

The work contained and presented here is our team's work and our team's work alone

# KICKSTARTER PROJECT REPORT

BY:

## TEAM 2

Balakrishnan Anisha,  
Chen Xi,  
Doddagaddavallinarasimhamurthy Amoghbharadwaj,  
Lakamsani Vinay Vihari,  
Lalanne Nerlande

# Table of Contents

<u>Problem Statement.....</u>	3
<u>Executive Summary.....</u>	4
<u>Data Dictionary.....</u>	7
<u>Data Cleaning.....</u>	11
<u>Inconsistent Data.....</u>	11
<u>Outlier Analysis.....</u>	12
<u>Missing Data .....</u>	15
<u>Interesting Patterns.....</u>	16
<u>Classification Modeling.....</u>	20
<u>Decision Tree.....</u>	20
<u>Boosted Tree.....</u>	22
<u>Discriminant Analysis.....</u>	23
<u>Neural Network.....</u>	25
<u>Clustering Analysis.....</u>	26
<u>Regression Modeling.....</u>	28
<u>Logistic Regression.....</u>	28
<u>Generalized Regression.....</u>	30
<u>Stepwise .....</u>	31
<u>Standard Least Square.....</u>	32
<u>Ensemble Modeling.....</u>	34
<u>Comparison of Models.....</u>	43
<u>Comparison of Cost of Errors.....</u>	44
<u>Proposed Model.....</u>	45
<u>Conclusion.....</u>	46
<u>Appendix.....</u>	50

## Problem Statement

The objective of our project is to predict if a campaign will be successful or unsuccessful before its completion using the Kickstarter funding platform.

	ID	name	category	main_category	currency	deadline	goal	launched	pledged	state	backers	country	usd pledged
Rows	323,750	32	100012970 Chris Eger Band - New ...	Music	USD	08/13/2014 8:59 AM	10000	07/14/2014 10:35 PM	13260	successful	92 US	13260	
All rows	323,750	33	1000131947 Arrows & Sound Debut ...	Indie Rock	USD	05/19/2012 1:04 AM	4000	04/19/2012 10:4 AM	8641.34	successful	157 US	8641.34	

## Executive Summary

The goal of this project is to predict the success or failure of a Kickstarter campaign at the time it is launched. We feel that this analysis is important since only 35% of campaigns reach their funding goals, Kickstarter has to know the factors that play a role in a project's outcome before launching. Also, we can suggest to backers as to which project is prone to be more successful. The projects that we analyzed fell into 15 categories and 13 columns. The categories are Art, Comics, Dance, Design, Fashion, Film & Video, Food, Games, Journalism, Music, Photography, Publishing, Technology and Theater.

Our team started with a dataset of 323,750 observations with 13 variables. Each project was given a status (state) of either canceled, failed, live, successful, suspended and undefined. In order to clean and analyze the data, we deleted all rows that had inconsistencies and errors, therefore reducing the dataset to 320,661. Our analysis started in 2009 and ended in 2016 with total pledged amount of over \$2 billion. The funding goal for projects ranged from \$0.01 to \$126,720,000. About 65 % projects failed to meet the minimum goal.

We found that the column “USD\_PLEDGED” has missing data, so we created a new column to reflect the amount of pledged USD. In addition, we analyzed the outliers for each column and decided to keep the outliers, because we need the information to build the model. Furthermore, there are observations where Backers equals 0 but Pledged amount is greater than zero. This is inconsistent as there should be at least one backer to pledge an amount for funding the project. As a result, we deleted rows where data was inconsistent.

Because we were interested in discovering trends or strong indicators that would distinguish between successful and failed projects, we focused on analyzing important statistics of various projects. We found that successful projects had funding goals averaging six times less than failed projects as well as Pledged amounts averaging eleven times more than failed projects.

Another identifiable trend is that the Pledged USD amount increases with the Duration of the campaign. [Image 1](#) shows that if the Duration is between 30 and 50 days, funding amount increases. Another trend revealed is that more Backers are attracted to projects originating in the United States as compared to originating in other countries ([Image 2](#)).

Also, [Image 3](#) reveals that more Backers are attracted to the project when its funding goal amount is small. Therefore, we can hypothesize that smaller projects are more likely to attract a larger number of backers. In addition, we found that more Backers are attracted when the campaign is paid for in USD ([Image 4](#)).

Another analysis was carried out as to whether certain categories performed better than others. ([Image 5 & 6](#)) As shown in image 5, the category of Games received the largest number of

Backers, whereas Crafts received the least number of Backers. As shown in [image 6](#), the least amount Pledged USD was for Crafts and Design and the maximum Pledged USD was for Design. Therefore, we conclude that Games is the category that's in demand.

[Image 7](#) shows additional data for average backers of the projects across all categories. Because of the 2009 recession, there was a decline in the number of Backers due to the high risk and fear dominating the investment world. As a matter of fact, [image 7](#) reveals that from 2012 – 2016 the number of investors of Backers increased.

[Image 8](#) shows the relationship between the number of backers and the success or failure of a campaign and reveals that when the number of backers is reduced, the campaign is more likely to fail, irrespective of the state of the economy.

Once we identified various patterns, we were able to build many models such as classification models, regression models and ensemble models. We tried different split percentages for training, validation and test data set so that we can better determine which model is best. We feel that the decision tree model is the best model since it is easier to explain and it provides the highest accuracy.

Finally, we calculated the mean of ratio of Total USD Pledged to the number of Backers, which is the estimated loss value for False Negative Cases. When mean equals \$68.32, the total amount being pledge directly will be considered as a loss. We also calculated the Cost of Error for False Negative which is \$68.32. Another calculation was in the case of not investing in a project that will be successful. In this case the investor will lose profit expect if he can get a general bank interest rate of 10%. On the other hand, if someone invested in the project, the profit could be at least 40% of the investment. However, the cost of error for false positive equals to \$20.49.

In conclusion, we came up with the following recommendations:

- Since the Country column has 3,797 observations (1%) with the value 'NO', it should be made a required field because the currency of Goal and Pledged amounts depend on that variable. In addition, data for the Country variable should be carefully collected. As a result, data for the Country variable should be carefully collected.
- For data to be valuable and accurate, Pledged amounts should always be associated with a Backer. For example, we identified 3,082 records that had zero Backers but some Pledged amounts. When this occurs, data is not useful. Therefore, the business should be sure that Backers are consistent with Pledged variables.

- Projects with Goal amounts over 2 Million should not be considered for launching a campaign since they are more likely to get Failed or Canceled as there are less chances of being backed.
- The columns Estimated Start Date and End Date should always be included in analyzing and modeling data so as to provide more accuracy in predicting duration of projects.
- To be more valuable to marketers, data should include specific location of project being launched.
- It is recommended that the company should take efforts to verify the Goal amount so that it is in line with the project and we do not use false values to make predictions.

## DATA DICTIONARY

The Kickstarter dataset (see below) contains 323,751 rows and 13 columns.

### EXISTING VARIABLES

Variable Name	Old Variable Type	Final Variable type	Description of the variable
ID	Continuous	Nominal	Internal Kickstarter ID
NAME	Nominal	Nominal	Name of the project
CATEGORY	Nominal	Nominal	Sub category under which the projects fall
MAIN_CATEGORY	Nominal	Nominal	Category under of project
CURRENCY	Nominal	Nominal	Currency used to support
DEADLINE	Nominal	Continuous	Deadline for crowdfunding
GOAL	Nominal	Continuous	Amount of money required to complete the project
LAUNCHED	Nominal	Continuous	Launched date for crowdfunding
PLEDGED	Nominal	Continuous	Amount pledged by crowd
STATE	Nominal	Nominal	The current condition the project is in
BACKERS	Nominal	Continuous	The number of people who have backed the project

COUNTRY	Nominal	Nominal	*The country where the project is being created.
USD PLEDGED	Nominal	Continuous	The value of pledged amount in USD currency with currency value as per the date when it was pledged

\*As per the old Data Dictionary, the Country column shows the country which originated the pledged amount. Since people can pledge money from more than one country. Hence, there cannot be one country name if the column contains the country from which money was pledged, the column should contain the country where the project is being conducted.

## NEW VARIABLES

Below is the list of new variables created for the purpose of Exploratory Data Analysis and Modeling.

Variable Name	Variable Type	Description of the variable
CURRENCY_RATE	Continuous	The foreign exchange rate of every country from the country column since 10/31/2018
FINAL STATE	Nominal	This column has value 1 for corresponding value on status column as 'Successful' and value 0 for corresponding status as 'Cancelled', 'Suspended', 'Failed'
DEADLINE_YEAR	Continuous	The year of the deadline of crowdfunding from the Deadline column
LAUNCHED_YEAR	Continuous	The year of launch of crowdfunding from the Launched column
GOAL_USD	Continuous	The Goal amount in USD currency
PLEDGED_USD	Continuous	The Pledged amount in USD currency

FUNDING_DURATION	Continuous	The difference of Launched and Deadline date of crowdfunding
RATIO_OF_PLEDGED_TO_GOAL	Continuous	The ratio of the pledged amount to the goal amount
RATIO_OF_PLEDGED_TO_BACKERS	Continuous	The ratio of pledged amount to the number of backers
RANGE SCALE[BACKERS]	Continuous	The number of Backers in the scale of 0 to 1
RANGE SCALE[GOAL_USD]	Continuous	The goal amount in USD in the scale of 0 to 1
RANGE SCALE[PLEDGED_USD]	Continuous	The pledged amount in USD in the scale of 0 to 1
RANGE SCALE[FUNDING DURATION]	Continuous	The duration between launched date and the deadline of crowdfunding

## DATA CLEANING

In order to explore the dataset, it is necessary for the data to be accurate, complete and relevant to the business so that we can predict the outcome of the project.

### 1. Resolve Inconsistencies

There are some observations where Backers equal zero but Pledged amount are greater than zero. This is inconsistent as there should be at least one backer to pledge an amount for funding the project.

Therefore, 3082 records with Backers = 0 and Pledged amount > 0 are removed.

ID	name	description of project 1	description of project 2	description of project 3	description of project 4	category	main_category	current currency
1	1000002330	The Songs of ...				Poetry	Publishing	GBP
2	1000004030	Where is Hank?				Narrative Film	Film & Video	USD
3	1000007540	ToshiCapital ...				Music	Music	USD
4	1000011046	Community Film ...				Film & Video	Film & Video	USD
5	1000014025	Monarch ...				Restaurants	Food	USD
6	1000023410	Support Solar ...				Food	Food	USD
7	1000030581	Chaser Strips ...				Drinks	Food	USD
8	1000034510	SPIN - Premium ...				Product Design	Design	USD
9	100004195	STUDIO IN THE ...				Documentary	Film & Video	USD
10	100004721	Of Jesus and ...				Nonfiction	Publishing	CAD
11	100005484	Lisa Lim New CD!				Indie Rock	Music	USD
12	1000055792	The Cottage ...				Crafts	Crafts	USD
13	1000056157	G-Spot Place for ...				Games	Games	USD
14	1000064360	Survival Rings				Design	Design	USD
15	1000064918	The Beard				Comic Books	Comics	USD
16	1000068480	Notes From ...				Art Books	Publishing	USD
17	1000070642	Mike Corey's ...				Music	Music	USD
18	1000071625	Boco Tea				Food	Food	USD
19	1000072011	CMUK Shoes:...				Fashion	Fashion	USD
20	1000082254	Alice in ...				Theater	Theater	USD
21	1000087442	Mountain brew: ...				Drinks	Food	NOK
22	1000091520	The Book Zoo - ...				Comics	Comics	USD
23	1000102741	Matt Cavanaugh ...				Music	Music	USD
24	1000103940	Superhero ...				DIY	Crafts	GBP
25	1000104688	Permaculture Skills				Webseries	Film & Video	CAD
26	1000104953	Rebel Army ...				Comics	Comics	GBP
27	1000115172	Daily Brew Coffee				Food Trucks	Food	GBP
28	1000117861	Ledr workbook: ...				Product Design	Design	USD

There are some observations with a Launched year as 1970 and funding duration as long as 30 years. This is impractical as the Kickstarter website was created in the year 2009 and any projects with launch date before 2009 will be bogus.

Therefore, these 7 records were also deleted.

The top window is titled "Select Rows - JMP Pro". It displays a search interface where "launched" is selected in a dropdown menu, and the search term "1970" is entered into a "contains" field. The "Action" buttons include OK, Cancel, Recall, and Help.

The bottom window is titled "ks projects 201612 data cleaned - JMP Pro". It shows a data table with columns: ID, name, description of project 1, description of project 2, description of project 3, description of project 4, category, main\_category, currency, deadline, and several other columns like description of project 5-8, goal, pledged, state, backers, country, and usd pledged. The table has 320,668 rows. A sidebar on the left shows the column structure and row counts: All rows (320,668), Selected (7), Excluded (0), Hidden (0), and Labelled (0).

## 2. Outlier Analysis

The GOAL column with amount greater than 2 million USD had project status as 'Failed' or 'Cancelled'. However, they cannot be omitted because they are genuine values and thus values above 2 million USD should be considered.

Some categories have records with very high Pledged values over 20 million USD. Since they are true values and are verified by the Kickstarter website, they are not omitted.

[Reference – Univariate Outlier Analysis Screenshots Appendix 1.1.](#)

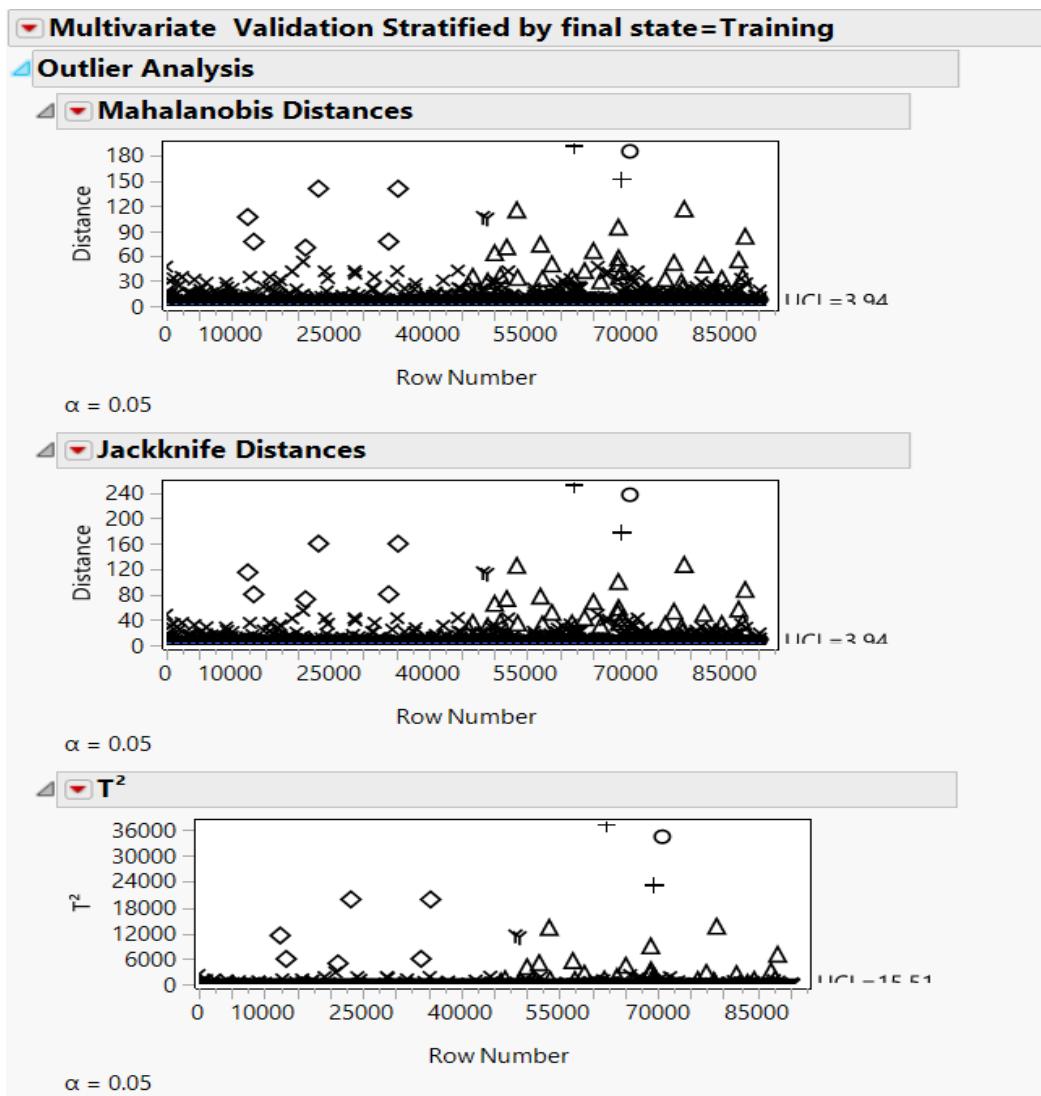
### Univariate outlier analysis:

Variable Name	Variable Type	Outlier Analysis
CURRENCY_RATE	Continuous	We are not using this variable for modeling
FINAL STATE	Nominal	As it is a categorical variable we cannot do outlier analysis
DEADLINE_YEAR	Continuous	There are no outliers
LAUNCHED_YEAR	Continuous	There are no outliers
GOAL_USD	Continuous	The Goal amount greater than 2 million USD have project status as Failed. But they cannot be omitted as they are true values.
PLEDGED_USD	Continuous	The Pledged amount in USD which are greater than 20million USD are genuine values thus these observations are considered for the predictions
FUNDING_DURATION	Continuous	After deleting observations with launched date as 1970, we have no outliers
RATIO_OF_PLEDGED_TO_GOAL	Continuous	This variable is not used for modeling
RATIO_OF_PLEDGED_TO_BACKERS	Continuous	This variable is not used for modeling
RANGE SCALE[BACKERS]	Continuous	This variable is not used for modeling
RANGE SCALE[GOAL_USD]	Continuous	Same as Goal_USD
RANGE SCALE[PLEDGED_USD]	Continuous	Same as Pledged_USD
RANGE SCALE [FUNDING DURATION]	Continuous	Same as Funding duration

## Multivariate outlier Analysis:

# Modelling prediction saved - Multivariate - JMP Pro

	deadline year	launched year	Range Scale[backers]	Range Scale[goal usd]	Range Scale[pledged usd]	Range Scale[funding duration]	ratio pledge of goal	ratio pledged by backers
deadline year	1.0000	0.9897	0.0153	0.0206	0.0243	-0.1643	0.0065	-0.0024
launched year	0.9897	1.0000	0.0159	0.0191	0.0247	-0.1856	0.0067	-0.0029
Range Scale[backers]	0.0153	0.0159	1.0000	0.0119	0.7975	-0.0005	0.0210	0.0113
Range Scale[goal usd]	0.0206	0.0191	0.0119	1.0000	0.0130	0.0203	-0.0007	0.0016
Range Scale[pledged usd]	0.0243	0.0247	0.7975	0.0130	1.0000	0.0099	0.0155	0.1009
Range Scale[funding duration]	-0.1643	-0.1856	-0.0005	0.0203	0.0099	1.0000	-0.0069	0.0299
ratio pledge of goal	0.0065	0.0067	0.0210	-0.0007	0.0155	-0.0069	1.0000	0.0021
ratio pledged by backers	-0.0024	-0.0029	0.0113	0.0016	0.1009	0.0299	0.0021	1.0000



From the above image, we can see few outliers but we are not removing them because all of them are true values and were verified by the Kickstarter website.

### **3. Missing data**

There are 465 observations with COUNTRY column labeled as 'NO', the corresponding USD\_PLEDGED values are missing. Since the currency exchange rate fluctuates, we use the exchange rate at 2018.10.31 to calculate the amount of pledged USD. We created a new column called PLEDGED\_USD. This newly created column is used for further analysis and modeling.

**The final data set after data preprocessing has a total of 320,661 observations in total.**

## INTERESTING PATTERNS

### 1. Pledged USD vs Funding duration

```
sns.lmplot('funding duration','pledged usd', df)
```

```
<seaborn.axisgrid.FacetGrid at 0x19494bcebe0>
```

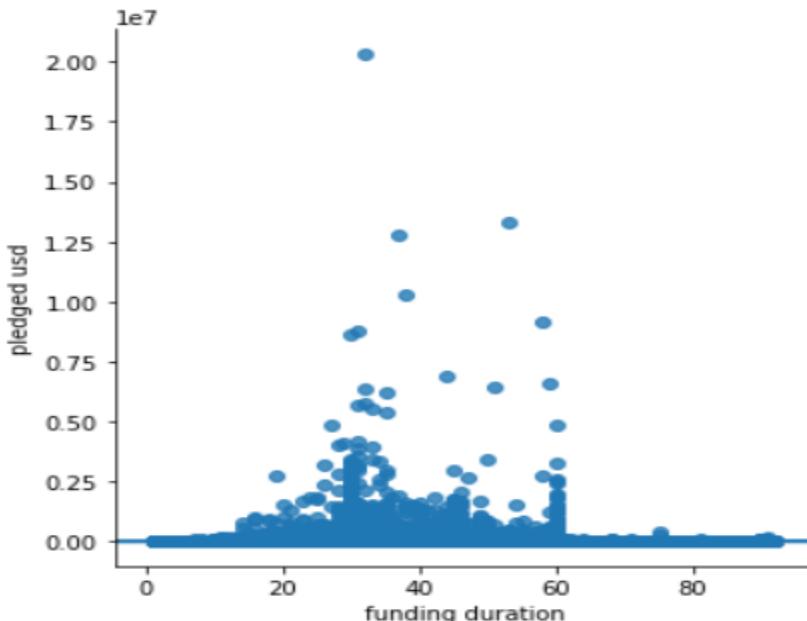


IMAGE1

We observe that If the duration is between 30-50, we get more funding. This is the ideal funding duration for all the projects. The funding duration should not be too small or too long for any Project.

