f (https://www.facebook.com/AnalyticsVidhya) | 🐦 (https://twitter.com/analyticsvidhya)

8+ (https://plus.google.com/+Analyticsvidhya/posts)

in (https://www.linkedin.com/groups/Analytics-Vidhya-Learn-everything-about-5057165) |

☰ |

Analytics Vidhya (http://www.analyticsvidhya.com)
Learn Everything About Analytics

(http://datahack.analyticsvidhya.com/contest/date-your-data)

Home (http://www.analyticsvidhya.com/) » Business Analytics (http://www.analyticsvidhya.com/blog/category/business-analytics)

Extracting right variables for your Regression model

BUSINESS ANALYTICS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/)   SAS
(HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/SAS/)

f (http://www.facebook.com/sharer.php?u=http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-
tracting%20right%20variables%20for%20your%20Regression%20model) 🐦 (https://twitter.com/home?
xtracting%20right%20variables%20for%20your%20Regression%20model+http://www.analyticsvidhya.com/blog/2013/12/extracting-
s-regression-model-2/) 8+ (https://plus.google.com/share?url=http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-
on-model-2/) P (http://pinterest.com/pin/create/button/?url=http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-
on-model-2/&media=&description=Extracting%20right%20variables%20for%20your%20Regression%20model)
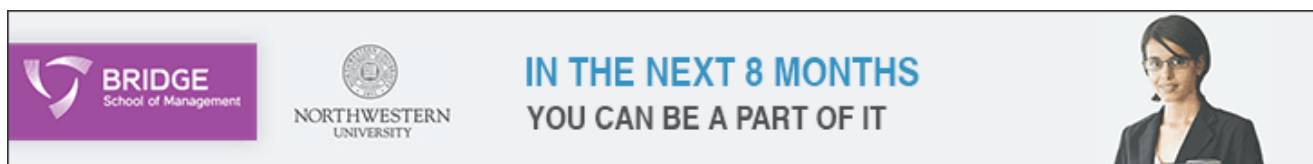
Getting the right variables in your model and cleaning them can make or break your model.

The precision of the model depends on the breadth (diversity) and depth (spread of data and correct transformations) of variables. This article will take you through some of the techniques used in the industry to create or transform variables. We will also cover the techniques used in the industry to select the right set of variables out of an exhaustive list created in our next article on the subject.

(http://i0.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/12/variables2.jpg)

[stextbox id="section"]Types of variables: [/stextbox]
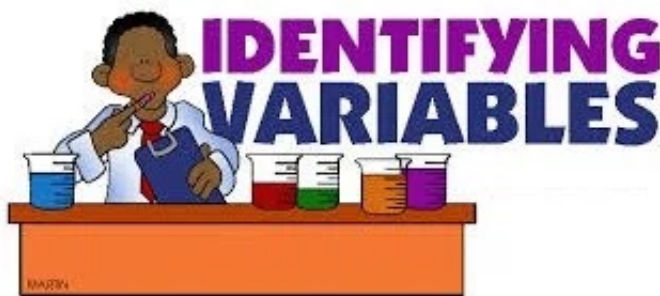
Let's categorize the possible variables to make the discussion easier and our analysis more structured. Following are the categories of variables we will discuss in this article :

1. Basic variable set

2. Derived variable set

3. Mathematically transformed variable set

4. Bin Variable set

5. Co-variant variable set

Types of variables

Basic variables	Derived variables	Mathematically transformed	Binned variables	Co-variant variable
• Hypothesis driven • Collect all available data • e.g. Demographics, behavioural, psychometric	• Combination of basic variables • e.g. Difference in dates, ratios	• Check all possible transformations • Sine, Cosine, Logarihmic, exponential	• Use bi-variate plots • Discover intervals where relation breaks / changes	• X-factor terms

(http://i2.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/12/types_of_variables.jpg)

Following is a business case we will consider for creating these variable sets :

Predict the total business of a insurance branch in next 3 months.

[stextbox id="section"]1. Creating the basic variable set : [/stextbox]

Get this step right and you are half done. There is no set procedure to get the right set of base variables. There are two approaches to get the set of exhaustive base variable set. Following are the approaches :

1. Create hypothesis which can possibly affect the dependent variable (Here we don't even care if the data for this variable is even available)

2. Enlist all variables available (Here we don't even care if this variable can possibly affect the dependent variable)

Here, we try to find all possible variable which can be collected in the analysis without thinking of shortlisting them for the final model. Each of these hypothesis falls in one of the following 3 categories :

a. **Demographic variables** : These variable defines quantifiable statistics for a data-point. In the current business case we will include variables like : Location of the branch, Number of Sales managers, Mix of designation in the branch etc.

b. **Behavioral variables** : These variables comes from the past performance of the subject. In the current case we will include variables like business done by branch in last quarter, Ticket size of the business done by branch, performance metrics of the sales managers in the branch.

c. **Psychometric variables** : For the current business case we will want to include variables like Net Promoter score of the branch, Employee satisfaction score of the branch etc. These types of variables which generally come from surveys are the psychometric variables.

