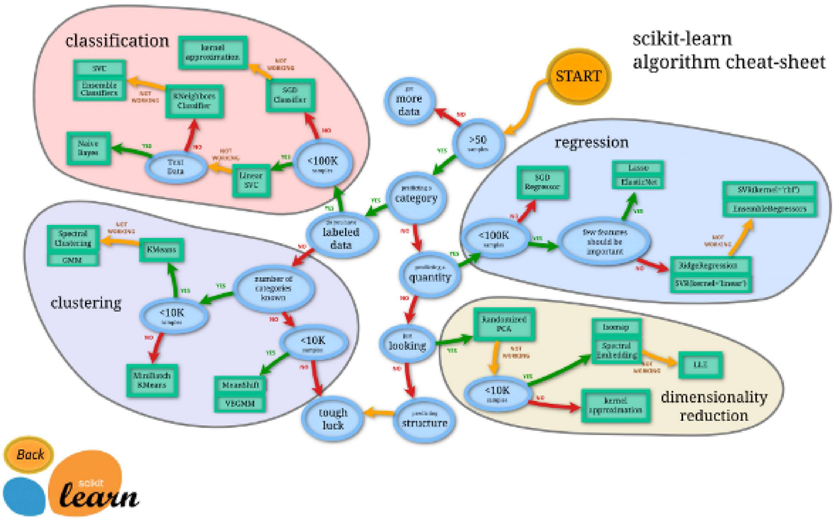


- **Statistical Parameters:** We also consider the p-values, information values, and other statistical metrics to select the right features.
- **PCA:** It helps to represent training data into lower dimensional spaces but still characterizes the inherent relationships in the data. It is a type of dimensionality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

5. Multiple Algorithms

There are many different algorithms in machine learning, but hitting 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 set 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 hyperparameters. These hyperparameters majorly influence the outcome of the learning process.

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

For example: In a random forest, we have various hyperparameters 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 hyperparameter tuning in detail. Below is a random forest scikit learn algorithm with a 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 produces better results. This can be achieved in 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 implement ensemble methods to improve the accuracy of your model. There are two good reasons for this:

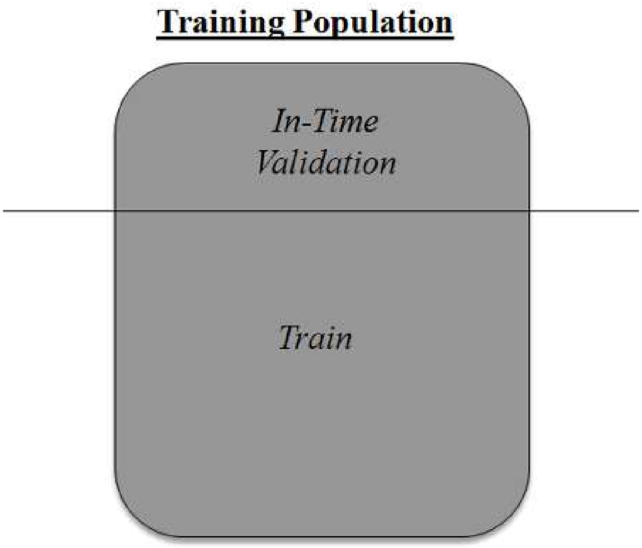
- They are generally more complex than traditional methods.
- The traditional methods give you a good base level from which you can improve and draw from to create your ensembles.

Caution!

Till here, we have seen methods that 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 the model's accuracy can be due to over-fitting too.

8. Cross Validation

To find the right answer to this question, we must use the **cross-validation** technique. Cross Validation is one of the most important concepts in data modeling. It says to 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".

Conclusion

The process of predictive modeling is tiresome. But, if you can think smart, you can outrun your fellow competition easily. Once you get the data set, follow these proven ways, and you'll surely get a robust machine-learning model. But, implementing 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 that can improve the accuracy of a predictive model. These methods are widely known but not used in sequence as defined above.

Key Takeaways

- Generate and test hypotheses to improve model performance.
- Clean and preprocess data to handle missing and outlier values.
- Use feature engineering techniques to create new features from existing data.
- Experiment with different model selection techniques to find the best model for your data.
- Perform hyperparameter tuning to optimize model performance.
- Consider using ensemble techniques to combine multiple models for better performance.
- Focus on practical learning and structured thinking to continuously improve your skills as a data scientist.

Frequently Asked Questions

Q1. How do you increase the accuracy of a regression model?

A. There are several ways to increase the accuracy of a regression model, such as collecting more data, relevant feature selection, feature scaling, regularization, cross-validation, hyperparameter tuning, adjusting the learning rate, and ensemble methods like bagging, boosting, and stacking.

Q2. How do you increase precision in machine learning?

A. To increase precision in machine learning:

- Improve the quality of training data.
- Perform feature selection to reduce noise and focus on important information.

- Optimize hyperparameters using techniques such as regularization or learning rate.
- Use ensemble methods to combine multiple models and improve precision.
- Adjust the decision threshold to control the trade-off between precision and recall.
- Use different evaluation metrics to better understand the performance of the model.

Q3. How can machine learning improve the accuracy of models?

A. Machine learning can improve the accuracy of models by finding patterns in data, identifying outliers and anomalies, and making better predictions. Additionally, ML algorithms can be used to automate many of the tasks associated with model creation which can lead to increased accuracy.

cross-validation Dimensionality Reduction Ensemble Model feature engineering feature selection

missing value treatment Outlier removal PCA

About the Author

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.

Our Top Authors

view more

Previous Post

Year in Review: Best of Analytics Vidhya from 2015

Next Post

New Year Resolutions for a Data Scientist

5 thoughts on "8 Ways to Improve Accuracy of Machine Learning Models (Updated 2023)"

raghava.r4u says:

December 31, 2015 at 12:08 pm

Superb Writing !!Great

Reply

Anjali says:

March 15, 2016 at 3:22 am

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.

Reply

Krish Gupta says:

September 15, 2022 at 3:45 pm

Nice content for learning

Reply

Krish Gupta says:

September 15, 2022 at 3:47 pm

Good content for learning

Reply

Krish Gupta says:

September 15, 2022 at 3:49 pm

It is the process or artificial intelligence so this is very useful for future

Reply

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name*

Email*

Website

□ Notify me of follow-up comments by email.

□ Notify me of new posts by email.

Submit

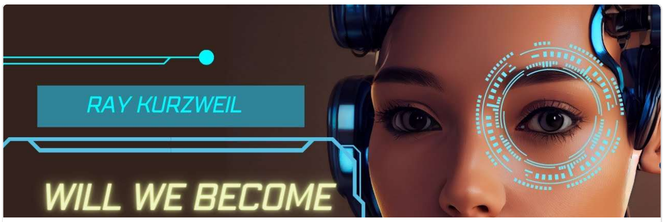
Top Resources

FreedomGPT: Personal, Bold and Uncensored Chatbot Running Locally on Your..

How to Use ChatGPT as a Data Scientist?

K.sabreena - APR 08, 2023

Aravindpai Pai - APR 08, 2023



Futurist Ray Kurzweil Claims Humans Will Achieve Immortality by 2030

K.sabreena - APR 06, 2023

Understand Random Forest Algorithms With Examples (Updated 2023)

Sruthi E R - JUN 17, 2021