### 2. Country Major vs Backers

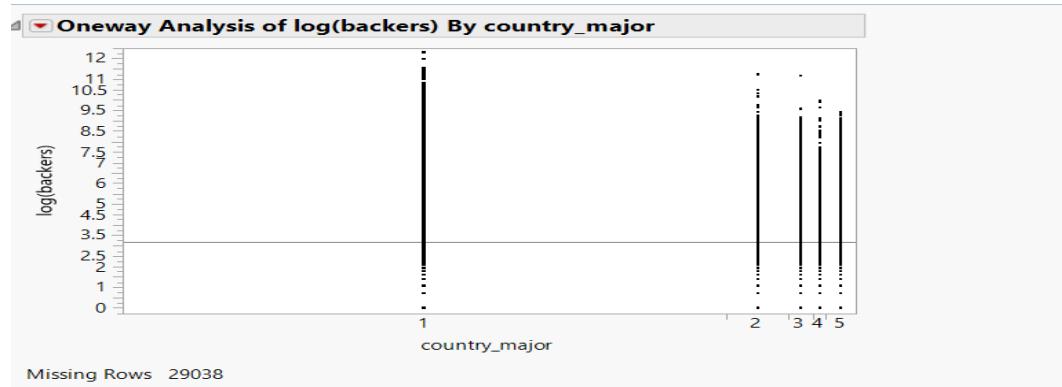


IMAGE2

We can hypothesize that more backers are more likely to invest in US project than any other country. It may be due to the trust that people tend to have in US projects. Also, the technological advantages that are more available in the US than anywhere in the world. Lastly, the United States tends to have a better reputation in regards to successful projects.

### 3. Goal vs Backers

```
plt.figure(figsize=(12,3))
sns.lmplot('goal','backers', data= df, size=2, aspect=4)
```

```
<seaborn.axisgrid.FacetGrid at 0x194958a0e48>
```

```
<Figure size 864x216 with 0 Axes>
```

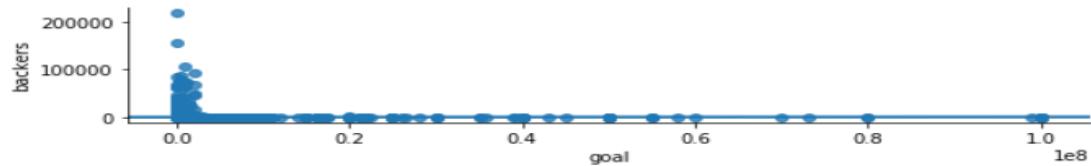


IMAGE3

We can expect more backers if the project's goal amount is small. Maybe small projects are likely to get more backers. Maybe people think that if the funding goal amount is large, it might take a long time to reach its goal and loses its interests.

### 4. Currency vs Backers

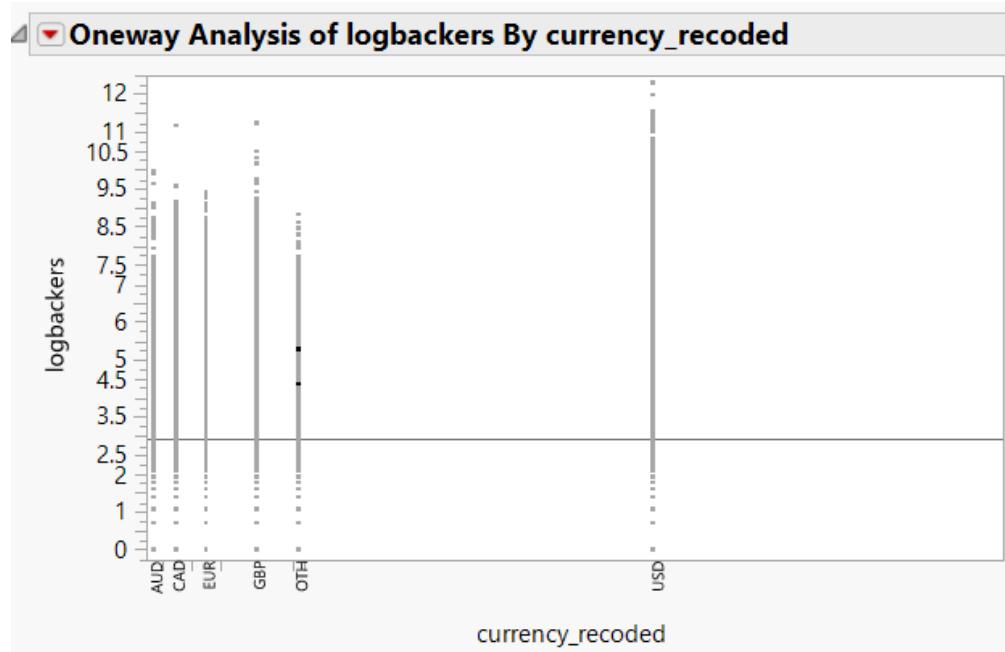


IMAGE4

We can expect more backers for USD currency. It may be due to the fact that the currency USD is universally accepted and it is well-known all over the world. Also, it may be due to the fact that the rate is also not as high as other renowned currencies like Euro, Pound.

## 5. Average backers for each category

```
backers = df.groupby('main_category').mean()['backers']

main_category
Art           39.424094
Comics        134.361882
Crafts         26.921023
Dance          43.272216
Design         237.192436
Fashion        62.283837
Film & Video   67.217146
Food            54.333427
Games           323.892388
Journalism      38.231994
Music           53.710210
Photography     37.304191
Publishing      52.686762
Technology      165.532824
Theater          47.408703
Name: backers, dtype: float64
```

IMAGE5

We see that Games category gets usually large number of Backers and Crafts category with the least number of backers. This maybe a general indicator for the current trend

## 6. Main Category vs Pledged USD

```
df.groupby('main_category').mean()['pledged_usd']

main_category
Art           3009.842974
Comics        6631.307194
Crafts         1538.531379
Dance          3391.538872
Design         22978.299055
Fashion        5531.053544
Film & Video   6128.339944
Food            4955.110715
Games           20129.557520
Journalism      2533.431828
Music           3790.114682
Photography     3156.833010
Publishing      3090.255067
Technology      20143.498302
Theater          3942.357402
Name: pledged usd, dtype: float64
```

IMAGE6

We get less pledged USD for Crafts Category and the max pledged USD is for Design category. Interestingly even though the average backers for Design and Technology was less than Games, we see that the amount raised by these categories are higher than the rest. So, we can conclude there is a big demand for Games, Technology and Design domains.

## 7. Launched Year vs Backers

```
|: df.groupby('launched year').mean()['backers']  
  
|: launched year  
2009    32.925508  
2010    38.683685  
2011    53.225331  
2012    105.526468  
2013    140.346329  
2014    92.373636  
2015    99.555476  
2016    123.476453  
Name: backers, dtype: float64
```

IMAGE7

Due to the recession in 2009-2011, we saw less backers on an average due to the risk of losing money and fear in economy. Apart from that, people are willing to back projects with a reasonable pledged amount. Unless something bad happens, we can have a good number of backers for a project.

## 8.Average backers for a project to be successful and failure in each year

```
df[df['final state'] ==1].groupby('launched year').mean()['backers']  
  
launched year  
2009    66.357513  
2010    77.586109  
2011    101.606442  
2012    215.789012  
2013    285.733790  
2014    252.528949  
2015    317.700873  
2016    288.854610  
Name: backers, dtype: float64  
  
df[df['final state'] ==0].groupby('launched year').mean()['backers']  
  
launched year  
2009    7.116000  
2010    8.526920  
2011    11.362221  
2012    20.750677  
2013    29.439372  
2014    18.816643  
2015    15.592420  
2016    19.508526  
Name: backers, dtype: float64
```

IMAGE8

We observe that if backers increase, the project is more likely to be successful. The increased number of backers might indicate that people's confidence in a project is high. Therefore, this might create a superficial demand for other backers and encourage them to invest on something new.

## Classification Models

With Stratified split of Training set – 0.4, Validation set – 0.4, Test set = 0.2

(We have considered this proportion as this split gave maximum accuracy and minimum False positive and False negative errors)

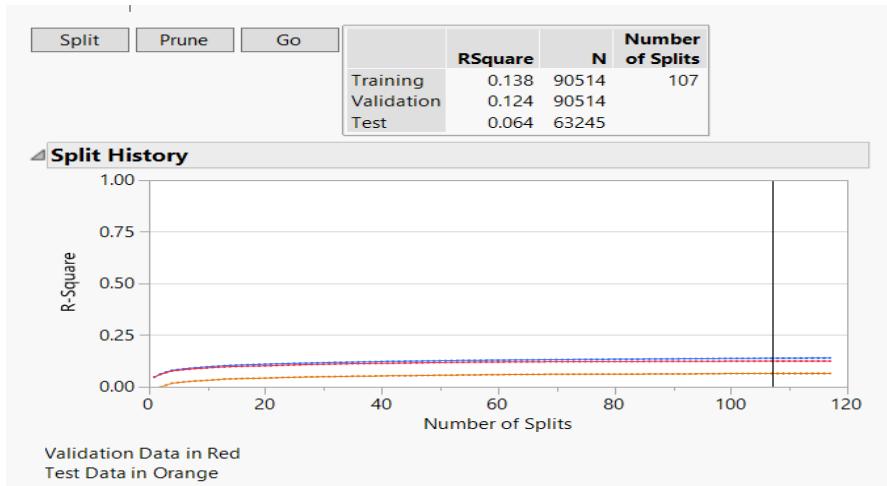
Reference – Decision Tree Model Appendix

### 1. Decision Tree

The screenshot shows the 'Partition - JMP Pro' dialog box for 'Recursive partitioning'. The interface is divided into several sections:

- Select Columns:** A list of 29 columns: RowID, ID, name, category, main\_category, currency, currency rate, deadline, goal, launched, pledged, state, final state, backers, country, deadline year, launched year, goal usd, pledged usd, funding duration, Range Scale[backers], Range Scale[goal usd], Range Scale[pledged usd], Range Scale[funding duration], ratio pledge of goal, ratio pledged by backers, Validation Stratified by final state, SortKey, and RowNumber. The 'Validation Stratified by final state' item is highlighted.
- Cast Selected Columns into Roles:**
  - Y, Response:** final state (optional)
  - X, Factor:** category, main\_category, currency, deadline, launched, county, deadline year, launched year, Range Scale[goal usd], Range Scale[funding duration] (optional)
  - Weight:** optional numeric
  - Freq:** optional numeric
  - Validation:** Validation Stratified by final state
  - By:** optional
- Action:** Buttons for OK, Cancel, Remove, Recall, and Help.
- Options:** Method dropdown set to 'Decision Tree', Validation Portion input field set to 0, and two checked checkboxes: 'Informative Missing' and 'Ordinal Restricts Order'.

## Results:



**Fit Details**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1379	0.1239	0.0644	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2320	0.2104	0.1106	$(1 - L(0)/L(\text{model}))^{(2/n)} / (1 - L(0)^{(2/n)})$
Mean -Log p	0.5975	0.6073	0.6101	$\sum -\text{Log}(p_{ij})/n$
RMSE	0.4541	0.4585	0.4607	$\sqrt{\sum (y_{ij} - p_{ij})^2/n}$
Mean Abs Dev	0.4126	0.4169	0.4177	$\sum  y_{ij} - p_{ij} /n$
Misclassification Rate	0.3269	0.3348	0.3443	$\sum (p_{ij} \neq p_{\text{Max}}) / n$
N	90514	90514	63245	n

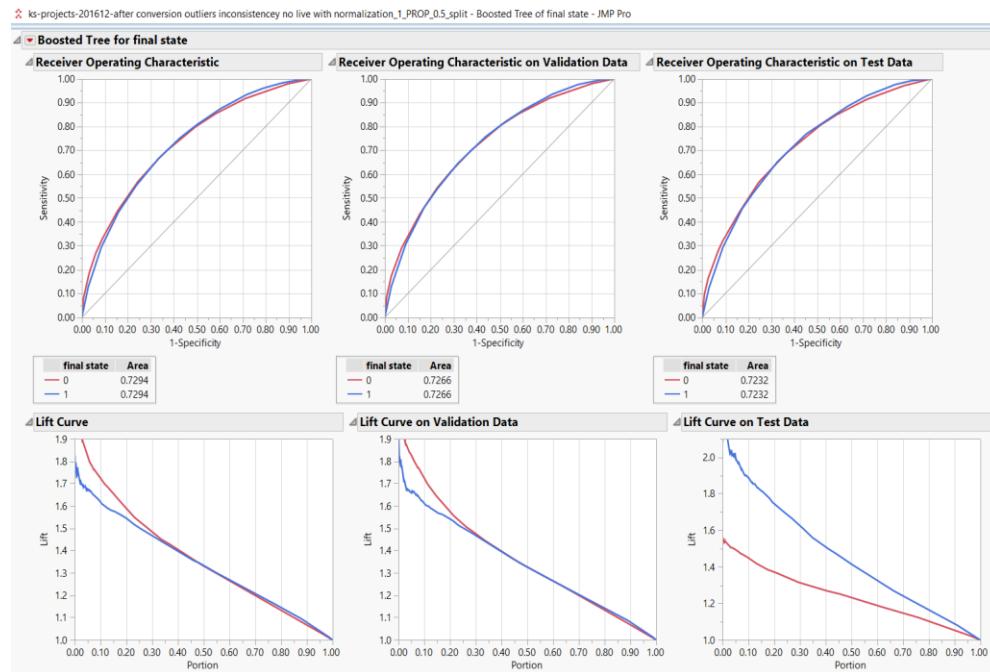
**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state	0 1	final state	0 1	final state	0 1
0	28930 16327	0	28585 16672	0	25693 14923
1	13259 31998	1	13635 31622	1	6855 15774

Dataset	Number of False Positive and False Negative	Accuracy
Training	29586	67.3%
Validation	30307	66.5%
Test	21778	65.5%

Obtaining an accuracy rate of 65.5% helped us to identify the best model.

## 2. Boosted Tree



### Overall Statistics

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1121	0.1098	0.0505	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.1919	0.1882	0.0875	$(1 - L(0)/L(\text{model}))^{(2/n)} / (1 - L(0)^{(2/n)})$
Mean -Log p	0.6155	0.6171	0.6192	$\sum -\text{Log}(\rho[j])/n$
RMSE	0.4622	0.4629	0.4643	$\sqrt{\sum (y[j] - \hat{y}[j])^2 / n}$
Mean Abs Dev	0.4429	0.4438	0.4442	$\sum  y[j] - \hat{y}[j]  / n$
Misclassification Rate	0.3342	0.3376	0.3473	$\sum (\rho[j] \neq \rho[\text{Max}]) / n$
N	90514	90514	63245	n

### Confusion Matrix

		Training		Validation		Test		
		Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count	
Actual	Predicted Count	0	1	0	1	0	1	
final state	final state	0	1	0	1	0	1	
0	28616	16641	0	28476	16781	0	25603	15013
1	13613	31644	1	13773	31484	1	6954	15675

Data set	Number of False positive and False Negative errors	Accuracy
Training	30254	66.57%
Validation	30554	66.24%
Test	21967	65.26%

This model has a bit less accuracy rate than the decision tree. The Decision tree model is better than the Boosted tree model.

### 3.Discriminant Analysis

Discriminant - JMP Pro

Identifying the group using the covariates

Select Columns

35 Columns

- RowID
- ID
- name
- category
- main\_category
- currency
- currency rate
- deadline
- goal
- launched
- pledged
- state
- final state
- backers
- county
- deadline year
- launched year
- goal usd
- pledged usd
- funding duration
- Range Scale[backers]
- Range Scale[goal usd]
- Range Scale[pledged usd]
- Range Scale[funding duration]
- ratio pledge of goal
- ratio pledged by backers
- Validation Stratified by final state
- SortKey
- RowNumber
- Prob(final state==1)

Cast Selected Columns into Roles

Action

OK

Cancel

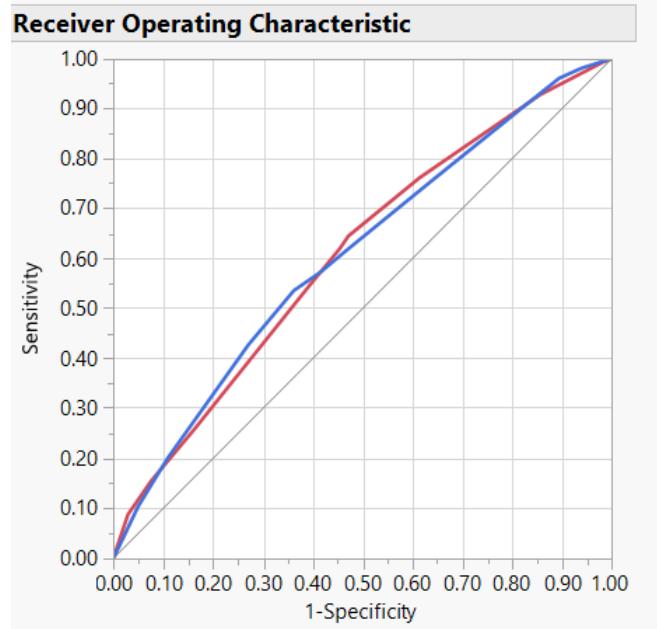
Remove

Recall

Help

Y, Covariates
deadline launched deadline year launched year Range Scale[goal usd] Range Scale[funding duration] <i>optional numeric</i>

X, Categories
final state <i>optional numeric</i>
Weight <i>optional numeric</i>
Freq <i>optional numeric</i>
Validation Validation Stratified by final state
By <i>optional</i>



**Discriminant Analysis**

Discriminant Method: Linear  
 Classification: final state  
 Validation: Validation Stratified by final state

**Canonical Plot**

**Discriminant Scores**

**Score Summaries**

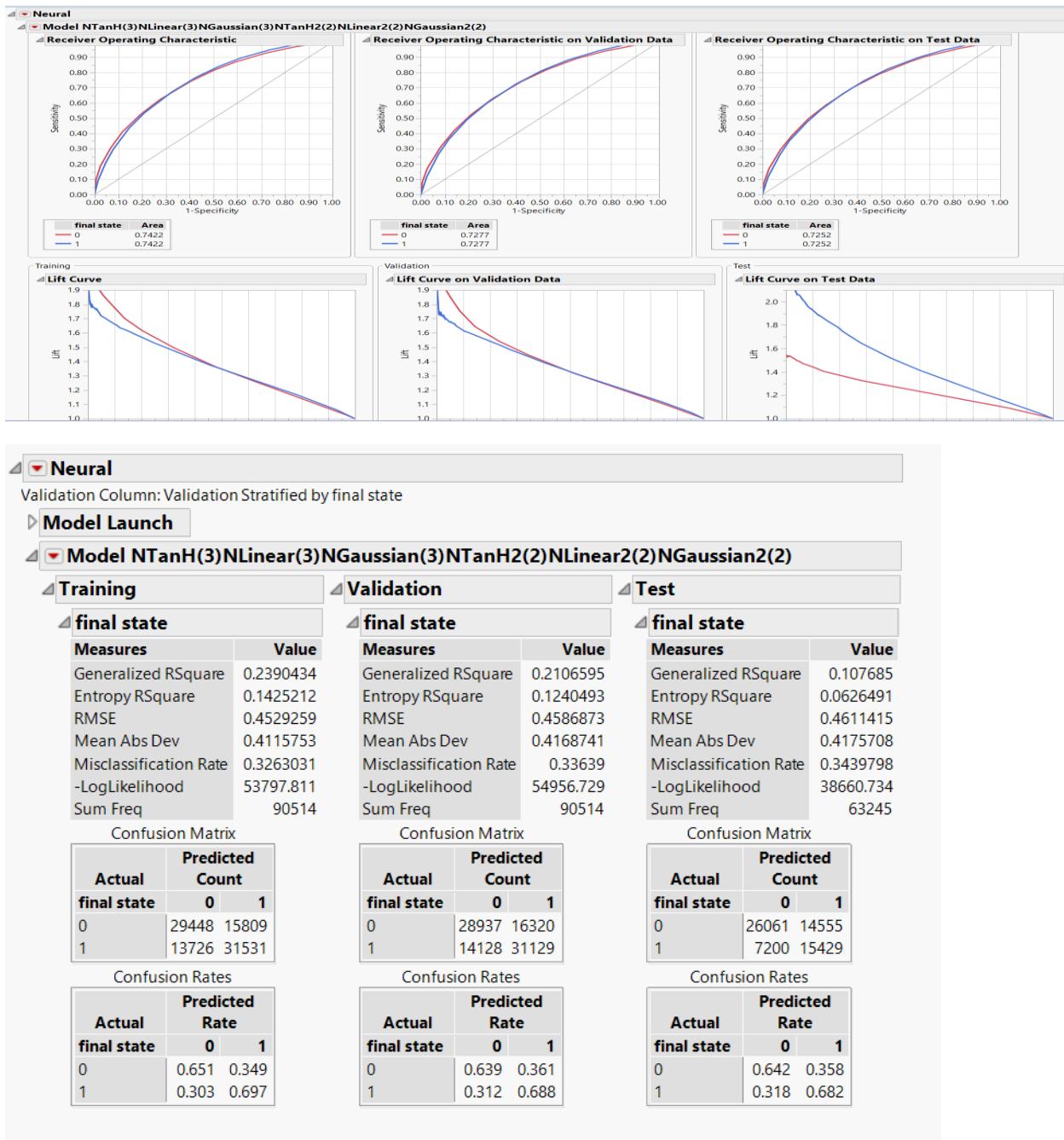
Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	90514	38237	42.2443	0.02757	122019
Validation	90514	37938	41.9140	0.02894	
Test	63245	26507	41.9116	-0.0332	

Training			Validation			Test		
Actual		Predicted Count	Actual		Predicted Count	Actual		Predicted Count
final state		0 1	final state		0 1	final state		0 1
0	26647	18610	0	26731	18526	0	23843	16773
1	19627	25630	1	19412	25845	1	9734	12895

Data set	Number of False positive and False Negative errors	Accuracy
Training	38237	57.75%
Validation	37938	58.06%
Test	26507	58.08%

We are getting very low accuracy rates as compared to other models. Even the number of False Negative errors is also high. We may not consider this model. This is the model where the testing accuracy is more than the training accuracy.

## 4.Neural Nets



Data set	Number of False positive and False Negative errors	Accuracy
Training	29535	67.36%
Validation	30448	66.36%
Test	21755	65.06%

This model has the same accuracy rate as the Decision tree. The neural net returned high accuracy rates versus the decision tree model. However, the testing accuracy is the same as the decision tree. This might be due to overfitting with the training dataset.

## 5. Cluster Analysis

The screenshot shows the 'K Means Cluster - JMP Pro' dialog box. On the left, under 'Select Columns', there is a list of 59 columns. One column, 'Validation Stratified by final state', is highlighted with a blue bar at the bottom. In the center, under 'Cast Selected Columns into Roles', the 'Y, Columns' section contains several optional numeric fields: deadline, launched, deadline year, launched year, Range Scale[goal usd], Range Scale[funding duration], and optional numeric. Below this, 'Weight' and 'Freq' are listed as optional numeric roles, and 'By' is listed as an optional role. On the right, the 'Action' panel includes buttons for OK, Cancel, Remove, Recall, and Help.