[stextbox id="section"]2. Creating the derived variable set : [/stextbox]

 After completing the list of the basic variables we move on to the derived variables. These variables have a better predictive power and are very stable. These variables are combination of more than one basic variable. Let's see how can we form derived variables for the case in hand. Following are the possible variables :

a .Revenue generated per resource = Revenue of the branch / Number of sales managers

b. Revenue on investment of the branch = Revenue of the branch / Total Cost of branch

c. Vintage of the branch = Today's date – Date of branch opening

d. Ratio of Senior to Junior employee = # Senior employee/ # Junior employee

This list can go on. Try hard to create all possible combination which are expected to influence the dependent variable. For some more ideas on derived variables, read this article (http://www.analyticsvidhya.com/blog/2013/11/simple-manipulations-extract-data/).

[stextbox id="section"]3. Creating the Mathematically transformed set : [/stextbox]

Till this step we already have all the basic and derived variables. Now is the time to find the best possible transformation for each of these variables. Try to check all possible mathematical transformation such as Sine, Cosine, Logarithmic, Exponential , Square, Square root etc. Once you have all the transformation for each of the variables, you have to choose the transformation that best mimics the dependent variable. This transformation should come from both business sense and statistical method. Following is what I do to choose the best transformation for each variable :

Make a regression model with only 2 steps (step-wise) and the actual dependent variable. Choose the transformation that enter the model in these 2 steps. Check if the transformation found as the best fit makes business sense and use it in the final model. For the current problem in hand Log(Total Revenue) is possibly a better variable than Total Revenue as the marginal addition on prediction will go down as the total revenue increases.

[stextbox id="section"]4. Creating the Bin Variable set : [/stextbox]

Creating bin variables is very essential in regression model. Find the bivariate plot between dependent variable and all other independent variables listed till now. Find the intervals where the relationship breaks in the bivariate. For the case in hand, lets say the performance of the branch is best if "Ratio of Senior to Junior employee" is between 0.9 to 1.1 . In such case we will create a bin variable between 0.9 and 1.1. It will look like as follows :

Bin = 1   for   0.9 <Ratio < 1.1

Bin = 0   otherwise

We will include this bin variable in the regression model.

[stextbox id="section"]5. Creating the Co-variant variable set : [/stextbox]
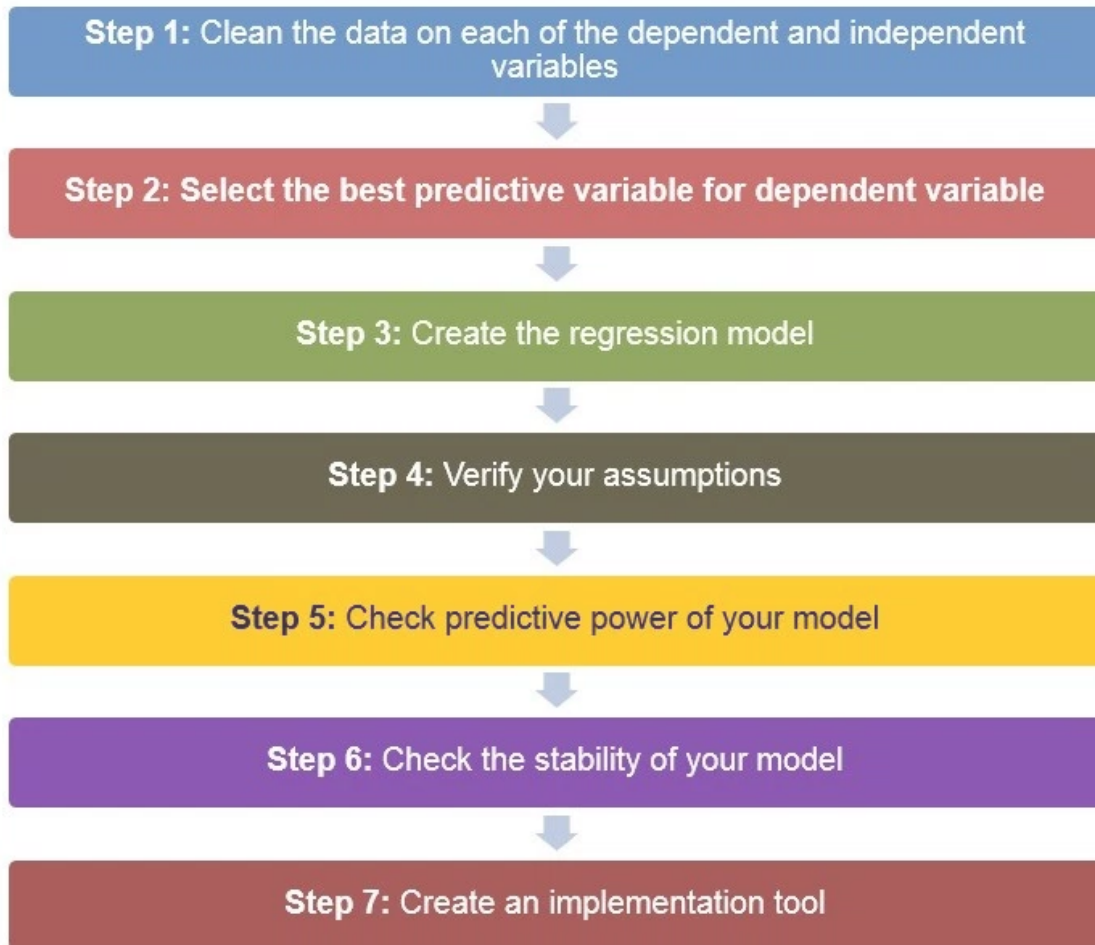
Here comes the X-factor for our regression model. This step incorporate the best cuts of a CART model and significantly raises the prediction power of the regression model. For more details on this technique, please read this article (http://www.analyticsvidhya.com/blog/2013/10/trick-enhance-power-regression-model-2/).

[stextbox id="section"]Next steps to complete the regression model : [/stextbox]

After we are done with the variable collection, following is the order to complete the regression model :

1. Clean the data on each of the dependent and independent variables.

2. Select the best predictive variable for the dependent variable

3. Create the regression model

4. Check the assumptions of the regression model using diagnostic plots (for more details, read this article (http://www.analyticsvidhya.com/blog/2013/12/residual-plots-regression-model/))

5. Check the predictive power of the model

6. Check the stability of the model

7. Create the implementation tool

Steps to complete regression model

Step 1: Clean the data on each of the dependent and independent variables

Step 2: Select the best predictive variable for dependent variable

Step 3: Create the regression model

Step 4: Verify your assumptions

Step 5: Check predictive power of your model

Step 6: Check the stability of your model

Step 7: Create an implementation tool

(http://i2.wp.com/www.analyticsvidhya.com/blog/wp-content/uploads/2013/12/steps_regression_model.jpg)

[stextbox id="section"]End notes : [/stextbox]

The more exhaustive our starting hypothesis the better is the predictive power of the model.Did you find the article useful? Share with us any other variable sets you incorporate in your model. Do let us know your thoughts about this article in the box below.

If you like what you just read & want to continue your analytics learning, subscribe to our emails (http://feedburner.google.com/fb/a/mailverify? uri=analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

Share this:

in (http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=linkedin&nb=1)
28

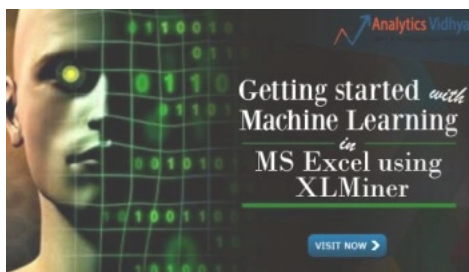f (http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=facebook&nb=1)
8

G+ (http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=google-plus-1&nb=1)

(http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=twitter&nb=1)

(http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=pocket&nb=1)

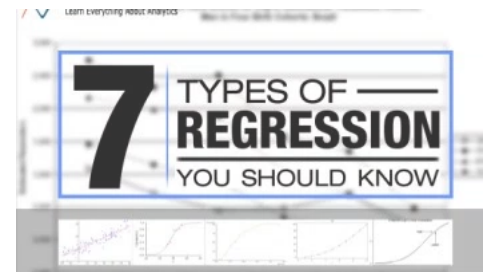(http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/?share=reddit&nb=1)

RELATED

(http://www.analyticsvidhya.com /blog/2015/11/started-machine-learning-ms-excel-xl-miner/)
Getting started with Machine Learning in MS Excel using XLMiner
(http://www.analyticsvidhya.com /blog/2015/11/started-machine-