## Modelling prediction saved - K Means Cluster - JMP Pro

Iterative Clustering Validation Stratified by final state=Training

Cluster Comparison

Method	NCluster	CCC	Best

Columns Scaled Individually

Control Panel

K Means NCluster=2

Columns Scaled Individually

Cluster Summary

Cluster	Count	Step	Criterion
1	90508	2	0
2	6		

Cluster Means

Cluster	deadline	launched	deadline year	launched year	Range Scale[goal usd]	Range Scale[funding duration]
1	3479910566	3476978842	2013.75535	2013.69192	0.00022401	0.36240546
2	3516698030	3513767980	2014.83333	2014.66667	0.56770833	0.35897436

Cluster Standard Deviations

Cluster	deadline	launched	deadline year	launched year	Range Scale[goal usd]	Range Scale[funding duration]
1	53121482.3	53320173.6	1.6867796	1.70218004	0.00275448	0.14050553
2	12131485.5	12160690.1	0.68718427	0.47140452	0.16397981	0.27587063

**RESULT** - Here, we find two clusters – one with 90,508 and the other with 6 observations. This shows that the model does not cluster successfully. In other words, the model was unsuccessful in forming valid clusters. Also, proper predictions may not be able to do since the original dataset had 45257 projects that were successful.

# REGRESSION MODELING

With Stratified split of Training set – 0.4, Validation set – 0.4, Test set = 0.2

## 1.Nominal Logistic regression model:

The screenshot shows the 'Fit Model' dialog in JMP Pro. In the 'Model Specification' section, under 'Select Columns', 59 columns are listed, including 'final state' as the Y variable. Under 'Personality', 'Nominal Logistic' is selected. In the 'Construct Model Effects' section, various model terms are specified: 'category', 'main\_category', 'currency', 'deadline', 'launched', 'country', 'deadline year', 'launched year', 'Range Scale[goal usd]', and 'Range Scale[funding duration]'. The 'Degree' is set to 2.

### Confusion Matrix

		Training		Validation		Test	
		Predicted		Predicted		Predicted	
Actual	Count	Actual	Count	Actual	Count	Actual	Count
final state		0	1	0	1	0	1
0	29003 16254	28933 16324	25968 14648				
1	14514 30743	14495 30762	7379 15250				

Data set	Number of False positive and False Negative Errors	Accuracy
Training	30768	66.00%
Validation	30819	65.95%
Test	22027	65.17%

We are getting low accuracy rate compared to classification models (decision tree). Even the number of False Negative error is also high compared to other models. This model results can be used to build Ensemble models.

Modelling prediction saved - Fit Nominal Logistic - JMP Pro

### Nominal Logistic Fit for final state

#### Effect Summary

Source	LogWorth	PValue
category	947.004	0.00000
Range Scale[goal usd]	279.250	0.00000
deadline	81.280	0.00000
launched	81.201	0.00000
Range Scale[funding duration]	75.647	0.00000
country	24.752	0.00000
currency	13.609	0.00000
main_category	5.042	0.00001
launched year	1.239	0.05764
deadline year	0.592	0.25611

[Remove](#) [Add](#) [Edit](#)  FDR

Converged in Gradient, 24 iterations

#### Iterations

#### Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	7238.567	196	14477.13	<.0001*
Full	55500.957			
Reduced	62739.524			

RSquare (U)	0.1154
AICc	111425
BIC	113410
Observations (or Sum Wgts)	90514

#### Fit Details

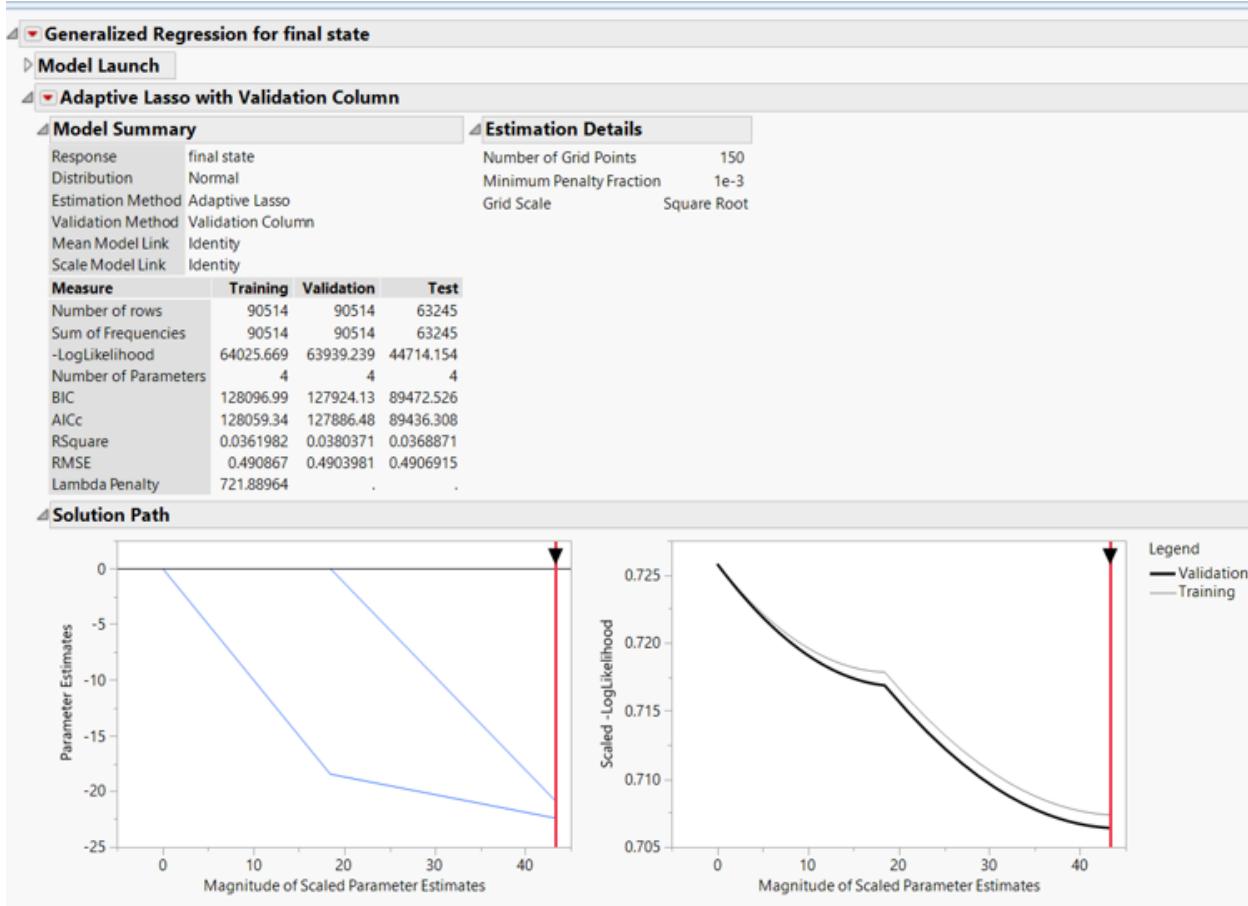
#### Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare	Prob>ChiSq
Lack Of Fit	90313	55499.571	110999.1	
Saturated	90509	1.386		Prob>ChiSq
Fitted	196	55500.957		<.0001*

#### Parameter Estimates

Term		Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	Biased	41.2723972	24201.187	0.00	0.9986
category[3D Printing]	Biased	-0.3417946	0.1865425	3.36	0.0669
category[Academic]	Biased	0.8685517	0.1860089	21.80	<.0001*
category[Accessories]	Biased	0.18734392	0.1096951	2.92	0.0877
category[Action]	Biased	1.82315007	0.8129966	5.03	0.0249*
category[Animals]	Biased	0.85156719	0.2687636	10.04	0.0015*

## 2.Generalized Regression:



We are getting very low R Square value versus other models because we are not using some of the nominal variables.

### 3. Step Wise regression:

Crossvalidation													
Source	RSquare	RASE	Freq										
Training Set	0.1242	0.46793	90514										
Validation Set	0.1242	0.46791	90514										
Test Set	0.0388	0.46996	63245										

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC	RSquare Validation	RMSE Validation	RSquare Test	RMSE Test
19818.799	90458	0.4680747	0.1242	0.1236	20.108722	56	119502.4	120038.8	0.1242	0.467911	0.0388	0.469964

Summary of Fit													
RSquare													0.12417
RSquare Adj													0.123637
Root Mean Square Error													0.468074
Mean of Response													0.5
Observations (or Sum Wgts)													90514

Analysis of Variance													
Source	DF		Sum of Squares	Mean Square	F Ratio								
Model	55		2809.776	51.0868	233.1741								
Error	90458		19818.724	0.2191	Prob > F								
C. Total	90513		22628.500		<.0001*								

The Stepwise model is getting the same R square as the Standard Least Squares. We can consider this model for creating ensemble models.

## 4.Standard Least Squares:

Modelling prediction saved - Fit Least Squares - JMP Pro

**Response final state**

Validation: Validation Stratified by final state

**Singularity Details**

Term	Details
Intercept	=category[3D Printing] + category[Academic] + category[Accessories] + ...
currency[CHF]	=currency[HKD] + country[CH] - country[HK]=currency[MXN] + country[...]

**Effect Summary**

Source	LogWorth	PValue
category	768.139	0.00000
deadline	79.428	0.00000
launched	79.319	0.00000
Range Scale[funding duration]	72.895	0.00000
country	42.859	0.00000
Range Scale[goal usd]	10.857	0.00000
main_category	4.600	0.00003
launched year	1.245	0.05688
deadline year	0.364	0.43219
currency	0.001	0.99860

[Remove](#) [Add](#) [Edit](#)  FDR

**Summary of Fit**

RSquare	0.125188
RSquare Adj	0.123289
Root Mean Square Error	0.468167
Mean of Response	0.5
Observations (or Sum Wgts)	90514

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	196	2832.814	14.4531	65.9418
Error	90317	19795.686	0.2192	<b>Prob &gt; F</b>
C. Total	90513	22628.500		<.0001*

**Parameter Estimates**

Term		Estimate	Std Error	t Ratio	Prob> t
Intercept	Biased	-12.22258	12.10534	-1.01	0.3126
category[3D Printing]	Biased	0.0294159	0.040431	0.73	0.4669
category[Academic]	Biased	-0.179375	0.038305	-4.68	<.0001*
category[Accessories]	Biased	-0.023833	0.023521	-1.01	0.3109
category[Action]	Biased	-0.25418	0.162243	-1.57	0.1172
category[Animals]	Biased	-0.164266	0.058739	-2.80	0.0052*
category[Animation]	Biased	-0.154749	0.158539	-0.98	0.3290

Effect Tests						
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
category	157	148	933.86711	28.7887	<.0001*	LostDFs
main_category	14	6	6.79999	5.1708	<.0001*	LostDFs
currency	12	5	0.05299	0.0484	0.9986	LostDFs
deadline	1	1	78.93857	360.1540	<.0001*	
launched	1	1	78.82850	359.6518	<.0001*	
deadline year	1	1	0.13522	0.6169	0.4322	
launched year	1	1	0.79477	3.6261	0.0569	
Range Scale[goal usd]	1	1	10.01585	45.6969	<.0001*	
Range Scale[funding duration]	1	1	72.33856	330.0417	<.0001*	
country	21	15	53.74880	16.3484	<.0001*	LostDFs

Crossvalidation			
Source	RSquare	RASE	Freq
Training Set	0.1252	0.46766	90514
Validation Set	0.1232	0.46818	90514
Test Set	0.0382	0.47010	63245

Model	Training R square	Validation R square	Test R square
Nominal Logistic Regression	0.1033	0.1024	0.1154
Generalized Regression	0.031982	0.0380371	0.0368871
Stepwise	0.1242	0.1242	0.0388
Standard Least Squares	0.1252	0.1232	0.0382

**RESULT –** We can conclude that the R square value is greater for Nominal Logistic Regression Model as compared to R square values of other three models. Thus, Nominal Logistic Regression Model is a better model than the other three models.

## Ensemble Models

With Stratified split of Training set – 0.4, Validation set – 0.4, Test set = 0.2

### 1. Decision Tree and Boosted tree Ensemble Model

Modelling prediction saved - Boosted Tree of final state - JMP Pro

**Boosted Tree for final state**

**Specifications**

Target Column:	final state	Number of training rows:	90514
Validation Column:	Validation Stratified by final state	Number of validation rows:	90514
Number of Layers:	50	Number of test rows:	63245
Splits per Tree:	3		
Learning Rate:	0.1		
Overfit Penalty:	0.0001		

**Overall Statistics**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1121	0.1098	0.0505	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.1919	0.1882	0.0875	$(1 - (L(0)/L(\text{model}))^{(2/n)})/(1 - L(0)^{(2/n)})$
Mean -Log p	0.6155	0.6171	0.6192	$\sum -\text{Log}(p[j])/n$
RMSE	0.4622	0.4629	0.4643	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4429	0.4438	0.4442	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3342	0.3376	0.3473	$\sum (p[j] \neq p_{\text{Max}})/n$
N	90514	90514	63245	n

**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted	Actual	Predicted	Actual	Predicted
final state	Count	final state	Count	final state	Count
final state	0 1	final state	0 1	final state	0 1
0	28616 16641	0	28476 16781	0	25603 15013
1	13613 31644	1	13773 31484	1	6954 15675

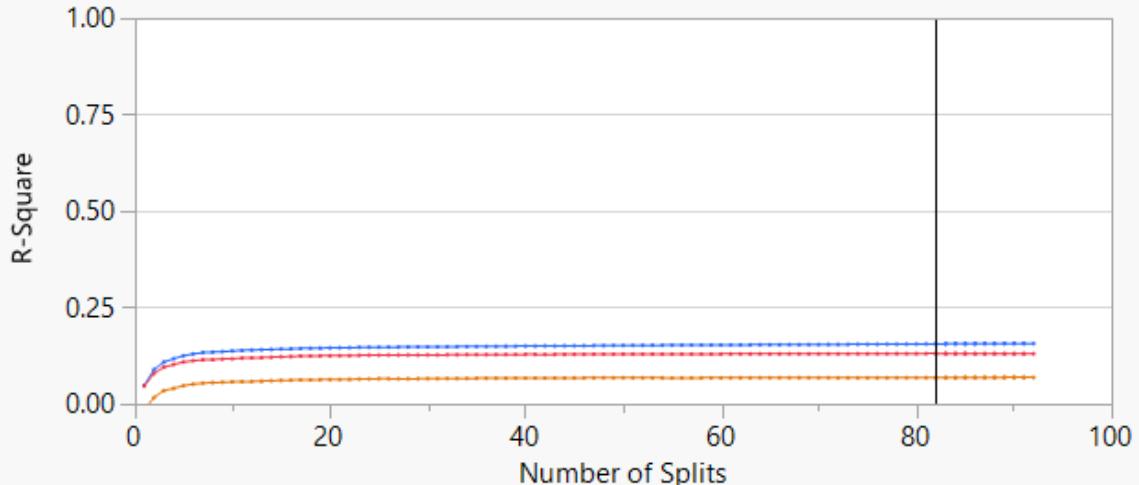
## 2. Neural and decision Tree Ensemble Model

### Confusion Matrix

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state	0 1	final state	0 1	final state	0 1
0	29070 16187	0	28625 16632	0	25549 15067
1	12371 32886	1	13014 32243	1	6600 16029

Split	Prune	Go	RSquare	N	Number of Splits
			Training	0.155	90514
			Validation	0.130	90514
			Test	0.068	63245

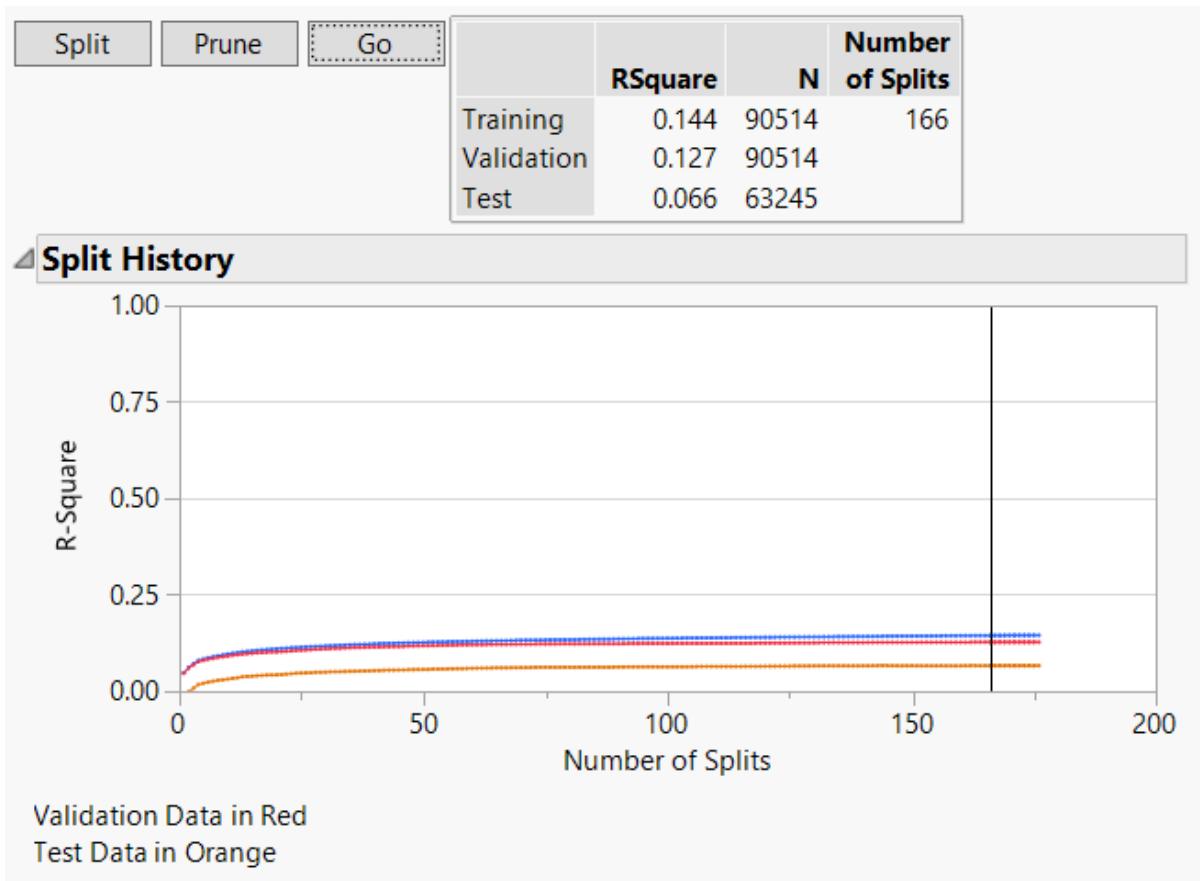
### Split History



Validation Data in Red

Test Data in Orange

### 3. Nominal Logistic and decision tree Ensemble Model

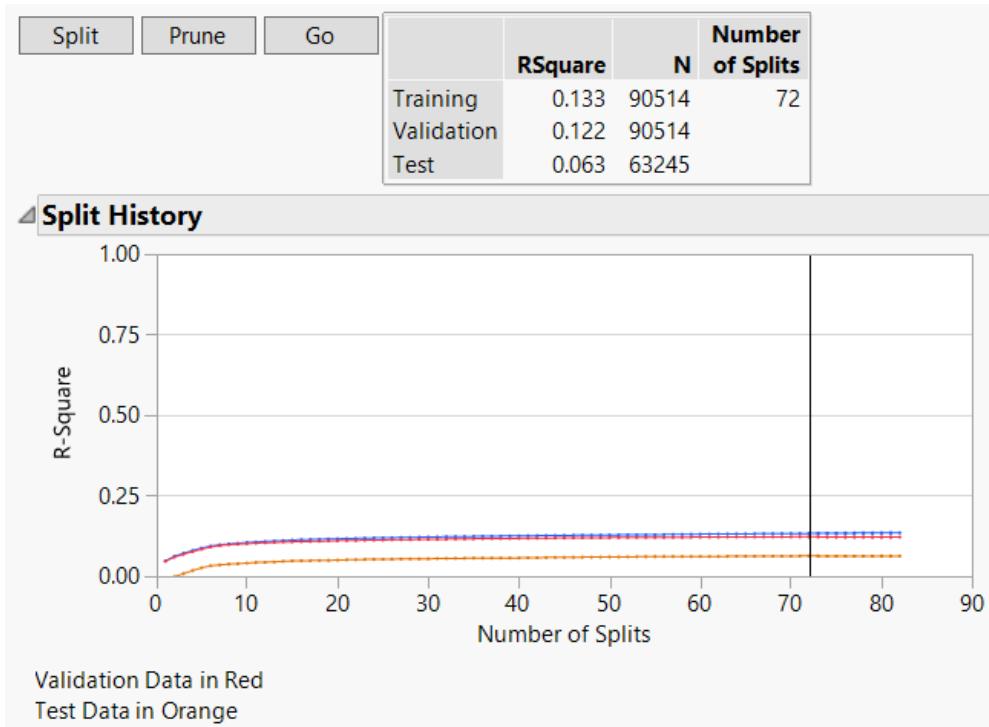


Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1445	0.1271	0.0659	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2420	0.2154	0.1131	$(1 - (L(0)/L(\text{model}))^{(2/n)})/(1 - L(0)^{(2/n)})$
Mean -Log p	0.5930	0.6050	0.6091	$\sum -\text{Log}(p[j])/n$
RMSE	0.4520	0.4575	0.4603	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4088	0.4140	0.4154	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3200	0.3316	0.3347	$\sum (p[j] \neq p_{\text{Max}})/n$
N	90514	90514	63245	n

**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state	0 1	final state	0 1	final state	0 1
0	30566 14691	0	30111 15146	0	26991 13625
1	14269 30988	1	14869 30388	1	7543 15086

#### 4. Step wise and decision Tree Ensemble Model



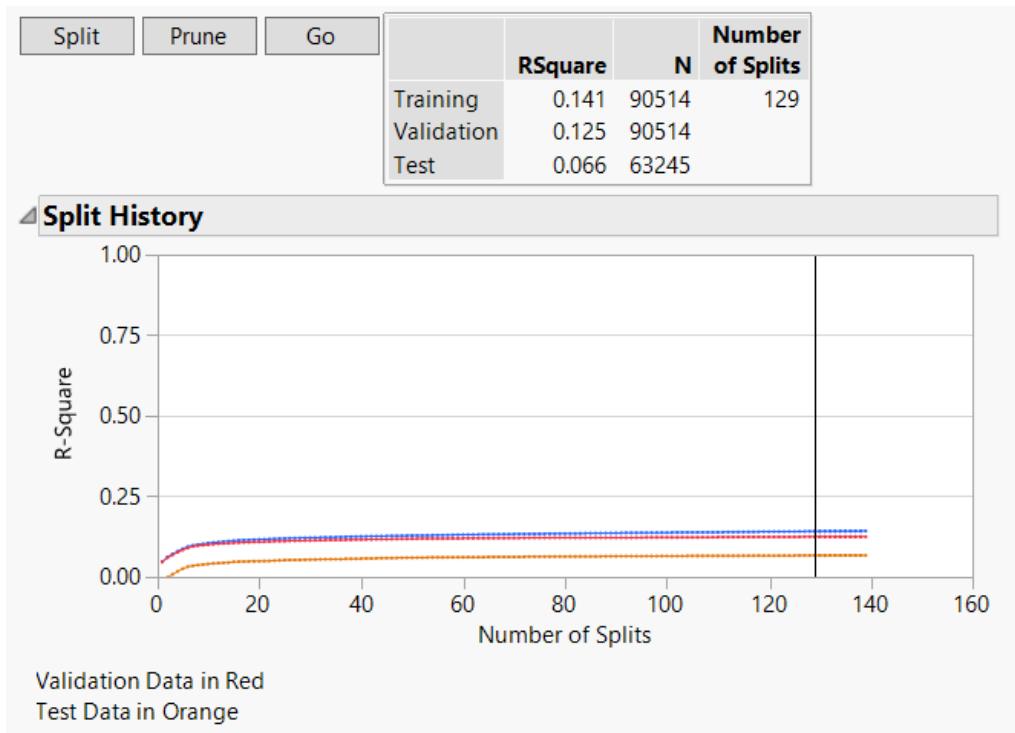
**Fit Details**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1329	0.1223	0.0628	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2243	0.2079	0.1079	$(1 - (L(0)/L(\text{model}))^{(2/n)}) / (1 - L(0)^{(2/n)})$
Mean -Log p	0.6011	0.6084	0.6112	$\sum -\text{Log}(\rho[j]) / n$
RMSE	0.4556	0.4591	0.4611	$\sqrt{\sum (y[j] - \rho[j])^2 / n}$
Mean Abs Dev	0.4153	0.4190	0.4196	$\sum  y[j] - \rho[j]  / n$
Misclassification Rate	0.3291	0.3355	0.3464	$\sum (\rho[j] \neq \rho_{\text{Max}}) / n$
N	90514	90514	63245	n

**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state	0 1	final state	0 1	final state	0 1
0	28539 16718	0	28299 16958	0	25497 15119
1	13073 32184	1	13410 31847	1	6787 15842

## 5.Standard Least Square and decision tree Ensemble Model



**Fit Details**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1414	0.1247	0.0664	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.2374	0.2117	0.1138	$(1 - (L(0)/L(\text{model}))^{(2/n)}) / (1 - L(0)^{(2/n)})$
Mean -Log p	0.5951	0.6067	0.6089	$\sum -\text{Log}(p[j])/n$
RMSE	0.4530	0.4582	0.4601	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4105	0.4158	0.4162	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3236	0.3332	0.3323	$\sum (p[j] \neq p_{\text{Max}})/n$
N	90514	90514	63245	n

**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted	Actual	Predicted	Actual	Predicted
final state	Count	final state	Count	final state	Count
0	31005	14252	0	30559	14698
1	15042	30215	1	15464	29793

## 6. Discriminant Analysis and Decision Tree Ensemble Model

### Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1383	0.1238	0.0623	$1 - \text{Loglike}(\text{model}) / \text{Loglike}(0)$
Generalized RSquare	0.2326	0.2102	0.1070	$(1 - (L(0) / L(\text{model}))^{(2/n)}) / (1 - L(0)^{(2/n)})$
Mean -Log p	0.5973	0.6074	0.6115	$\sum -\text{Log}(\rho[j]) / n$
RMSE	0.4539	0.4586	0.4613	$\sqrt{\sum (y[j] - \rho[j])^2 / n}$
Mean Abs Dev	0.4122	0.4170	0.4181	$\sum  y[j] - \rho[j]  / n$
Misclassification Rate	0.3247	0.3354	0.3434	$\sum (\rho[j] \neq \rho_{\text{Max}}) / n$
N	90514	90514	63245	n

### Confusion Matrix

Training		Validation		Test	
Actual	Predicted	Actual	Predicted	Actual	Predicted
final state	Count	final state	Count	final state	Count
final state	0 1	final state	0 1	final state	0 1
0	29488 15769	0	29011 16246	0	26095 14521
1	13622 31635	1	14111 31146	1	7197 15432

Split	Prune	Go		RSquare	N	Number of Splits
			Training	0.138	90514	124
			Validation	0.124	90514	
			Test	0.062	63245	

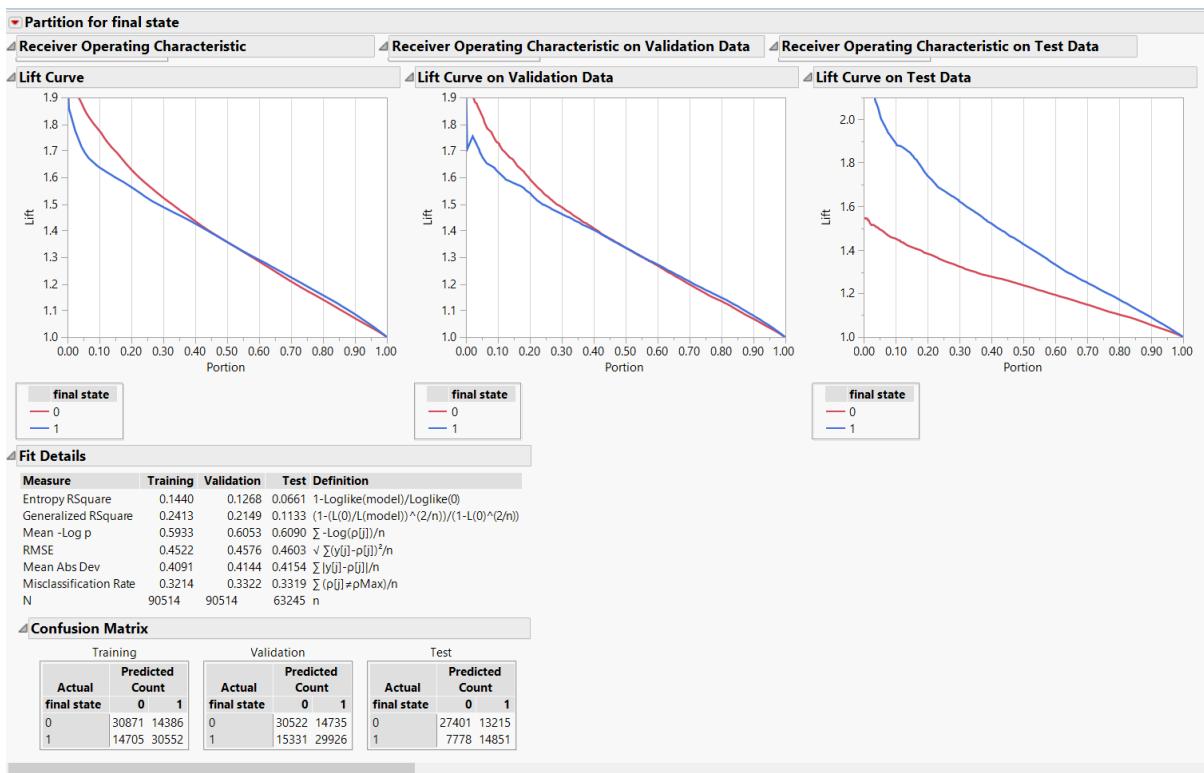
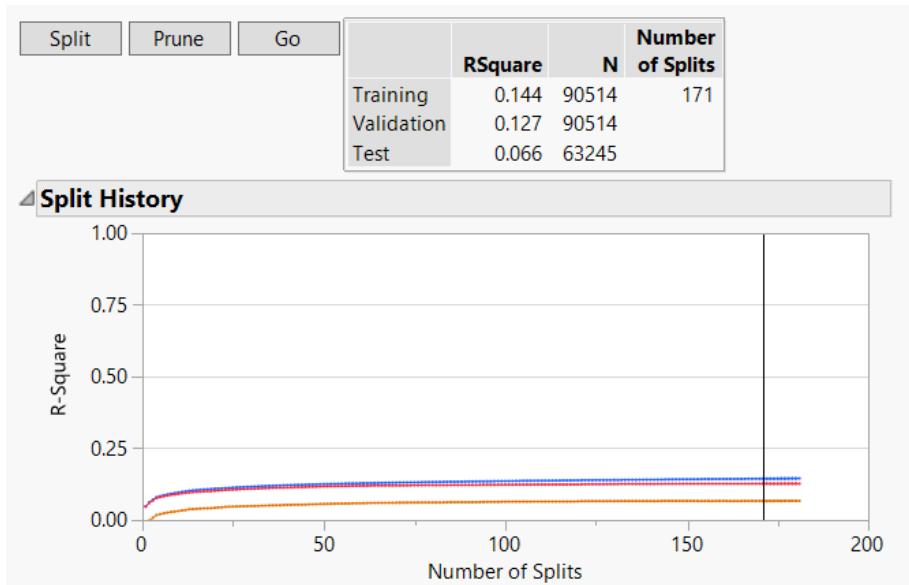
## 7. Discriminant Analysis and Neural Nets Ensemble Model

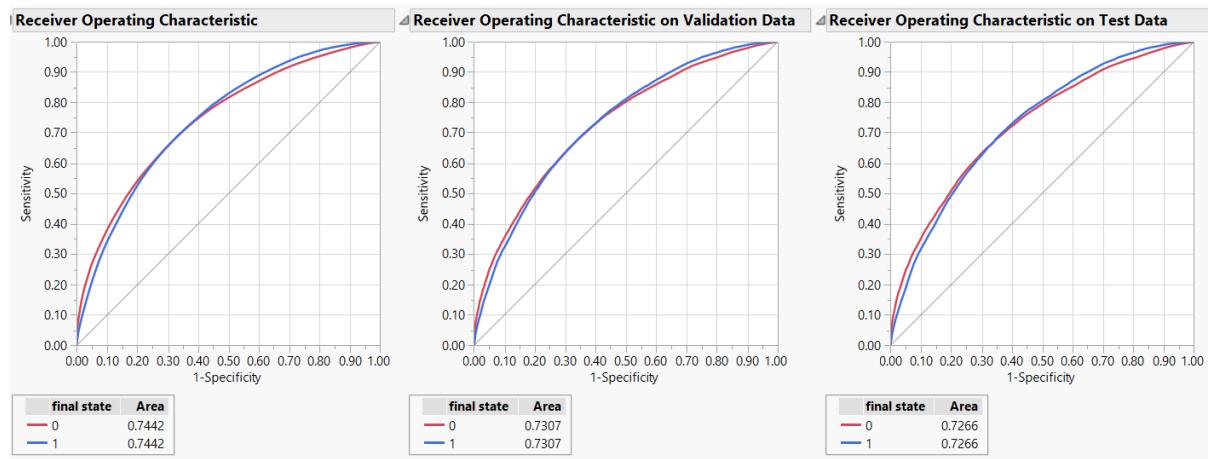
Model NTanH(2)NLinear(2)NGaussian(2)NTanH2(1)NLinear2(1)NGaussian2(1)																																
Training	Validation	Test																														
<b>final state</b>	<b>final state</b>	<b>final state</b>																														
<b>Measures</b>	<b>Measures</b>	<b>Measures</b>																														
Generalized RSquare	0.1921176	Generalized RSquare	0.1858493																													
Entropy RSquare	0.112233	Entropy RSquare	0.1082817																													
RMSE	0.4622009	RMSE	0.4636056																													
Mean Abs Dev	0.4295484	Mean Abs Dev	0.4310439																													
Misclassification Rate	0.3448638	Misclassification Rate	0.3481671																													
-LogLikelihood	55698.081	-LogLikelihood	55945.979																													
Sum Freq	90514	Sum Freq	90514																													
Confusion Matrix	Confusion Matrix	Confusion Matrix																														
<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Count</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>29138 16119</td></tr> <tr> <td>1</td><td>15096 30161</td></tr> </tbody> </table>	Predicted		Actual	Count	final state	0 1	0	29138 16119	1	15096 30161	<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Count</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>28962 16295</td></tr> <tr> <td>1</td><td>15219 30038</td></tr> </tbody> </table>	Predicted		Actual	Count	final state	0 1	0	28962 16295	1	15219 30038	<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Count</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>26163 14453</td></tr> <tr> <td>1</td><td>7616 15013</td></tr> </tbody> </table>	Predicted		Actual	Count	final state	0 1	0	26163 14453	1	7616 15013
Predicted																																
Actual	Count																															
final state	0 1																															
0	29138 16119																															
1	15096 30161																															
Predicted																																
Actual	Count																															
final state	0 1																															
0	28962 16295																															
1	15219 30038																															
Predicted																																
Actual	Count																															
final state	0 1																															
0	26163 14453																															
1	7616 15013																															
Confusion Rates	Confusion Rates	Confusion Rates																														
<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Rate</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>0.644 0.356</td></tr> <tr> <td>1</td><td>0.334 0.666</td></tr> </tbody> </table>	Predicted		Actual	Rate	final state	0 1	0	0.644 0.356	1	0.334 0.666	<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Rate</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>0.640 0.360</td></tr> <tr> <td>1</td><td>0.336 0.664</td></tr> </tbody> </table>	Predicted		Actual	Rate	final state	0 1	0	0.640 0.360	1	0.336 0.664	<table border="1"> <thead> <tr> <th colspan="2">Predicted</th> </tr> <tr> <th>Actual</th><th>Rate</th></tr> </thead> <tbody> <tr> <td>final state</td><td>0 1</td></tr> <tr> <td>0</td><td>0.644 0.356</td></tr> <tr> <td>1</td><td>0.337 0.663</td></tr> </tbody> </table>	Predicted		Actual	Rate	final state	0 1	0	0.644 0.356	1	0.337 0.663
Predicted																																
Actual	Rate																															
final state	0 1																															
0	0.644 0.356																															
1	0.334 0.666																															
Predicted																																
Actual	Rate																															
final state	0 1																															
0	0.640 0.360																															
1	0.336 0.664																															
Predicted																																
Actual	Rate																															
final state	0 1																															
0	0.644 0.356																															
1	0.337 0.663																															

## 8. Generalized Regression and Stepwise Ensemble Model

Crossvalidation			
Source	RSquare	RASE	Freq
Training Set	0.1242	0.46793	90514
Validation Set	0.1242	0.46791	90514
Test Set	0.0387	0.46998	63245

## 9. Generalized Regression and Decision Tree Ensemble Model





## 10. Generalized Regression and Standard Least Squares Ensemble Model

**Response final state**  
Validation: Validation Stratified by final state

**Singularity Details**

Term	Details
Intercept	=category[3D Printing] + category[Academic] + category[Accessories] + ...

**Effect Summary**

Source	LogWorth	PValue
category	770.026	0.00000
deadline	79.559	0.00000
launched	79.446	0.00000
currency	20.956	0.00000
Range Scale[goal usd]	10.784	0.00000
main_category	4.562	0.00003
launched.year	1.349	0.04480
deadline.year	0.399	0.39925
Prediction Generalized regreesion	.	.
Range Scale[funding duration]	.	.

[Remove](#) [Add](#) [Edit](#)  FDR

**Summary of Fit**

RSquare	0.122813
RSquare Adj	0.121055
Root Mean Square Error	0.468763
Mean of Response	0.5
Observations (or Sum Wgts)	90514

**Analysis of Variance**

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	181	2779.065	15.3540	69.8737
Error	90332	19849.435	0.2197	<b>Prob &gt; F</b>
C. Total	90513	22628.500		<.0001*

## COMPARISON OF MODELS

Models	False Negative for Test	False Positive for Test	R Square Training	Accuracy Training	R Square validation	Accuracy validation	R Square Test	Accuracy Test
<b>Decision Tree (DT)</b>	16327	13259	0.2320	67.3	0.2104	66.5	0.1106	65.5
<b>Boosted Tree (BT)</b>	16641	13613	0.1919	66.57	0.1882	66.24	0.0875	65.26
<b>Discriminant Analysis (DA)</b>	18610	19627		57.75		58.06		58.08
<b>Neural Nets (NN)</b>	15809	13726	0.239	67.36	0.21065	66.36	0.1076	65.06
<b>Nominal logistic regression (NL)</b>	-	-	0.1154	66.00		65.95		65.17
<b>Generalized regression</b>	-	-	0.036198	-	0.03803	-	0.03688	-
<b>Step wise (SW)</b>	-	-	0.1242	-	0.1242	-	0.0388	-
<b>Standard Least Squares (SL)</b>	-	-	0.1252	-	0.1232	-	0.0382	-
<b>Ensemble 01 DT vs BT</b>	28616	16641	0.1919	66.5	0.1882	66.24	0.0875	65.26
<b>Ensemble 02 NN vs DT</b>	16187	12371	0.155	68.004	0.130	67.2	0.068	65.7
<b>Ensemble 03 NL vs DT</b>	14691	14269	0.2420	68.004	0.2154	66.839	0.1131	66.53
<b>Ensemble 04 SW vs DT</b>	16718	13073	0.133	67.08	0.122	66.44	0.063	65.3
<b>Ensemble 05 SL vs DT</b>	14252	15042	0.141	67.63	0.125	66.67	0.066	66.77
<b>Ensemble 06 DA vs DT</b>	14521	7197	0.2326	67.5	0.2102	66.4	0.1070	65.66
<b>Ensemble 07 DA vs NN</b>	14453	7616	0.192	65.51	0.1858	65.183	0.085006	65.10
<b>Ensemble 08 GR vs SW</b>	-	-	0.124	-	0.124	-	0.0387	-
<b>Ensemble 09 GR vs DT</b>	13215	7778	0.2413	67.86	0.2149	66.78	0.1133	65.5
<b>Ensemble 10 GR vs SL</b>	-	-	0.1228	-	0.1214	-	0.0371	-

\*Since Regression models consider only continuous target variables, only R square values can be compared.

## COST OF FALSE POSITIVE AND FALSE NEGATIVE ERROR COMPARISON FOR TEST DATASET

<b>Models</b>	<b>False Negative</b>	<b>False Positive</b>	<b>Total Cost of Error</b>	<b>R Square for Test dataset</b>	<b>Accuracy (%)</b>
<b>Decision Tree (DT)</b>	6855	14923	\$1159998.31	0.1106	65.5
<b>Boosted Tree (BT)</b>	6954	15013	\$1168175.62	0.0875	65.26
<b>Discriminant Analysis (DA)</b>	9734	16773	\$1345381.02		58.08
<b>Neural Nets (NN)</b>	7200	15429	\$1201637.28	0.1076	65.06
<b>Nominal logistic regression (NL)</b>	-	-	-		65.17
<b>Generalized regression</b>	-	-	-	0.03688	-
<b>Step wise (SW)</b>	-	-	-	0.0388	-
<b>Standard Least Squares (SL)</b>	-	-	-	0.0382	-
<b>Ensemble 01 DT vs BT</b>	6954	15013	\$1168175.62	0.0875	65.26
<b>Ensemble 02 NN vs DT</b>	6600	15067	\$1164611.44	0.068	65.7
<b>Ensemble 03 NL vs DT</b>	7543	13625	\$1085416.07	0.1131	66.53
<b>Ensemble 04 SW vs DT</b>	6787	15119	\$1171995.71	0.063	65.3
<b>Ensemble 05 SL vs DT</b>	7858	13156	\$1059828.34	0.066	66.77
<b>Ensemble 06 DA vs DT</b>	14521	7197	\$789234.33	0.1070	65.66
<b>Ensemble 07 DA vs NN</b>	14453	7616	\$816467.09	0.085006	65.10
<b>Ensemble 08 GR vs SW</b>	-	-	-	0.0387	-
<b>Ensemble 09 GR vs DT</b>	13215	7778	\$802168.31	0.1133	65.5
<b>Ensemble 10 GR vs SL</b>	-	-	-	0.0371	-

## PROPOSED MODEL – DECISION TREE MODEL

Split proportion: Training set – 0.4 Validation set – 0.4 Test set – 0.2

Cut-off value	Training False negative	Training False Positive	Training Accuracy	Validation False negative	Validation False Positive	Validation accuracy	Test False negative	Test False Positive	Test Accuracy (%)
0.5	16327	13259	67.3	16672	13635	66.51	14923	6855	65.56
0.4	8710	22221	65.82	9000	22578	65.11	8185	11383	69.06
0.3	32870	2479	60.94	33194	2673	60.37	29832	1331	49.27
0.6	8710	22221	65.82	9000	22578	65.11	8185	11383	69.06

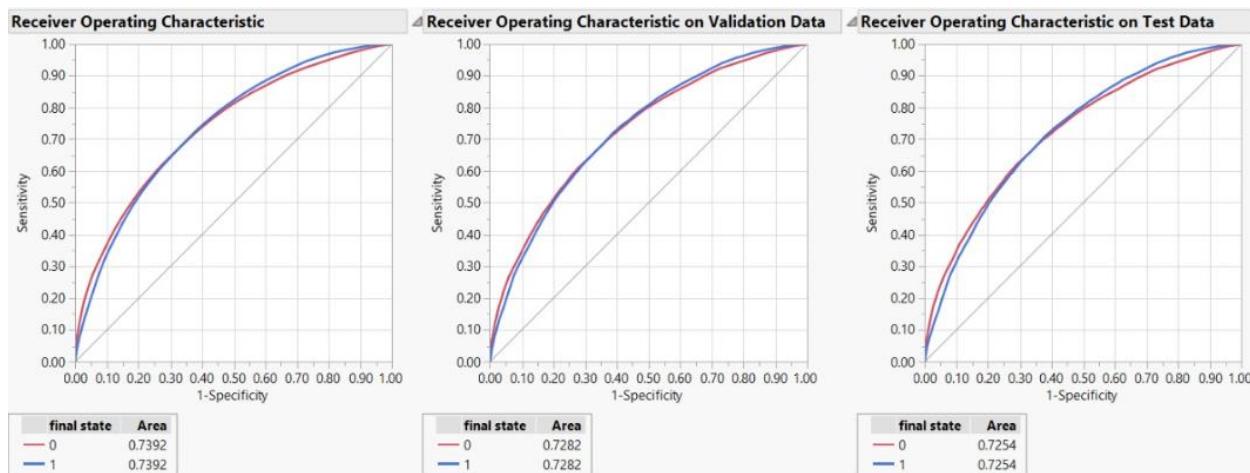
## CONCLUSION

### CASE 01:(Predicting Success before the start of funding date)

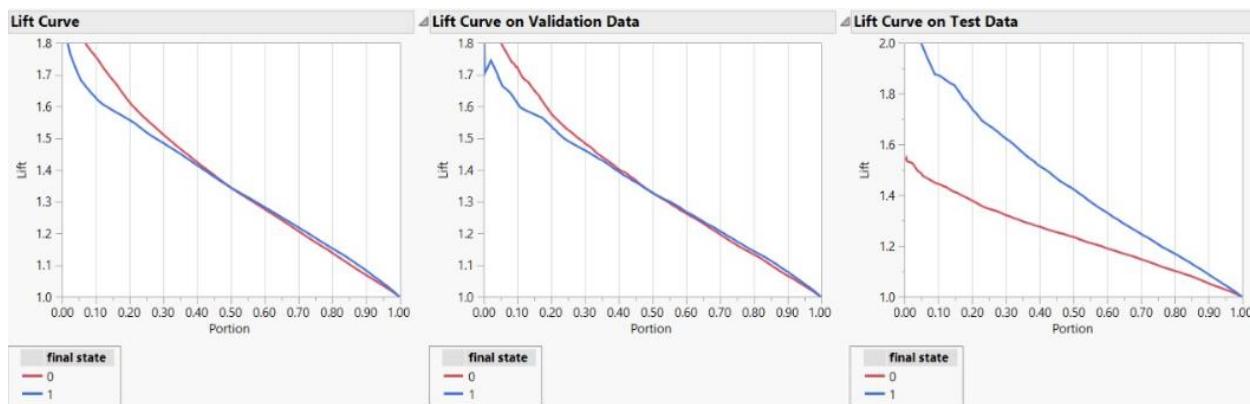
#### Cost of Errors:

Considering Cost of False Negative (Actual - Successful, predicted – Unsuccessful) as \$20.49 and cost of False Positive (Actual - Unsuccessful, Predicted – Successful) as \$68.32

Cut-off value	Test False Negative	Test False Positive	Total cost of error	Accuracy (%)
0.5	14923	6855	\$774195.4	65.56
0.4	8185	11383	\$945397.21	69.06
0.3	29832	1331	\$702191.6	49.27
0.6	8185	11383	\$945397.21	69.06



We are getting Area under curve as 0.72 for the test data set and even same for the validation set.