(http://www.analyticsvidhya.com /blog/2015/06/hackathon-practice-guide-analytics-vidhya/)
The Hackathon Practice Guide by Analytics Vidhya
(http://www.analyticsvidhya.com /blog/2015/06/hackathon-practice-guide-analytics-vidhya/)

(http://www.analyticsvidhya.com /blog/2015/08/comprehensive-guide-regression/)
7 Types of Regression Techniques you should know!
(http://www.analyticsvidhya.com /blog/2015/08/comprehensive-guide-regression/)

learning-ms-excel-xl-miner/)
In "Business Analytics"

In "Analytics Vidhya"

In "Business Analytics"

TAGS: ANALYTICS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ANALYTICS/), BUSINESS ANALYTICS (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BUSINESS-ANALYTICS/), CREATE VARIABLES (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/CREATE-VARIABLES/), LINEAR-REGRESSION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LINEAR-REGRESSION/), LOGISTIC REGRESSION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/LOGISTIC-REGRESSION/), REGRESSION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/REGRESSION/), SELECT VARIABLES (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/SELECT-VARIABLES/), VARIABLE SELECTION (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/VARIABLE-SELECTION/)

Previous Article
Being paranoid about data accuracy! (http://www.analyticsvidhya.com/blog/2013/12/paranoid-data-accuracy-about/)

Next Article
Highlights of 2013 (http://www.analyticsvidhya.com/blog/2013/12/highlights-2013/)

(http://www.analyticsvidhya.com/blog/author/tavish1/)

Author

Tavish Srivastava
(http://www.analyticsvidhya.com/blog/author/tavish1/)

I am Tavish Srivastava, a post graduate from IIT Madras in Mechanical Engineering. I have more than two years of work experience in Analytics. My experience ranges from hands on analytics in a developing country like India to convince banking partners with analytical solution in matured market like US. For last two and a half years I have contributed to various sales strategies, marketing strategies and Recruitment strategies in both Insurance and Banking industry.

6 COMMENTS

Rachit says: REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/?REPLYTOCOM=4124#RESPOND)
JANUARY 4, 2014 AT 4:08 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-4124)

Hi Kunal,

This article is helpful and gives structured understanding of Regression Model.
I would also request you to explain these steps through a case study which can help us in understanding how we
1. clean data
2. select predictors (challenges of dummy variables in case of categorical variables).
3. Verify assumptions
4. Model Validation

Kunal Jain (http://www.analyticsvidhya.com) says: REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/?REPLYTOCOM=4173#RESPOND)
JANUARY 9, 2014 AT 5:46 PM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-4173)

Rachit,

Thanks for the suggestion. Would cover this through one of the articles in coming days. Stay tuned!

Thanks,
Kunal

rohit says: REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/?REPLYTOCOM=6764#RESPOND)
MAY 4, 2014 AT 7:27 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-6764)

HI Kunal,

Can you just give example of transformation of variables and how do we interpret this transformed variables.

For example:

Y=10 Log(x1)+1000/(x2)+100 sqrt(x3)

And also how should we decide which tranformation goes in favour of our model.

Thanks,
Rohit

Tavish Srivastava says: REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/?REPLYTOCOM=6774#RESPOND)
MAY 4, 2014 AT 10:47 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-6774)

Rohit,
When you have a transformed variable in your kitty, the relationship between x and y is no longer linear. Hence the interpretation of coefficient is made in terms of the transformed variable. In your example, a unit increase in the value of log x1 increases the value of y by 10. Coming to your second question, an easy way to choose transformed variables is to build a step wise regression model for all the transformed variables coming from one input variable and choose the final 2 variables entering the model. This makes sure that the cgi squared value of the transformed variable is high. If the original variable has a much higher chi squared value, the transformed variables will not add much value to your model.

Hope this helps.
Tavish

All you need to know to start a career in analytics (http://www.analyticsvidhya.com/blog/2014/12/start-career-analytics-industry/) says:
DECEMBER 7, 2014 AT 9:24 PM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-48269)
[…] 1. http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/ (http://www.analyticsvidhya.com/blog/2013/12/extracting-variables-regression-model-2/) […]

Arindam says: REPLY (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/?REPLYTOCOM=68242#RESPOND)
JANUARY 16, 2015 AT 11:26 AM (HTTP://WWW.ANALYTICSVIDHYA.COM/BLOG/2013/12/EXTRACTING-VARIABLES-REGRESSION-MODEL-2/#COMMENT-68242)

How do you interpret the trignometric function of independent var in business sense. Can you explain with an example.

Thanks.