**CONCLUSION** - Since the estimated cost of error of each type is an approximate value, and it varies from project to project. Thus, it is not right to choose the best model solely on the basis of least cost of error. We should also consider accuracy along with cost of each type of error.

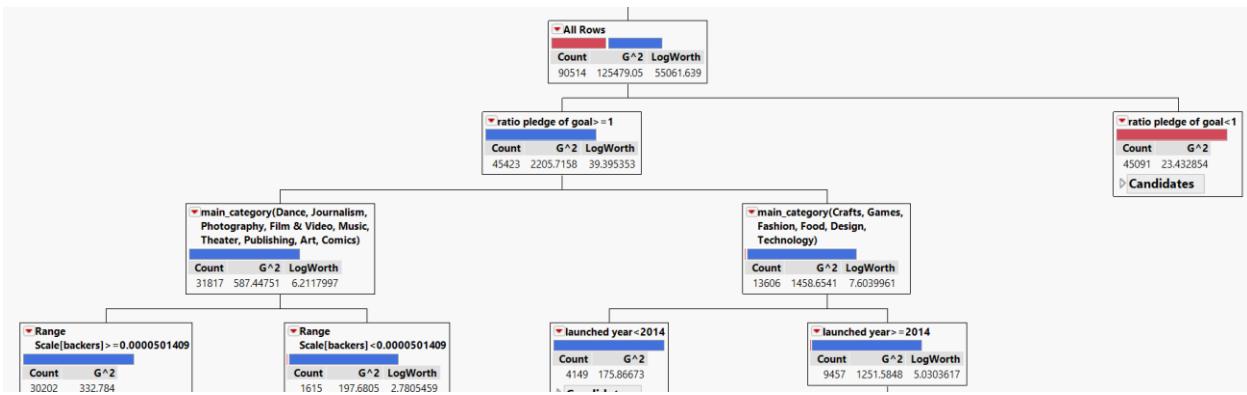
Since, the number of False positive errors are less and good accuracy of 65.56% for cut-off value 0.5, thus this model performs better than other models and we propose Decision tree model with cut-off value of 0.5 as our final model for prediction.

## **CASE 02: (Predicting Success just before the deadline of the project by considering pledged USD and number of backers as Explanatory variables)**

We are suggesting the person to invest in a project on the day before deadline of the project which helps model to make use of the pledged USD and number of backers.

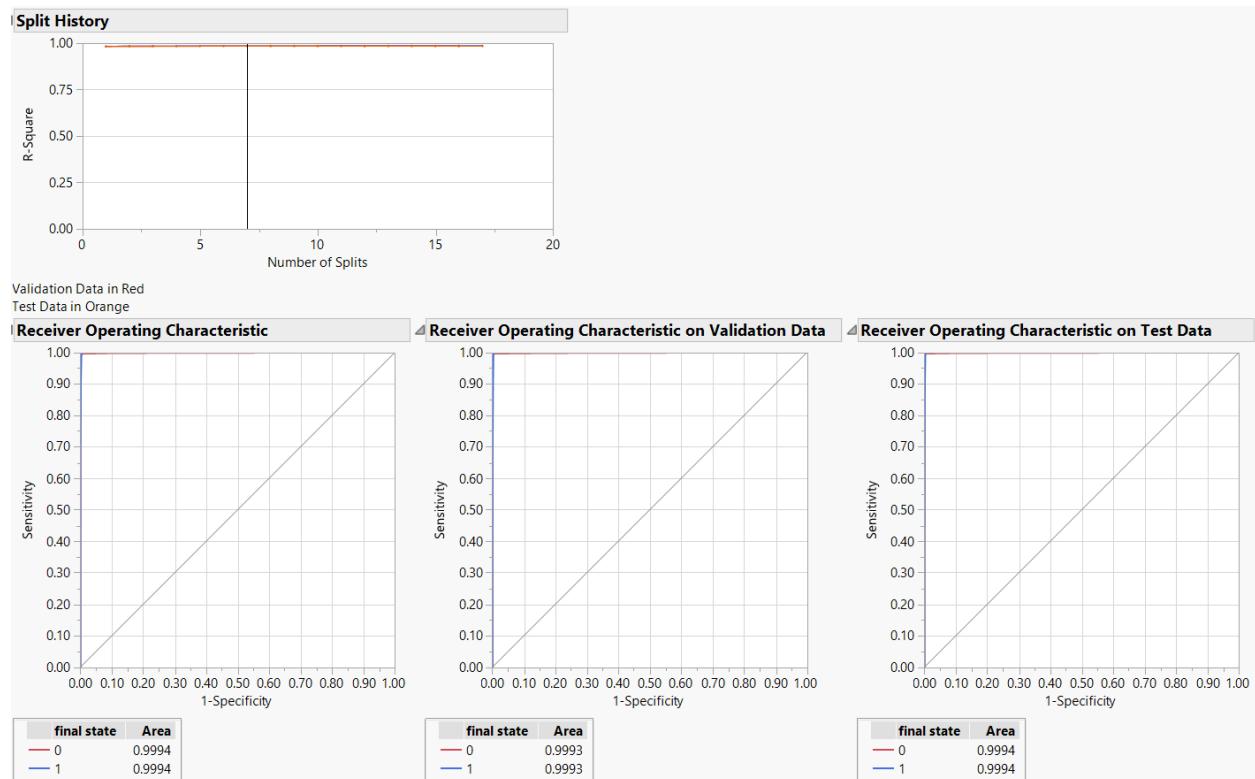
Considering the below variables for building decision tree model.

The screenshot shows the JMP Pro software interface for recursive partitioning. The 'Select Columns' section displays a list of 70 columns, with 'final state' highlighted. The 'Cast Selected Columns into Roles' section shows 'Y, Response' set to 'final state optional'. The 'X, Factor' section lists various columns including category, main\_category, currency, deadline, launched, country, deadline year, launched year, Range Scale[backers], Range Scale[goal usd], Range Scale[pledged usd], Range Scale[funding duration], ratio pledge of goal, ratio pledged by backers, and Weight. The 'Action' panel contains OK, Cancel, Remove, Recall, and Help buttons. The 'Options' section at the bottom includes Method (Decision Tree), Validation Portion (0), and two checked checkboxes: Informative Missing and Ordinal Restricts Order.



The ratio of pledge to goal does the major split for the success rate of the project.

The next major factor in deciding the success is main category of the project.



If we see area under the curve, we are getting 0.9994 sq. units in all the training, validation and testing datasets. We can conclude that the model is performing well with more accuracy.

## Confusion Matrix

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state		final state		final state	
0	45090 167	0	45093 164	0	40478 138
1	1 45256	1	3 45254	1	1 22628

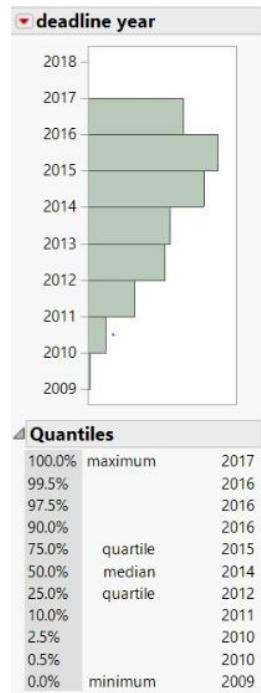
Also, we can see that there are less number of each type of errors, in total there are 139 errors in test dataset.

Overall, we obtain 99.765% accuracy. So, if someone is going to invest in starter project just before funding duration then they can be 99.765% accurate in their prediction.

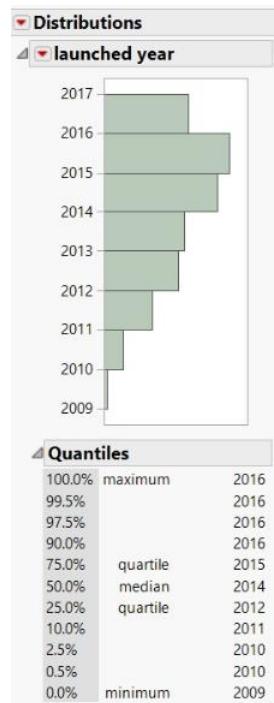
## APPENDIX

### APPENDIX 1.1 - OUTLIER ANALYSIS DISTRIBUTION FOR NEW VARIABLES

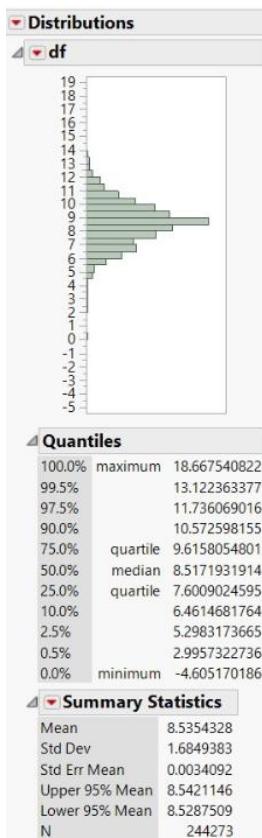
#### Deadline Distribution



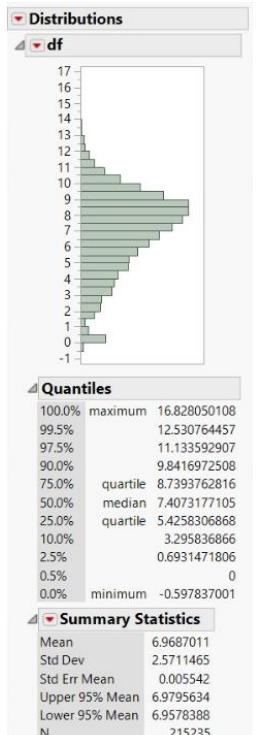
Launched year Distribution - Launched year with 1970 was deleted.



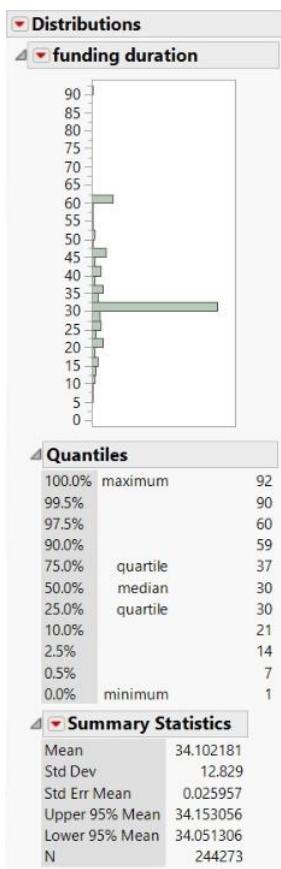
## Goal\_USD Distribution



## Pledged\_USD



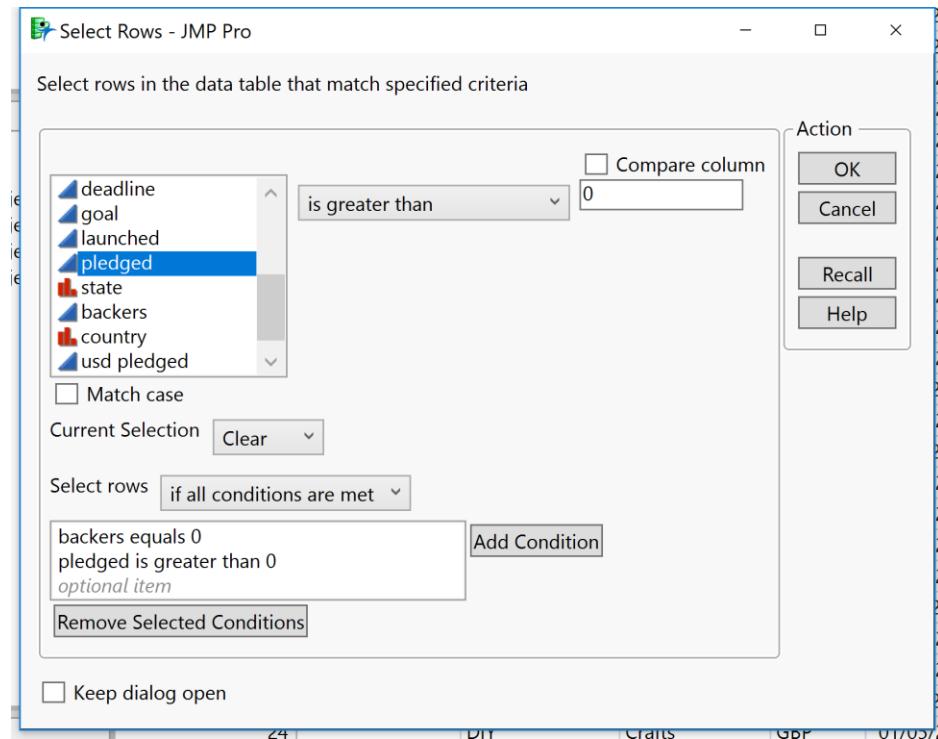
## FUNDING DURATION



## APPENDIX 1.2

### Step 1

Delete the inconsistency – Backers = 0, Pledged amount is having some value



File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

ks-projects-201612-data... Source

Columns (17/1)

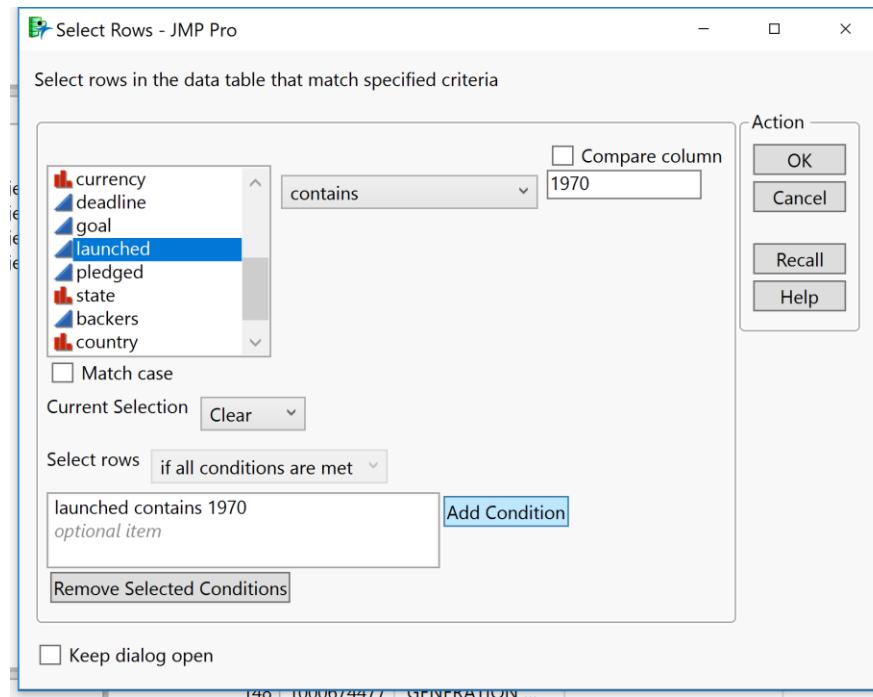
ID	name	description of project 1	description of project 2	description of project 3	description of project 4	category	main_category	currency
1	1000002330	The Songs of ...				Poetry	Publishing	GBP
2	1000004038	Where is Hank?				Narrative Film	Film & Video	USD
3	1000007540	ToshiCapital ...				Music	Music	USD
4	1000011046	Community Film ...				Film & Video	Film & Video	USD
5	1000014025	Monarch ...				Restaurants	Food	USD
6	1000023410	Support Solar ...				Food	Food	USD
7	1000030581	Chaser Strips ...				Drinks	Food	USD
8	1000034518	SPIN - Premium ...				Product Design	Design	USD
9	1000041951	STUDIO IN THE ...				Documentary	Film & Video	USD
10	100004721	Of Jesus and ...				Nonfiction	Publishing	CAD
11	100005484	Lisa Lim New CD!				Indie Rock	Music	USD
12	1000055792	The Cottage ...				Crafts	Crafts	USD
13	1000056157	G-Spot Place for ...				Games	Games	USD
14	1000064365	Survival Rings				Design	Design	USD
15	1000064918	The Beard				Comic Books	Comics	USD
16	1000069480	Notes From ...				Art Books	Publishing	USD
17	1000070642	Mike Corey's ...				Music	Music	USD
18	1000071625	Boco Tea				Food	Food	USD
19	1000072011	CMUK. Shoes: ...				Fashion	Fashion	USD
20	1000082254	Alice in ...				Theater	Theater	USD
21	1000087442	Mountain brew: ...				Drinks	Food	NOK
22	1000091520	The Book Zoo - ...				Comics	Comics	USD
23	1000102741	Matt Cavanaugh ...				Music	Music	USD
24	1000103940	Superhero ...				DIY	Crafts	GBP
25	1000104688	Permaculture Skills				Webseries	Film & Video	CAD
26	1000104953	Rebel Army ...				Comics	Comics	GBP
27	1000115172	Daily Brew Coffee				Food Trucks	Food	GBP
28	1000117861	Ledr workbook: ...				Product Design	Design	USD

ks-projects-201612-data cleaned - JMP Pro

ID	name	description of project 1	description of project 2	description of project 3	description of project 4	category	main_category	currency
125	100057323	David LaRocca: ...				Illustration	Art	USD
126	1000577059	Calendars for Cats				Photobooks	Photography	USD
127	10005784	DOGMA - Short ...				Drama	Film & Video	GBP
128	1000581546	Project U-Neek				Art	Art	GBP
129	1000581804	Architecture & ...				Architecture	Design	GBP
130	1000586849	#NotMyPresident...				Apparel	Fashion	USD
131	1000590709	Musical Light Suit				Hardware	Technology	USD
132	1000600012	YA science ...				Young Adult	Publishing	GBP
133	1000600526	Rising Star - ...				Hip-Hop	Music	GBP
134	1000602156	HYDRA DICE				Games	Games	USD
135	1000620551	Islamic World - ...				Mobile Games	Games	GBP
136	1000624031	Hangar 1 ...				Flight	Technology	USD
137	1000627718	Adèle's Heart is ...				Theater	Theater	USD
138	1000629643	ODIN: Android ...				Hardware	Technology	USD
139	1000638151	Jenny's ...				Fine Art	Photography	GBP
140	1000639526	Kaleidoscope Man				Film & Video	Film & Video	GBP
141	1000644119	Operación Douve				Action	Film & Video	EUR
142	1000648117	42 Sketches 2013				Art Books	Publishing	USD
143	1000648918	ADVENT SAGA, ...				Video Games	Games	USD
144	1000650960	Crap Amidst ...				Playing Cards	Games	USD
145	1000654344	Packable, ...				Fashion	Fashion	USD
146	1000656794	VoxCube - 8x8x8 ...				Makerspaces	Technology	AUD
147	1000659557	New world ...				Punk	Music	USD
148	1000674477	"GENERATION ...				Documentary	Film & Video	USD
149	1000682369	Rack and Ruin, ...				Fiction	Publishing	GBP
150	1000684975	Help me prepare ...				Painting	Art	USD
151	1000697657	The Forever ...				Hardware	Technology	CAD
152	1000699196	COVENANT				Jazz	Music	USD

## Step 2

Delete the records with launch date having the year 1970



ks-projects-201612-data cleaned - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

Source

Columns (17/1)

Rows

	ID	name	description of project 1	description of project 2	description of project 3	description of project 4	category	main_category	currency	deadline
125	125	ccca: ...					Illustration	Art	USD	10/22/2015 4:30 ...
126	126	Cats					Photobooks	Photography	USD	12/29/2015 2:56 ...
127	127	Short ...					Drama	Film & Video	GBP	03/20/2015 1:00 ...
128	128	Week					Art	Art	GBP	05/08/2013 1:06 ...
129	129	re & ...					Architecture	Design	GBP	09/28/2014 7:47 ...
130	130	incident...					Apparel	Fashion	USD	12/22/2016 4:46 ...
131	131	ht Suit					Hardware	Technology	USD	06/28/2011 9:09 ...
132	132	...					Young Adult	Publishing	GBP	06/16/2016 1:05 ...
133	133	- ...					Hip Hop	Music	GBP	02/22/2013 10:32 ...
134	134	CE					Games	Games	USD	06/06/2012 7:42 ...
135	135	orld - ...					Mobile Games	Games	GBP	08/19/2014 12:52 ...
136	136	...					Flight	Technology	USD	10/18/2014 7:40 ...
137	137	art is ...					Theater	Theater	USD	06/18/2012 2:33 ...
138	138	roid ...					Hardware	Technology	USD	07/08/2014 5:59 ...
139	139						Fine Art	Photography	GBP	08/29/2014 6:14 ...
140	140	pe Man					Film & Video	Film & Video	GBP	12/22/2012 3:59 ...
141	141	Douve					Action	Film & Video	EUR	05/01/2016 7:09 ...
142	142	is 2013					Art Books	Publishing	USD	04/07/2013 10:16 ...
143	143	AGA, ...					Video Games	Games	USD	07/31/2014 4:01 ...
144	144	st ...					Playing Cards	Games	USD	03/22/2016 12:36 ...
145	145	...					Fashion	Fashion	USD	06/07/2014 1:00 ...
146	146	8x8x8 ...					Makerspaces	Technology	AUD	04/01/2016 12:15 ...
147	147	...					Punk	Music	USD	07/05/2016 1:46 ...
148	148	ION ...					Documentary	Film & Video	USD	12/09/2012 4:57 ...
149	149	uin ...					Fiction	Publishing	GBP	06/17/2015 9:52 ...
150	150	repare ...					Painting	Art	USD	04/24/2016 9:25 ...
151	151	r ...					Hardware	Technology	CAD	10/23/2013 11:36 ...
152	152	T					Jazz	Music	USD	12/07/2014 6:02 ...

ks-projects-201612-data cleaned - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

Source

Columns (17/1)

Rows

	ID	name	description of project 1	description of project 2	description of project 3	description of project 4	category	main_category	currer
238	1014518705	Eduin and the ...					Narrative Film	Film & Video	USD
2389	1014523997	PlayOn ...					Music	Music	USD
2390	101453305	HOLOS - Symbol ...					Public Art	Art	USD
2391	101453314	Social Media ...	A Short Film From ...				Shorts	Film & Video	USD
2392	101453652	"The Naked ...					Photobooks	Photography	USD
2393	1014569071	Search For Love ...					Film & Video	Film & Video	USD
2394	1014577403	Rock Paper ...					Publishing	Publishing	USD
2395	1014585959	ZVIDA ...					Public Art	Art	USD
2396	1014610497	THE HUNT FOR ...					Print	Journalism	USD
2397	1014610768	Of Gods and ...					Comic Books	Comics	USD
2398	1014620304	Æ - Be the ...					Comics	Comics	CAD
2399	1014632933	Dungeonnmans: ...					Video Games	Games	USD
2400	1014638860	Drag Seeker - ...					People	Photography	USD
2401	101466029	MITCH ...					Art	Art	USD
2402	1014680652	Stories of the ...					People	Photography	USD
2403	1014686065	Publish Poor ...					Nonfiction	Publishing	USD
2404	1014692205	Orin Acute ...					Gadgets	Technology	CAD
2405	1014694874	(BullyTube )					Web	Journalism	USD
2406	1014702475	Dead Film Society					Television	Film & Video	USD
2407	1014708315	CREME DE LA ...					Food	Food	USD
2408	1014721567	CySec - Alliance ...					Technology	Technology	EUR
2409	1014722794	Audio Killer ...					Horror	Film & Video	USD
2410	101472381	ArtPrize 2014 - ...					Painting	Art	USD
2411	1014726003	Mr. Babes Needs ...					Painting	Art	USD
2412	1014738284	EMBER wear Ski ...					Wearables	Technology	GBP
2413	1014739191	Wind - A Short ...					Narrative Film	Film & Video	USD
2414	1014750154	SEKTRA_A ...					Fiction	Publishing	USD
2415	1014756598	LIVING COLOR - ...					Shorts	Film & Video	USD

### Step 3

Change the state failed, cancelled, suspended, undefined to 0 and successful to 1

Count	Old Values (6)	New Values (2)
168522	failed	0
32398	cancelled	0
1476	suspended	0
683	undefined	1
113143	successful	1
4439	live	1

### Step 4

Delete the observations with status as live and name the data set as “Delete the live”

Row	launched	pledged	state	final state	backers	country	deadline year	launched year	goal usd	pledged usd	funding duration	ratio pledge to goal
1	11 4:46 PM	0	cancelled	0	0	US	2011	2011	0.01	0	36	0
2	09 7:54 AM	100	successful	1	6	US	2009	2009	0.01	100	9	10000
3	12 7:23 AM	0	failed	0	0	US	2012	2012	0.15	0	51	0
4	11 3:59 PM	0	failed	0	0	US	2011	2011	0.5	0	7	0
5	10 11:46 ...	0	failed	0	0	US	2010	2010	1	0	82	0
6	11 12:21 ...	0	failed	0	0	GB	2014	2014	1.28	0	60	0
7	16 9:17 PM	0	failed	0	0	US	2016	2016	1	0	60	0
8	12 8:10 PM	0	failed	0	0	US	2012	2012	1	0	60	0
9	15 1:25 AM	0	cancelled	0	0	US	2015	2015	1	0	60	0
10	16 4:56 PM	0	failed	0	0	US	2016	2016	1	0	60	0
11	16 11:34 ...	0	live	1	0	US	2017	2017	1	0	60	0
12	14 12:11 ...	0	cancelled	0	0	US	2015	2014	1	0	60	0
13	12 1:26 AM	0	failed	0	0	US	2012	2012	1	0	60	0
14	15 6:25 PM	0	failed	0	0	US	2015	2015	1	0	60	0
15	11 7:07 PM	0	failed	0	0	US	2011	2011	1	0	55	0
16	11 7:19 PM	0	cancelled	0	0	US	2011	2011	1	0	54	0
17	15 11:20 ...	0	cancelled	0	0	US	2016	2015	1	0	50	0
18	15 4:23 PM	0	cancelled	0	0	BE	2016	2015	1.13	0	45	0
19	15 12:36 ...	0	failed	0	0	US	2015	2015	1	0	44	0
20	14 7:25 AM	0	failed	0	0	US	2014	2014	1	0	41	0
21	15 2:38 PM	0	failed	0	0	US	2015	2015	1	0	40	0
22	15 1:55 PM	0	cancelled	0	0	US	2016	2015	1	0	36	0
23	16 6:00 PM	0	failed	0	0	US	2016	2016	1	0	33	0
24	16 3:14 PM	0	cancelled	0	0	GB	2016	2016	1.28	0	32	0
All rows	320,661											
Selected	4,439											
Excluded	0											
Hidden	0											
Labelled	0											

3.ks-projects-201612-after conversion outliers inconsistency and delete live - JMP Pro

	ID	name	category	main_category	currency	currency rate	deadline	goal	laur
Source	1	68856463 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/2012
	2	620302213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/20
	3	9572984 Nana	Shorts	Film & Video	USD	1	03/16/2012 6:23 AM	0.15	01/25/20
	4	219760504 RocknRoll NoisePollution	Documentary	Film & Video	USD	1	07/19/2011 3:59 PM	0.5	07/12/20
	5	1316126964 Antipledge	Performance Art	Art	USD	1	04/01/2010 6:00 PM	1	01/09/2012
	6	1087894007 Lets Build Riders Company To Be Known Globally	Product Design	Design	GBP	1.28	11/16/2014 12:21 AM	1	09/11/2012
	7	112629146 Till the End of Time	Nonfiction	Publishing	USD	1	11/11/2016 9:17 PM	1	09/12/20
	8	149339585 "Marching on Memphis..."	Film & Video	Film & Video	USD	1	07/16/2012 8:10 PM	1	05/17/20
	9	1494614896 Poker Paranoia (Canceled)	Fashion	Fashion	USD	1	06/11/2015 1:25 AM	1	04/12/20
	10	1542552782 Adventures of Christiana, Mercy, and Family and ...	Publishing	Publishing	USD	1	10/22/2016 4:56 PM	1	08/23/20
	11	2059635082 Strange Graphs New Album (Canceled)	Electronic Music	Music	USD	1	02/02/2015 12:11 AM	1	12/04/20
	12	209708644 SLICE three: FUGUE	Documentary	Film & Video	USD	1	04/14/2012 1:26 AM	1	02/14/20
	13	2140232586 The Guide to Being unsuccessful	Art	USD	1	05/02/2015 6:25 PM	1	03/03/20	
	14	808941435 Springfield Hemp Fest	Performance Art	Art	USD	1	08/22/2011 10:28 PM	1	06/28/20
	15	294171855 Chesterfield Sculpture Project	Sculpture	Art	USD	1	09/02/2011 6:00 AM	1	07/10/20
	16	128494572 Yetticate Academy (Canceled)	Publishing	Publishing	USD	1	01/21/2016 11:20 PM	1	12/02/20
	17	204741853 Help EDM project 'Chav Fury' to get off (Canceled)	Electronic Music	Music	EUR	1.13	01/29/2016 4:23 PM	1	12/15/20
	18	2064457704 Fare thee well GRATEFUL DEAD documentary	Photobooks	Photography	USD	1	04/03/2015 12:36 AM	1	02/18/2012
	19	1054378026 Reece Ran's WINTER	Fiction	Publishing	USD	1	10/30/2014 6:25 AM	1	09/19/20
	20	1882394705 JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...	Fiction	Publishing	USD	1	12/31/2015 9:54 PM	1	11/21/20
	21	1357482430 . cliff (Canceled)	Tabletop Games	Games	USD	1	01/31/2016 1:55 PM	1	12/26/20
	22	310936687 SMILE #UWokeUp	Nonfiction	Publishing	USD	1	08/29/2016 7:36 PM	1	07/21/20
	23	378468728 Things to do in London (Canceled)	Zines	Publishing	GBP	1.28	02/29/2016 1:00 AM	1	01/28/20
	24	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...	Horror	Film & Video	USD	1	11/01/2014 7:59 AM	1	09/30/2012
	25	944578907 Spore Animated (Canceled)	Comedy	Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/20
	26	1873533970 Game Development	Video Games	Games	USD	1	04/08/2015 11:30 PM	1	03/08/20
	27	998486147 Half Bloods Series	Drama	Film & Video	USD	1	04/17/2016 4:57 AM	1	03/17/20
	28	1711779658 Cameron Hill Vlon	Film & Video	Film & Video	GRP	1.28	10/19/2016 5:15 PM	1	09/19/20

## Step 5

Delete the observations with status as undefined and use Decision tree modeling

3.ks-projects-201612-after conversion outliers inconsistency and delete live - JMP Pro

	ID	name	category	main_category	currency	currency rate	deadline	goal	laur
Source	1	68856463 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/2012
	2	620302213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/20
	3	9572984 Nana	Shorts	Film & Video	USD	1	03/16/2012 6:23 AM	0.15	01/25/20
	4	219760504 RocknRoll NoisePollution	Documentary	Film & Video	USD	1	07/19/2011 3:59 PM	0.5	07/12/20
	5	1316126964 Antipledge	Performance Art	Art	USD	1	04/01/2010 6:00 PM	1	01/09/2012
	6	1087894007 Lets Build Riders Company To Be Known Globally	Product Design	Design	GBP	1.28	11/16/2014 12:21 AM	1	09/11/2012
	7	112629146 Till the End of Time	Nonfiction	Publishing	USD	1	11/11/2016 9:17 PM	1	09/12/20
	8	149339585 "Marching on Memphis..."	Film & Video	Film & Video	USD	1	07/16/2012 8:10 PM	1	05/17/20
	9	1494614896 Poker Paranoia (Canceled)	Fashion	Fashion	USD	1	06/11/2015 1:25 AM	1	04/12/20
	10	1542552782 Adventures of Christiana, Mercy, and Family and ...	Publishing	Publishing	USD	1	10/22/2016 4:56 PM	1	08/23/20
	11	2059635082 Strange Graphs New Album (Canceled)	Electronic Music	Music	USD	1	02/02/2015 12:11 AM	1	12/04/20
	12	209708644 SLICE three: FUGUE	Documentary	Film & Video	USD	1	04/14/2012 1:26 AM	1	02/14/20
	13	2140232586 The Guide to Being unsuccessful	Art	USD	1	05/02/2015 6:25 PM	1	03/03/20	
	14	808941435 Springfield Hemp Fest	Performance Art	Art	USD	1	08/22/2011 10:28 PM	1	06/28/20
	15	294171855 Chesterfield Sculpture Project	Sculpture	Art	USD	1	09/02/2011 6:00 AM	1	07/10/20
	16	128494572 Yetticate Academy (Canceled)	Publishing	Publishing	USD	1	01/21/2016 11:20 PM	1	12/02/20
	17	204741853 Help EDM project 'Chav Fury' to get off (Canceled)	Electronic Music	Music	EUR	1.13	01/29/2016 4:23 PM	1	12/15/20
	18	2064457704 Fare thee well GRATEFUL DEAD documentary	Photobooks	Photography	USD	1	04/03/2015 12:36 AM	1	02/18/2012
	19	1054378026 Reece Ran's WINTER	Fiction	Publishing	USD	1	10/30/2014 6:25 AM	1	09/19/20
	20	1882394705 JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...	Fiction	Publishing	USD	1	12/31/2015 9:54 PM	1	11/21/20
	21	1357482430 . cliff (Canceled)	Tabletop Games	Games	USD	1	01/31/2016 1:55 PM	1	12/26/20
	22	310936687 SMILE #UWokeUp	Nonfiction	Publishing	USD	1	08/29/2016 7:36 PM	1	07/21/20
	23	378468728 Things to do in London (Canceled)	Zines	Publishing	GBP	1.28	02/29/2016 1:00 AM	1	01/28/20
	24	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...	Horror	Film & Video	USD	1	11/01/2014 7:59 AM	1	09/30/2012
	25	944578907 Spore Animated (Canceled)	Comedy	Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/20
	26	1873533970 Game Development	Video Games	Games	USD	1	04/08/2015 11:30 PM	1	03/08/20
	27	998486147 Half Bloods Series	Drama	Film & Video	USD	1	04/17/2016 4:57 AM	1	03/17/20
	28	1711779658 Cameron Hill Vlon	Film & Video	Film & Video	GRP	1.28	10/19/2016 5:15 PM	1	09/19/20

4.ks-projects-201612-after conversion outliers inconsistency and delete live and undefined - JMP Pro												
File	Edit	Tables	Rows	Cols	DOE	Analyze	Graph	Tools	Add-Ins	View	Window	Help
4.ks-projects-201612-aft...		ID	name		category	main_category	currency	currency rate	deadline	goal	laur	
Columns (27/1)		1	688564643 Word-of-mouth publishing: get "Corruptions" out ...		Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/2012	
deadline		2	620302113 LOVELAND Round 6: A Force More Powerful		Conceptual Art	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/2010	
goal		3	9572984 Nana		Shorts	Film & Video	USD	1	03/16/2012 6:23 AM	0.15	01/25/2012	
launched		4	219760504 RocknRoll NoisePollution		Documentary	Film & Video	USD	1	07/19/2011 3:59 PM	0.5	07/12/2012	
pledged		5	1316126964 Antipledge		Performance Art	Art	USD	1	04/01/2010 6:00 PM	1	01/09/2011	
state		6	108/894007 Lets Build Riders Company To Be Known Globally		Product Design	Design	GBP	1.28	11/16/2014 12:21 AM	1	09/11/2012	
final state		7	112629146 Till the End of Time		Nonfiction	Publishing	USD	1	11/11/2016 9:17 PM	1	09/12/2012	
backers		8	149339585 "Marching on Memphis..."		Film & Video	Film & Video	USD	1	07/16/2012 8:10 PM	1	05/17/2012	
country		9	1494614896 Poker Paranoia (Canceled)		Fashion	Fashion	USD	1	06/11/2015 1:25 AM	1	04/12/2012	
deadline year		10	1542552782 Adventures of Christiana, Mercy, and Family and ...		Publishing	Publishing	USD	1	10/22/2016 4:56 PM	1	08/23/2012	
launched year		11	2059635082 Strange Graphs New Album (Canceled)		Electronic Music	Music	USD	1	02/02/2015 12:11 AM	1	12/04/2012	
pledged year		12	209708644 SLICE three FUGUE		Documentary	Film & Video	USD	1	04/14/2012 1:26 AM	1	02/14/2012	
goal usd		13	2140232586 The Guide to Being Unsuccessful		Art	Art	USD	1	05/02/2015 6:25 PM	1	03/03/2012	
pledged usd		14	808941435 Springfield Hemp Fest		Performance Art	Art	USD	1	08/22/2011 10:28 PM	1	06/28/2012	
funding duration		15	294171855 Chesterfield Sculpture Project		Sculpture	Art	USD	1	09/02/2011 6:00 AM	1	07/10/2012	
ratio pledge to goal		16	128494572 Yeticate Academy (Canceled)		Publishing	Publishing	USD	1	01/21/2016 11:20 PM	1	12/02/2012	
ratio pledge to backers		17	207471853 Help EDM project 'Chav Fury' to get off (Canceled)		Electronic Music	Music	EUR	1.13	01/29/2016 4:23 PM	1	12/15/2012	
Range Scale[backers]+		18	206457704 Fare thee well GRATEFUL DEAD documentary		Photobooks	Photography	USD	1	04/03/2015 12:36 AM	1	02/18/2012	
Range Scale[goal usd]+		19	1054378026 Reece Ran's WINTER		Fiction	Publishing	USD	1	10/30/2014 6:25 AM	1	09/19/2012	
Range Scale[pledged usd]+		20	1882394705 JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...		Fiction	Publishing	USD	1	12/31/2015 9:54 PM	1	11/21/2012	
Range Scal...ng duration+*		21	1357482430 cfff (Canceled)		Tabletop Games	Games	USD	1	01/31/2016 1:55 PM	1	12/26/2012	
Range Scal...dgo to goal+*		22	310936687 SMILE #UwokeUp		Nonfiction	Publishing	USD	1	08/29/2016 7:36 PM	1	07/21/2012	
Range Scal...e to backers+*		23	378460728 Things to do in London (Canceled)		Zines	Publishing	GBP	1.28	02/29/2016 1:00 AM	1	01/28/2012	
Rows		24	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...		Horror	Film & Video	USD	1	11/01/2014 7:59 AM	1	09/30/2012	
All rows		315,539			Comedy	Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/2012	
Selected		0	944578907 Spore Animated (Canceled)		Video Games	Games	USD	1	04/08/2015 11:30 PM	1	03/08/2012	
Excluded		0	1873533970 Game Development		Drama	Film & Video	USD	1	04/17/2016 4:57 AM	1	03/17/2012	
Hidden		0	27 998486147 Half Bloods Series		Film & Video	Film & Video	GRP	1.28	10/19/2016 9:51:15 PM	1	09/19/2012	
Labelled		0	28 1711779658 Cameron Hill Vlon									

## Step 6

Delete the observations with status as suspended

5.ks-projects-201612-after conversion outliers inconsistency and delete live and undefined and suspend - JMP Pro

	pledge to goal	ratio pledge to backers	Range Scale[backers]	Range Scale[goal usd]	Range Scale[pledged ...]	Range Scale[funding duration]	Range Scale[rat...pled...]	Range Scale[rat...dge ...]
17	0	0	0	8.75e-9	0	0.4835164835	0	0
18	0	0	0	7.734375e-9	0	0.4725274725	0	0
19	0	0	0	7.734375e-9	0	0.4395604396	0	0
20	0	0	0	7.734375e-9	0	0.4285714286	0	0
21	0	0	0	7.734375e-9	0	0.3846153846	0	0
22	0	0	0	7.734375e-9	0	0.3516483516	0	0
23	0	0	0	9.921875e-9	0	0.3406593407	0	0
24	0	0	0	7.734375e-9	0	0.3406593407	0	0
25	0	0	0	7.734375e-9	0	0.3406593407	0	0
26	0	0	0	7.734375e-9	0	0.3296703297	0	0
27	0	0	0	7.734375e-9	0	0.3296703297	0	0
28	0	0	0	9.921875e-9	0	0.3186813187	0	0
29	0	0	0	9.921875e-9	0	0.3186813187	0	0
30	0	0	0	9.921875e-9	0	0.3186813187	0	0
31	0	0	0	9.921875e-9	0	0.3186813187	0	0
32	0	0	0	8.75e-9	0	0.3186813187	0	0
33	0	0	0	7.734375e-9	0	0.3186813187	0	0
34	0	0	0	7.734375e-9	0	0.3186813187	0	0
35	0	0	0	7.734375e-9	0	0.3186813187	0	0
36	0	0	0	7.734375e-9	0	0.3186813187	0	0
37	0	0	0	7.734375e-9	0	0.3186813187	0	0
38	0	0	0	7.734375e-9	0	0.3186813187	0	0
39	0	0	0	7.734375e-9	0	0.3186813187	0	0
40	0	0	0	7.734375e-9	0	0.3186813187	0	0
41	0	0	0	7.734375e-9	0	0.3186813187	0	0
42	0	0	0	7.734375e-9	0	0.3186813187	0	0
43	0	0	0	7.734375e-9	0	0.3186813187	0	0
44	0	0	0	7.734375e-9	0	0.3186813187	0	0

## APPENDIX 1.3

This part includes the different situations we tried to find the best data set for modeling.