LEAVE A REPLY

Connect with:

f (http://www.analyticsvidhya.com/wp-login.php?
action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=http%3A%2F%2Fv
variables-regression-model-2%2F)

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

SUBMIT COMMENT

☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

TOP USERS

Rank	Name	Points
1	Nalin Pasricha (http://datahack.analyticsvidhya.com/user/profile/Nalin)	3478
2	SRK (http://datahack.analyticsvidhya.com/user/profile/SRK)	3364
3	Aayushmnit (http://datahack.analyticsvidhya.com/user/profile/aayushmnit)	3075
4	binga (http://datahack.analyticsvidhya.com/user/profile/binga)	2623
5	vikash (http://datahack.analyticsvidhya.com/user/profile/vikash)	2190

More Rankings (http://datahack.analyticsvidhya.com/users)

POPULAR POSTS

- Free Must Read Books on Statistics & Mathematics for Data Science
  (http://www.analyticsvidhya.com/blog/2016/02/free-read-books-statistics-mathematics-data-science/)
- A Complete Tutorial to Learn Data Science with Python from Scratch
  (http://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/)
- Essentials of Machine Learning Algorithms (with Python and R Codes)
  (http://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/)
- A Complete Tutorial on Time Series Modeling in R
  (http://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/)
- Complete guide to create a Time Series Forecast (with Codes in Python)
  (http://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/)
- 4 tricky SAS questions commonly asked in interview
  (http://www.analyticsvidhya.com/blog/2013/11/4-sas-tricky-analytics-interview/)
- SAS vs. R (vs. Python) – which tool should I learn?
  (http://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/)
- 7 Important Model Evaluation Error Metrics Everyone should know
  (http://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/)

(http://imarticus.org/programs/business-analytics-

professional/)

RECENT POSTS

(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch (http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)
MANISH SARASWAT , FEBRUARY 28, 2016

(http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-

segmentation/)

Guide to Build Better Predictive Models using Segmentation (http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)
GUEST BLOG , FEBRUARY 26, 2016

(http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-

salary-report-2016/)

Quick Insights: India Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

KUNAL JAIN , FEBRUARY 24, 2016

(http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/)

India Exclusive: Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/)

KUNAL JAIN , FEBRUARY 22, 2016

GET CONNECTED

4,162
FOLLOWERS
(http://www.twitter.com/analyticsvidhya)

11,954
FOLLOWERS
(http://www.facebook.com/Analyticsvidhya)

g+ **915**
FOLLOWERS
(https://plus.google.com/+Analyticsvidhya)

Email
SUBSCRIBE
(http://feedburner.google.com/fb/a/mailverify?

(http://www.analyticsvidhya.com/blog/2016/02/quick-

ABOUT US

insights-analytics-big-data-salary-report-2016/)

For those of you, who are wondering what is "Analytics Vidhya", "Analytics" can be defined as the science of extracting insights from raw data. The spectrum of analytics starts from capturing data and evolves into using insights / trends from this data to make informed decisions.

STAY CONNECTED

4,162
FOLLOWERS
(http://www.twitter.com/analyticsvidhya)

11,954
FOLLOWERS
(http://www.facebook.com/Analyticsvidhya)

g+ **915**
FOLLOWERS
(https://plus.google.com/+Analyticsvidhya)

Email
SUBSCRIBE
(http://feedburner.google.com/fb/a/mailverify?
uri=analyticsvidhya)

LATEST POSTS

(http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-

science-scratch/)

A Complete Tutorial to learn Data Science in R from Scratch (http://www.analyticsvidhya.com/blog/2016/02/complete-tutorial-learn-data-science-scratch/)
MANISH SARASWAT , FEBRUARY 28, 2016

(http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-

segmentation/)

Guide to Build Better Predictive Models using Segmentation (http://www.analyticsvidhya.com/blog/2016/02/guide-build-predictive-models-segmentation/)
GUEST BLOG , FEBRUARY 26, 2016

(http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)

Quick Insights: India Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/quick-insights-analytics-big-data-salary-report-2016/)
KUNAL JAIN , FEBRUARY 24, 2016

(http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/)

India Exclusive: Analytics and Big Data Salary Report 2016 (http://www.analyticsvidhya.com/blog/2016/02/analytics-big-data-salary-report-2016/)
KUNAL JAIN , FEBRUARY 22, 2016

QUICK LINKS

Home (http://www.analyticsvidhya.com/)

About Us (http://www.analyticsvidhya.com/about-me/)

Our team (http://www.analyticsvidhya.com/about-me/team/)

Privacy Policy (http://www.analyticsvidhya.com/privacy-policy/)

Refund Policy (http://www.analyticsvidhya.com/refund-policy/)

Terms of Use (http://www.analyticsvidhya.com/terms/)

TOP REVIEWS