1. We used the data set “delete the live (state)” to build the model by using the decision tree model
  - a. Use Stratified Split proportion: Training set- 0.3 Validation set- 0.3 Test set - 0.4

3.ks-projects-201612-after conversion outliers inconsistency and delete live - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

Stratified Split - JMP Pro

Select Column: final state

Specify Data Proportions:

- Training Set: 0.3
- Validation Set: 0.3
- Test Set: 0.4
- Alter Proportions in Both Training and Validation

Action: OK, Cancel, Help

evaluations done

	RowID	ID	name	category	main_category	currency	currency rate	deadline	goal	launched	pledged	state	final state
14	808941435	Springfield Hemp Fest		Performance Art	Art	USD	1	12/13/2011 4:46 PM	0.01	11/07/2		0	
15	294171855	Chesterfield Sculpture Project		Sculpture	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/2		0	
16	12894572	Yetticate Academy (Cancelled)		Publishing	Publishing	USD	1	03/16/2012 6:23 AM	0.15	01/25/2		0	
17	204741853	Help EDM project 'Chav Fury' to get off (Cancelled)		Electronic Music	Music	EUR	1.13	01/29/2016 3:59 PM	0.5	07/12/2		0	
18	2064457704	Fare thee well GRATEFUL DEAD documentary		Photobooks	Photography	USD	1	04/01/2010 6:00 PM	1	01/09/		0	
19	1054378026	Reeces Ran's WINTER		Fiction	Publishing	USD	1	10/22/2016 4:56 PM	1	08/23/2		0	
20	1882394705	JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...		Fiction	Publishing	USD	1	12/31/2015 12:11 AM	1	12/04/		0	
21	1357482430	cffff (Cancelled)		Tabletop Games	Games	USD	1	01/31/2016 1:55 PM	1	12/26/2		0	
22	310936687	SMILE #UWokeUp		Nonfiction	Publishing	USD	1	08/29/2016 7:36 PM	1	07/27/2		0	
23	37846828	Things to do in London (Cancelled)		Zines	Publishing	GBP	1.28	02/29/2016 1:00 AM	1	01/28/2		0	
24	515995141	THE SOPHIA MOVIE - A Suspense/Horror Film ...		Horror	Film & Video	USD	1	11/01/2014 7:59 AM	1	09/30/		0	
25	4944578907	Spore Animated (Cancelled)		Comedy	Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/2		0	
26	1873533970	Game Development		Video Games	Games	USD	1	04/08/2015 11:30 PM	1	03/08/		0	

3.ks-projects-201612-after conversion outliers inconsistency and delete live\_1\_PROP\_0.5 - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

3.ks-projects-201612-aft... >

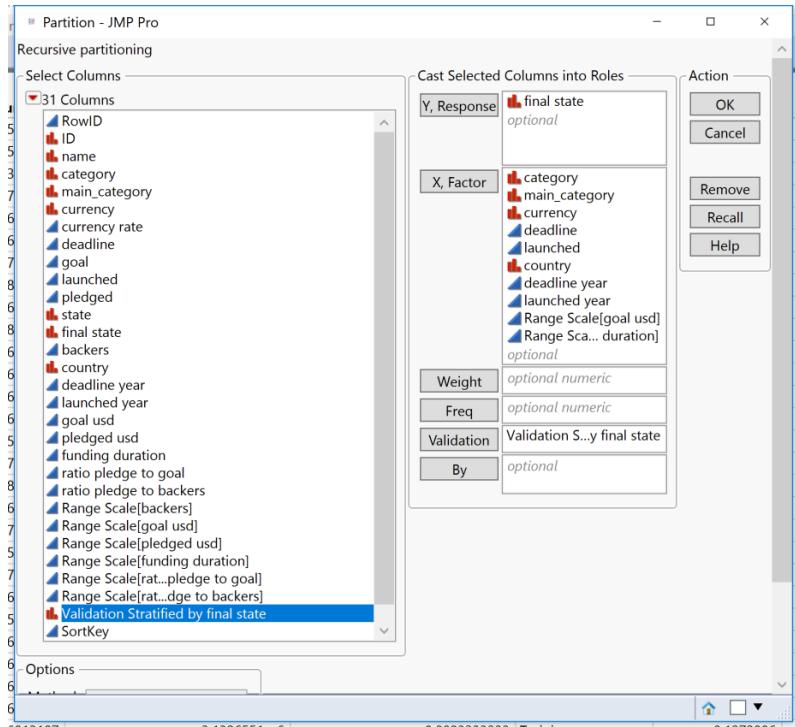
Source

Columns (31/0)

RowID	ID	name	category	main_category	currency	currency rate	deadline	goal	launched	pledged	state	final state	
1	297169	1672450808	ErgStick - ...	Gadgets	Technology	GBP	1.28	06/28/2015 1:00 ...	71337	05/21/2015 12:43...	9978	failed	0
2	283915	965498196	You're Just ...	Horror	Film & Video	USD	1	04/22/2015 11: ...	47000	03/13/2015 12:11...	0	cancelled	0
3	275747	704785626	How to Obtain ...	Publishing	Publishing	USD	1	06/26/2015 9:59 ...	35000	04/21/2015 9:59 ...	30	failed	0
4	58007	1872101823	The Pikes Peak ...	Classical Music	Music	USD	1	05/11/2014 5:19 ...	1300	04/17/2014 3:00 ...	601	failed	0
5	182545	1074484037	Wild by Nature ...	Nature	Photography	USD	1	12/25/2014 10:07...	8000	11/25/2014 10:07...	0	failed	0
6	238657	707722833	Believe a USC ...	Shorts	Film & Video	USD	1	06/22/2013 8:00 ...	16000	05/23/2013 8:00 ...	235.01	cancelled	0
7	257088	1384171269	Raining Light ...	Restaurants	Food	EUR	1.13	12/07/2015 6:49 ...	24700	11/16/2015 6:49 ...	2006	cancelled	0
8	45503	1145008426	Glass Christmas ...	Glass	Crafts	USD	1	06/21/2016 4:06 ...	1000	06/04/2016 4:31 ...	205	failed	0
9	201848	862662482	Invictus ...	Live Games	Games	USD	1	05/15/2015 8:04 ...	10000	04/15/2015 8:04 ...	150	failed	0
10	88683	1968273847	Real Life Magazine	Publishing	Publishing	GBP	1.28	10/22/2016 2:46 ...	2500	10/08/2016 2:46 ...	1	failed	0
11	37361	224561142	"Immersive ...	Experimental	Theater	USD	1	12/04/2015 4:33 ...	898	11/04/2015 4:33 ...	0	failed	0
12	147272	1968562975	Baffles.com - ...	3D Printing	Technology	NZD	0.65	03/10/2015 9:05 ...	5000	02/08/2015 10:05...	557	failed	0
13	131929	1831722293	HEALER (Cancelled)	Fiction	Publishing	USD	1	07/02/2012 6:24 ...	4500	06/02/2012 6:24 ...	975	cancelled	0
14	72595	1999879373	sMedGE	Product Design	Design	USD	1	08/15/2016 2:46 ...	2000	07/16/2016 2:46 ...	0	failed	0
15	309998	1131825016	Survival Horror ...	Video Games	Games	USD	1	07/16/2011 6:47 ...	200000	06/01/2016 6:47 ...	41	failed	0
16	123259	401110579	A beautiful ...	Art	Art	USD	1	08/28/2014 9:24 ...	4000	08/07/2014 9:24 ...	0	cancelled	0
17	149100	2198323232	Arancini ...	Food Trucks	Food	GBP	1.28	07/18/2016 1:20 ...	5000	07/04/2016 1:20 ...	1241	failed	0
18	286563	904709753	Toss'd - Farm ...	Food	Food	USD	1	10/25/2014 6:04 ...	50000	09/25/2014 6:04 ...	15	failed	0
19	11882	727413761	EVERYTHING - ...	Art	Art	GBP	1.28	06/08/2013 2:40 ...	300	05/14/2013 2:40 ...	38	cancelled	0
20	265829	553307431	the ALMIGHTY	Narrative Film	Film & Video	USD	1	06/22/2012 7:34 ...	25900	05/08/2012 7:34 ...	1531.25	failed	0
21	69106	1444451537	Expanding ...	Graphic Design	Design	USD	1	09/15/2014 6:59 ...	1700	08/26/2014 3:43 ...	0	failed	0
22	7811	546994405	Second Chance - ...	Film & Video	Film & Video	GBP	1.28	03/21/2014 9:02 ...	200	02/19/2014 10:02...	65	failed	0
23	197223	985560357	The God Project	Nonfiction	Publishing	USD	1	11/05/2016 11:47...	10000	09/22/2016 12:47...	1	failed	0
24	59423	615272367	The Cupcake ...	Food	Food	USD	1	07/04/2015 6:59 ...	1500	06/02/2015 7:17 ...	0	failed	0

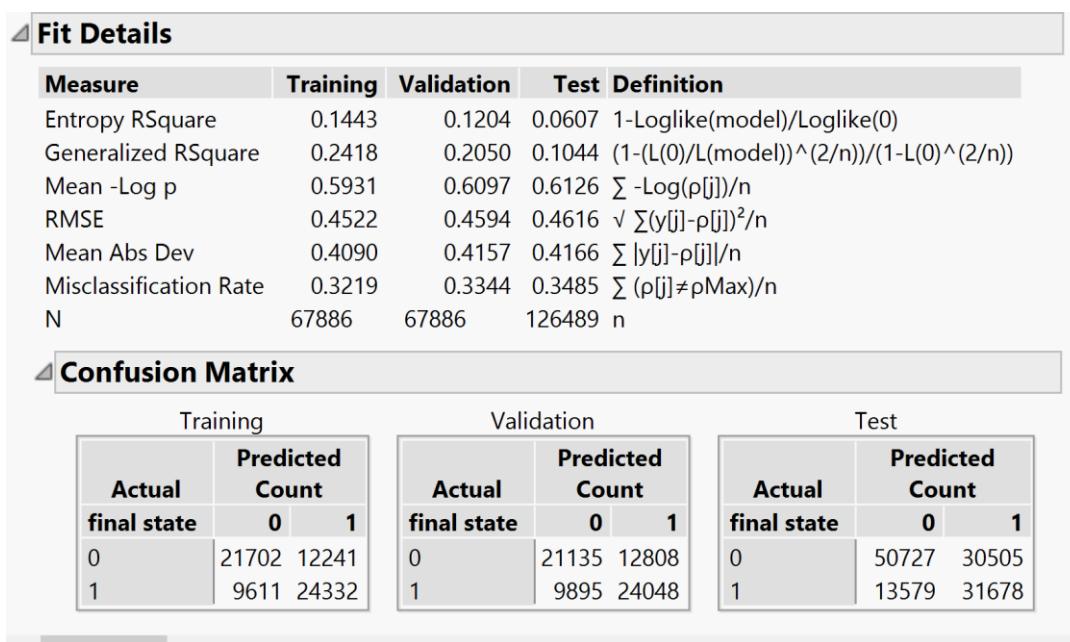
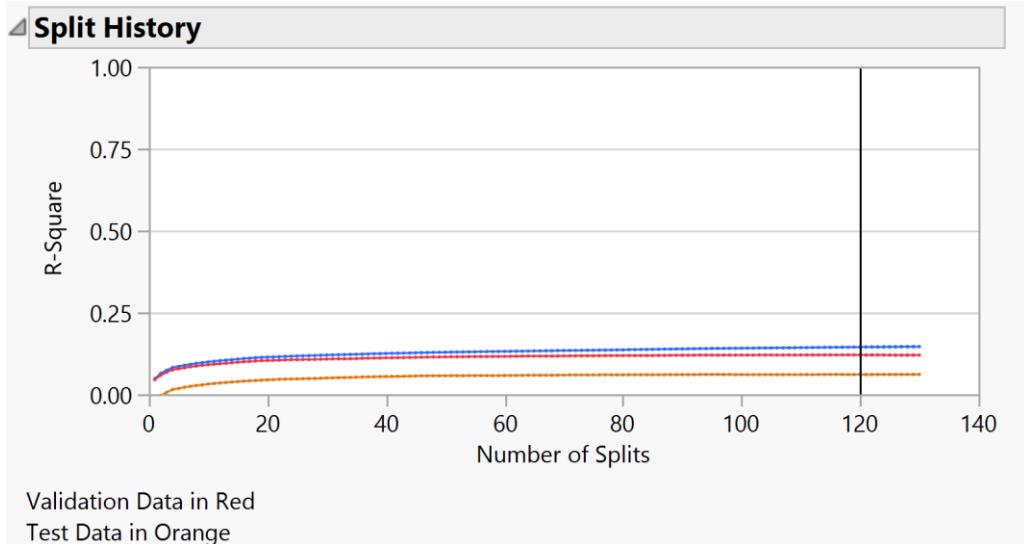
Sample size: 262,261

Input variables:



Results:

	RSquare	N	Number of Splits
Training	0.144	67886	120
Validation	0.120	67886	
Test	0.061	126489	



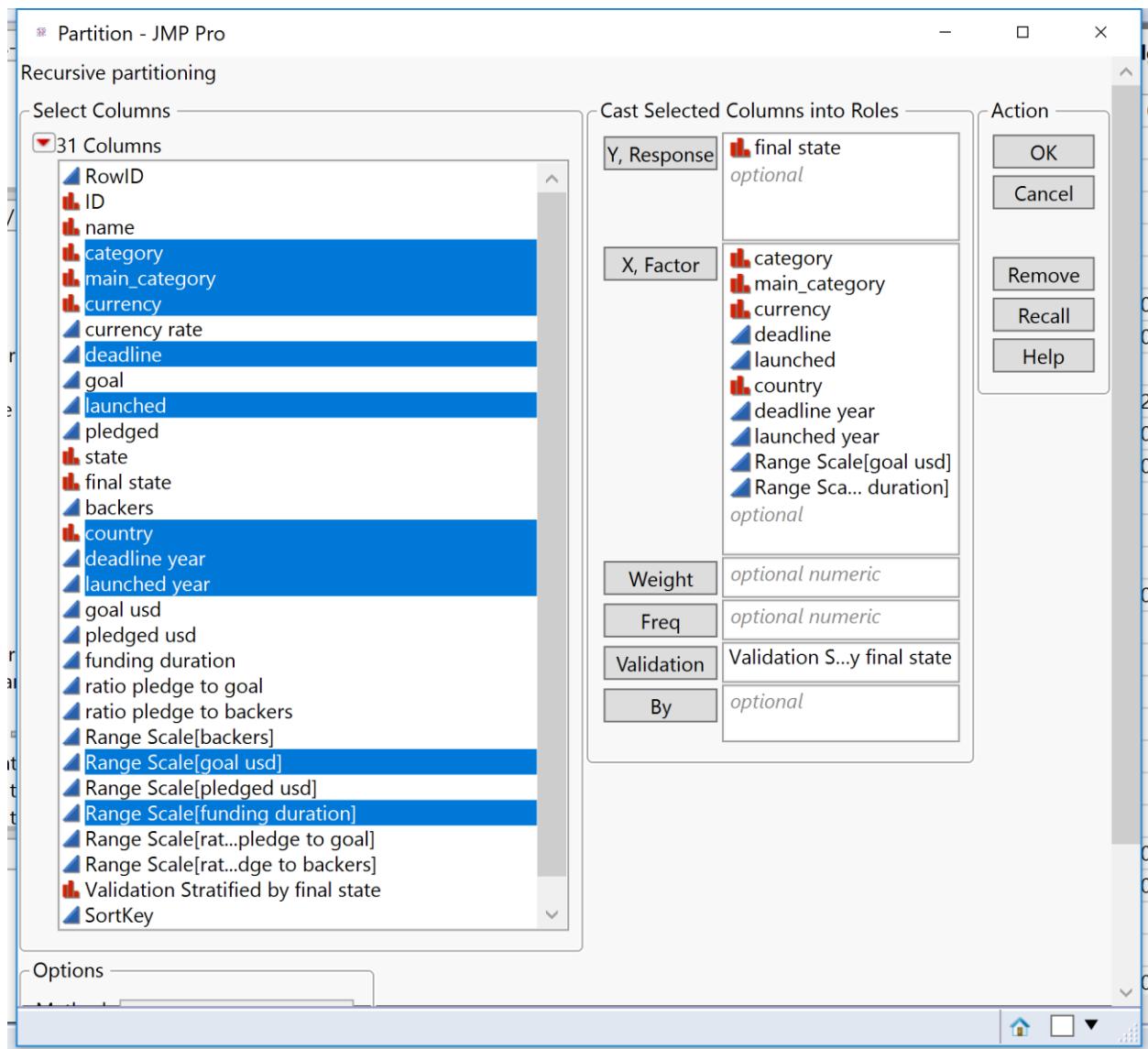
We are getting good accuracy 67.8% for training data set and 65.15% for test data set.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	21852	67.81%
Validation	22703	66.56%
Test	44084	65.15%

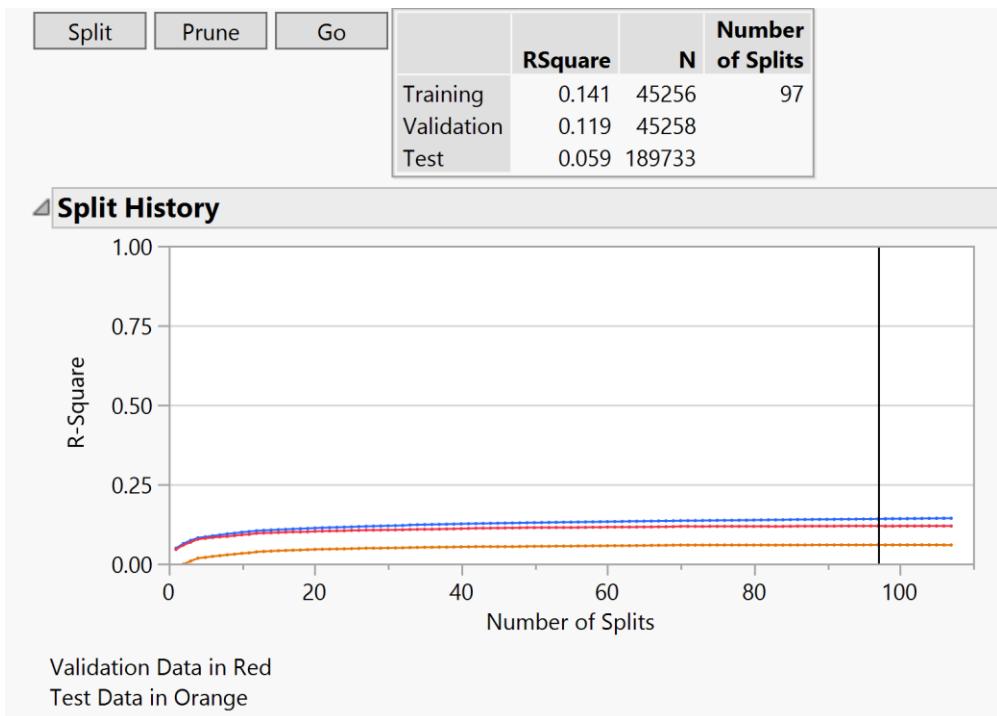
b. Use Stratified Split Proportion: Training set - 0.2 Validation set- 0.2 Test set - 0.6

Sample size: 280,247

## Input variables:



Results:



**Fit Details**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1410	0.1179	0.0586	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2367	0.2010	0.1010	$(1-(L(0)/L(model)))^{(2/n)})/(1-L(0)^{(2/n)})$
Mean -Log p	0.5954	0.6114	0.6139	$\sum -\log(p[j])/n$
RMSE	0.4531	0.4603	0.4624	$\sqrt{\sum(y[j]-\rho[j])^2/n}$
Mean Abs Dev	0.4108	0.4171	0.4179	$\sum  y[j]-\rho[j] /n$
Misclassification Rate	0.3251	0.3387	0.3487	$\sum (\rho[j] \neq \rho_{Max})/n$
N	45256	45258	189733	n

**Confusion Matrix**

Training		Validation		Test	
		Predicted Count		Predicted Count	
Actual	Predicted	Actual	Predicted	Actual	Predicted
final state	Count	final state	Count	final state	Count
final state	0 1	final state	0 1	final state	0 1
0	14649 7979	0	14194 8435	0	76728 45119
1	6735 15893	1	6894 15735	1	21032 46854

We are getting good accuracy 67.5% for training data set and 65.13% for test data set. This split provides less accuracy ( $65.13\% < 67.5\%$ ) for test data set comparing the split “Train 0.3 Validation 0.3 Test 0.4”. Therefore, we keep the split “Train 0.3 Validation 0.3 Test 0.4”.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	14714	67.49%

Validation	15329	66.13%
Test	66151	65.13%

### c. Use Stratified Split Proportion: Training set- 0.4 Validation set - 0.4 Testing set - 0.2

Screenshot of JMP Pro showing the 'Stratified Split' dialog and the resulting data table.

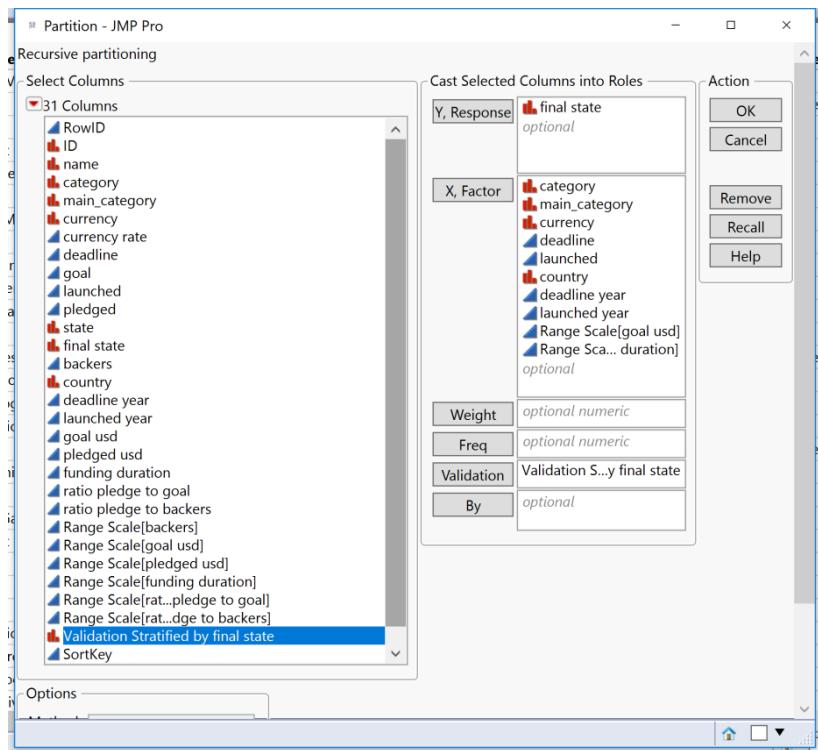
The 'Specify Data Proportions' dialog shows:

- Training Set: 0.4
- Validation Set: 0.4
- Test Set: 0.2
- Alter Proportions in Both Training and Validation (radio button selected)

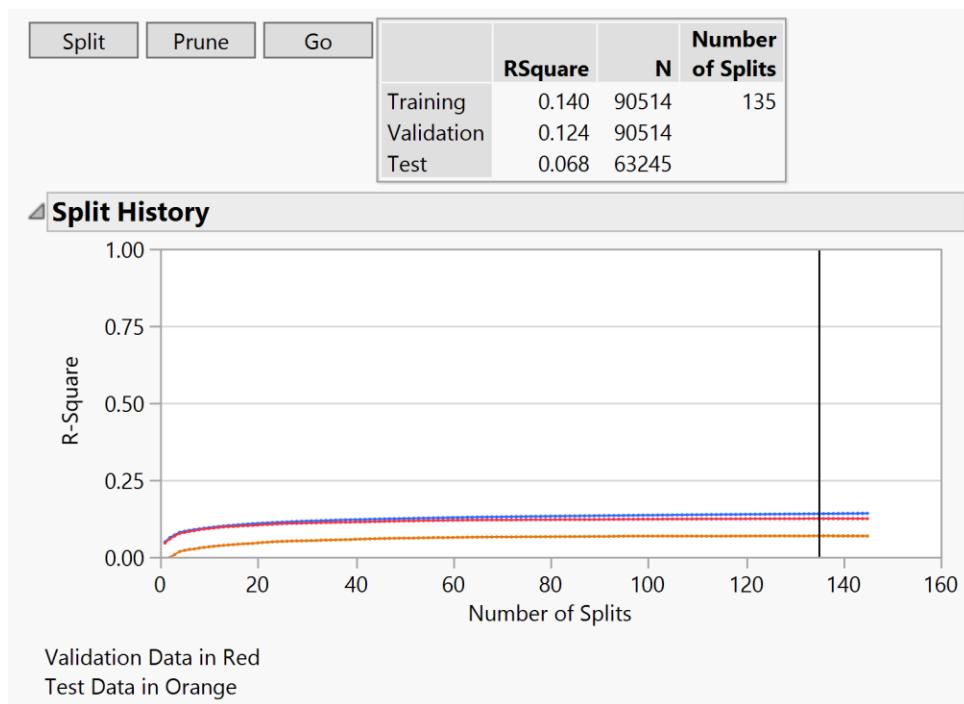
The resulting data table (8.ks-projects-201612-after conversion outliers inconsistency and delete live\_0.4&0.4&0.2 - JMP Pro) contains 244,273 rows and includes columns such as RowID, ID, name, category, main\_category, currency, currency rate, deadline, goal, launched, pledged, state, final state, backers, country, deadline year, launched year, goal usd, pledged usd, funding duration, ratio pledge to goal, ratio pledge to backers, and Ratio Scale(backers).

Sample size: 244,273

Input variables:



### Results:



### Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1399	0.1243	0.0682	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2350	0.2110	0.1169	$(1 - (L(0)/L(\text{model}))^{(2/n)})/(1 - L(0)^{(2/n)})$
Mean -Log p	0.5962	0.6070	0.6076	$\sum -\text{Log}(p[j])/n$
RMSE	0.4536	0.4583	0.4595	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4116	0.4160	0.4158	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3254	0.3340	0.3417	$\sum (p[j] \neq p_{\text{Max}})/n$
N	90514	90514	63245	n

### Confusion Matrix

Training			Validation			Test		
Actual		Predicted Count	Actual		Predicted Count	Actual		Predicted Count
final state	0	1	final state	0	1	final state	0	1
0	29186	16071	0	28736	16521	0	25923	14693
1	13386	31871	1	13708	31549	1	6915	15714

We are getting good accuracy 67.5% for training data set and 65.83% for test data set. This split provides more accuracy ( $65.83\% > 65.15\%$ ) for test data set comparing the split “Train 0.3 Validation 0.3 Test 0.4”. Therefore, we keep the split “Train 0.4 Validation 0.4 Test 0.2” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	29457	67.46%
Validation	30229	66.60%
Test	21608	65.83%

2. We used the data set “all (state)” and “undefined is unsuccessful” to build the model by using the decision tree model

a. Use Stratified split proportion: Training set - 0.2 Validation set - 0.2 Test set - 0.6

ID	name	category	main_category	currency	currency rate	deadline	goal	laur
1	68856463 Word of mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/2C
		Video	USD	1	12/04/2009 7:07 AM	0.01	11/25/20	
		Video	USD	1	03/16/2012 6:23 AM	0.15	01/25/20	
		Video	USD	1	07/19/2011 3:59 PM	0.5	07/12/2C	
		USD	USD	1	04/01/2010 6:00 PM	1	01/09/2I	
		GBP	USD	1.28	11/16/2014 12:21 A...	1	09/17/2I	
		ing	USD	1	11/11/2016 9:17 PM	1	09/12/2C	
		Video	USD	1	07/16/2012 8:10 PM	1	05/17/2C	
		ing	USD	1	06/11/2015 1:25 AM	1	04/12/20	
		ing	USD	1	10/22/2016 4:56 PM	1	08/23/2C	
		USD	USD	1	02/03/2017 11:34 P...	1	12/05/2I	
		USD	USD	1	02/02/2015 12:11 A...	1	12/04/2I	
		Video	USD	1	04/14/2012 1:26 AM	1	02/14/20	
		USD	USD	1	05/02/2015 6:25 PM	1	03/03/2C	
		USD	USD	1	08/22/2011 10:28 P...	1	06/28/2C	
		USD	USD	1	09/02/2011 6:00 AM	1	07/10/2C	
		USD	USD	1	01/21/2016 11:20 P...	1	12/03/2I	

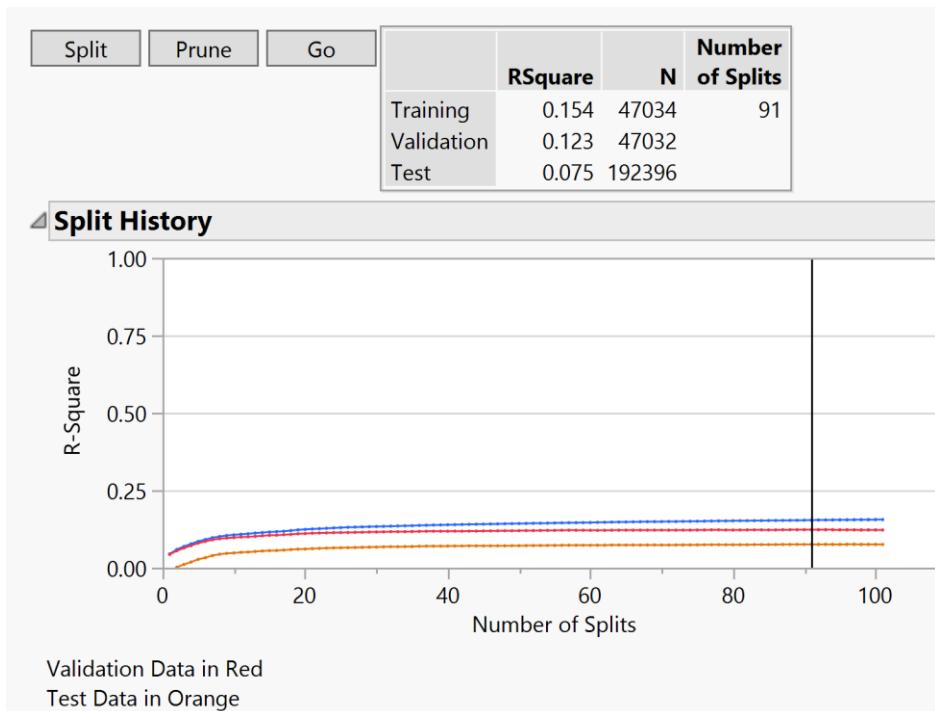
RowID	name	category	main_category	currency	currency rate	deadline	goal	launched	pledged	state	final state	backers	country	deadline year	launched year	goal usd	pledged usd	funding duration	ratio pledge to goal	ratio pledge to backers	Ratio Scale[backers]	Validation Stratified by final state	SortKey	RowNumber
1	0.3186813187	0	0	0.17509448	17100																			
2	0.1978021978	2.3974401e-6	0.00355	0.52312089	17101																			
3	0.3186813187	3.7193518e-7	0.005042	0.89409496	17102																			
4	0.3186813187	0	0	0.18762859	17103																			
5	0.3186813187	4.1639987e-6	0.0051083882	0.83149348	17104																			
6	0.3186813187	0	0	0.65852478	17105																			
7	0.3186813187	0	0	0.30784203	17106																			
8	0.3186813187	1.2959136e-6	0.1	0.36927241	17107																			
9	0.6153846154	1.9179521e-7	0.01	0.11682099	17108																			
10	0.3186813187	1.198201e-6	0.0041666667	0.51077155	17109																			
11	0.3186813187	0	0	0.14542792	17110																			
12	0.2197802198	1.3699658e-7	0.0005	0.67315707	17111																			
13	0.2087912088	8.790614e-10	0.00055	0.87029881	17112																			
14	0.2637362637	4.2770332e-7	0.0031857143	0.41577818	17113																			
15	0.9450549451	2.3441637e-7	0.00275	0.93803896	17114																			
16	0.3186813187	1.8893869e-7	0.0102888889	0.18038406	17115																			
17	0.2967032967	1.4192846e-6	0.000925	0.87951357	17116																			
18	0.967032967	5.5620611e-7	0.00725	0.34771159	17117																			
19	0.4835164835	1.2466689e-6	0.004875	0.72733438	17118																			
20	0.3186813187	0	0	0.334794	17119																			
21	0.3186813187	2.9968002e-8	0.0005	0.34441359	17120																			
22	0.4835164835	9.5097606e-8	0.0025	0.04308642	17121																			
23	0.3406593407	2.60362e-6	0.0011195876	0.83034582	17122																			
24	0.3186813187	9.3404268e-7	0.0088545455	0.26691124	17123																			

Sample size: 286,462

## Input variables:

The screenshot shows the 'Partition - JMP Pro' dialog. On the left, under 'Select Columns', there is a list of 31 columns: RowID, ID, name, category, main\_category, currency, currency rate, deadline, goal, launched, pledged, state, final state, backers, country, deadline year, launched year, goal usd, pledged usd, funding duration, ratio pledge to goal, ratio pledge to backers, Range Scale[backers], Range Scale[goal usd], Range Scale[pledged usd], Range Scale[funding duration], Range Scale[rat...pledge to goal], Range Scale[rat...dge to backers], Validation Stratified by final state, and SortKey. The 'Validation Stratified by final state' column is highlighted with a blue selection bar at the bottom. On the right, the 'Cast Selected Columns into Roles' section is displayed. It includes fields for 'Y, Response' (final state, optional), 'X, Factor' (category, main\_category, currency, deadline, launched, country, deadline year, launched year, Range Scale[goal usd], Range Sca... duration, optional numeric), 'Weight' (optional numeric), 'Freq' (optional numeric), 'Validation' (Validation S...y final state, optional), and 'By' (optional). Action buttons OK, Cancel, Remove, Recall, and Help are located on the far right.

## Results:



Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1536	0.1228	0.0754	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2557	0.2087	0.1290	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5867	0.6081	0.6076	$\sum -\text{Log}(\rho_{ij})/n$
RMSE	0.4493	0.4587	0.4594	$\sqrt{\sum (y_{ij}-\rho_{ij})^2/n}$
Mean Abs Dev	0.4039	0.4121	0.4121	$\sum  y_{ij}-\rho_{ij} /n$
Misclassification Rate	0.3178	0.3363	0.3417	$\sum (\rho_{ij} \neq \rho_{jMax})/n$
N	47034	47032	192396	n

Confusion Matrix									
		Training			Validation			Test	
		Predicted Count				Predicted Count			
Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted	Actual	Predicted
final state	Count	final state	Count	final state	Count	final state	Count	final state	Count
0	15442	0	8075	0	15114	0	77860	0	43987
1	6873	1	16644	1	7416	1	21746	1	48803

We are getting good accuracy 68.2% for training data set and 65.83% for test data set. This split provides same accuracy (65.83% = 65.83%) for test data set comparing the split “Train 0.4 Validation 0.4 Test 0.2”. Therefore, we keep the split “Train 0.4 Validation 0.4 Test 0.2” of “delete the live (state)” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	14948	68.22%
Validation	15818	66.37%
Test	65733	65.83%

b. Use Stratified split proportion: Training set - 0.4 Validation set - 0.4 Testing set - 0.2

2.ks-projects-201612-after conversion outliers inconsistency include all state - JMP Pro

The screenshot shows a JMP Pro interface with a data table titled "2.ks-projects-201612 aft...". A "Stratified Split - JMP Pro" dialog is open, containing a "Select Column" dropdown with various project metrics like state, final state, backers, country, deadline year, launched year, goal usd, pledged usd, funding duration, ratio pledge to goal, and ratio pledge to backers. Below it are sections for "Specify Data Proportions" (Training Set: 0.4, Validation Set: 0.4, Test Set: 0.2) and "Action" (OK, Cancel, Help). A "Select Focal Group" dropdown shows the value 0.

ID	name	category	main_category	currency	currency rate	deadline	goal	laur
1	68856463 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD		1 12/13/2011 4:46 PM	0.01	11/07/2C
2	62030213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD		1 12/04/2009 7:07 AM	0.01	11/25/20
3	9572984 Nana	Shorts	Film & Video	USD		1 03/16/2012 6:23 AM	0.15	01/25/20
4	219760504 RocknRoll NoisePollution	Documentary	Film & Video	USD		1 07/19/2011 3:59 PM	0.5	07/12/2C
5	1316126964 Antipledge	Performance Art	Art	USD		1 04/01/2010 6:00 PM	1	01/09/2C
6	1087894007 Lets Build Riders Company To Be Known Globally	Product Design	Design	GBP	1.28	11/16/2014 12:21 AM	1	09/11/2C
7	112629146 Till The End of Time	Nonfiction	Publishing	USD		1 11/11/2016 9:17 PM	1	09/12/2C
8	1493398444						1	05/17/2C
9	1494614224						1	04/12/20
10	1542552124						1	08/23/2C
11	1694496124						1	12/05/2C
12	2059637124						1	12/04/2C
13	2097081124						1	02/14/20
14	2140234124						1	03/03/2C
15	808941124						1	06/28/2C
16	294171124						1	07/10/2C
17	128494124						1	12/02/2C
18	204741124						1	12/15/2C
19	206445124						1	02/18/2C
20	1054378124						1	09/19/20
21	188239124						1	11/21/2C
22	135748124						1	12/26/2C
23	310938124						1	07/27/2C
24	378468728 Things to do in London (Canceled)	Zines	Publishing	GBP	1.28	02/29/2016 1:00 AM	1	01/28/2C
25	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...	Horror	Film & Video	USD	1	11/01/2014 7:59 AM	1	09/30/2C
26	944578907 Spore Animated (Canceled)	Comedy	Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/2C
27	1873533970 Game Development	Video Games	Games	USD	1	04/08/2015 11:30 PM	1	03/08/20
28	998486147 Half Blondie Series	Drama	Film & Video	USD	1	04/17/2016 4:57 AM	1	03/17/2C

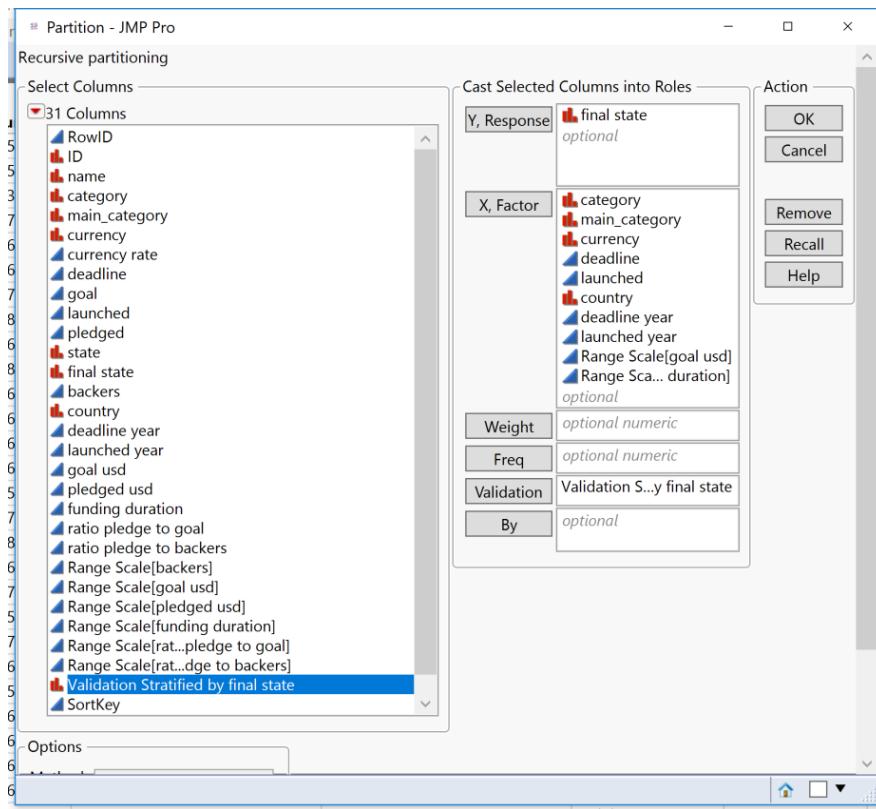
10.ks-projects-201612-after conversion outliers inconsistency include all state\_0.4&0.4&0.2 - JMP Pro

The screenshot shows a JMP Pro interface with a data table titled "10.ks-projects-201612 aft...". A "Stratified Split - JMP Pro" dialog is open, containing a "Select Column" dropdown with various project metrics like range scale, ratio scale, validation stratified by, sort key, and row number. Below it are sections for "Specify Data Proportions" (Training Set: 0.4, Validation Set: 0.4, Test Set: 0.2) and "Action" (OK, Cancel, Help). A "Select Focal Group" dropdown shows the value 0.

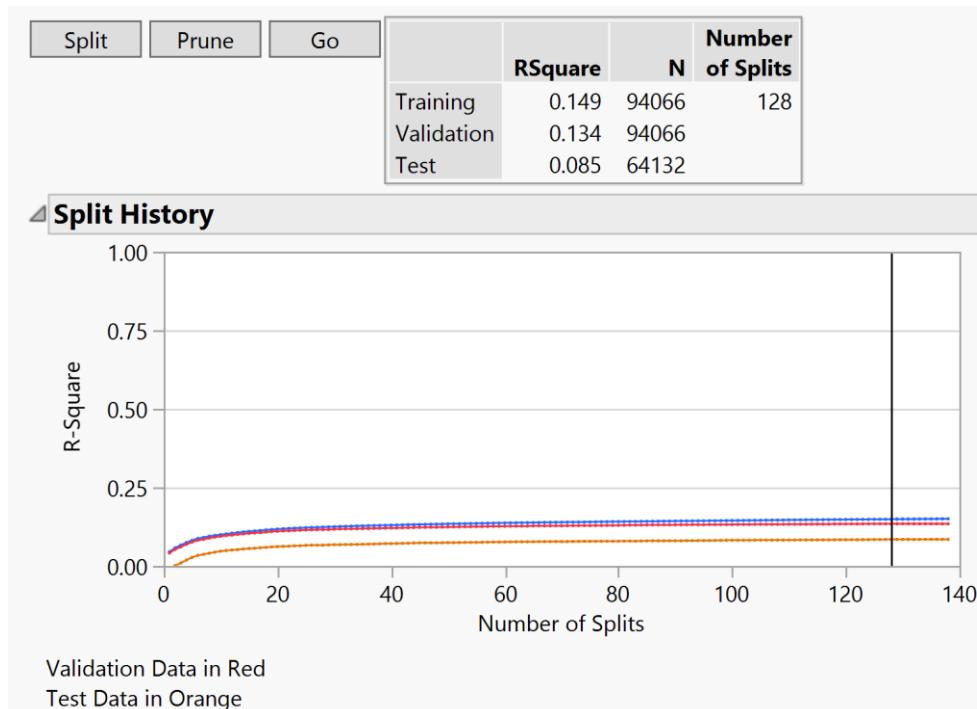
RowID	Range Scale[funding duration]	Range Scale[rat...pledge to goal]	Range Scale[rat...dge to backers]	Validation Stratified by ...	SortKey	RowNumber
1	0.3186813187	0	0	Training	0.02684093	34199
2	0.3186813187	4.79488e-10	0.0001	Training	0.66488748	34200
3	0.3186813187	1.0046416e-7	0.00275	Training	0.73869925	34201
4	0.3186813187	0	0	Training	0.03816636	34202
5	0.3186813187	5.1781707e-7	0.00054	Training	0.29356995	34203
6	0.4945054945	1.2633467e-6	0.0063789474	Training	0.70036406	34204
7	0.3736263736	0	0	Training	0.9443838	34205
8	0.4835164835	2.602935e-7	0.00168625	Training	0.16721263	34206
9	0.3406593407	0	0	Training	0.81016593	34207
10	0.6483516484	1.841234e-6	0.006	Training	0.99968047	34208
11	0.3186813187	0	0	Training	0.318087	34209
12	0.3186813187	3.8359042e-8	0.000339	Training	0.96180566	34210
13	0.3186813187	1.065529e-8	0.005	Training	0.62372626	34211
14	0.3186813187	6.7275859e-7	0.0024	Training	0.63025823	34212
15	0.1426571429	1.9179521e-7	0.002	Training	0.11591011	34213
16	0.1978021978	0	0	Training	0.52773466	34214
17	0.6483516484	2.4281274e-6	0.0068432432	Training	0.57989324	34215
18	0.1648351648	8.2197948e-8	0.0003	Training	0.86837071	34216
19	0.3186813187	1.9179521e-7	0.0007333333	Training	0.97017073	34217
20	0.6483516484	3.8359042e-8	0.0076	Training	0.71132551	34218
21	0.6263736264	0	0	Training	0.70857475	34219
22	0.3186813187	2.3974401e-8	0.0025	Training	0.13205792	34220
23	0.3186813187	2.075242e-6	0.0104709677	Training	0.08421387	34221
24	0.3186813187	5.0346243e-8	0.0021	Training	0.11444272	34222
25	0.3186813187	3.2669118e-6	0.0023227273	Training	0.08302784	34223
26	0.1426571429	9.5897606e-9	0.0005	Training	0.89473711	34224
27	0.3736263736	1.5535412e-6	0.003456	Training	0.96636023	34225
28	0.3186813187	1.775929e-6	0.003264	Training	0.83685534	34226

Sample size: 252,264

Input variables:



### Results:



Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1489	0.1344	0.0846	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2487	0.2266	0.1439	$(1 - (L(0)/L(\text{model})))^{(2/n)}/(1 - L(0)^{(2/n)})$
Mean -Log p	0.5899	0.6000	0.6016	$\sum -\text{Log}(\rho_{ij})/n$
RMSE	0.4508	0.4554	0.4567	$\sqrt{\sum (y_{ij} - \rho_{ij})^2/n}$
Mean Abs Dev	0.4065	0.4110	0.4111	$\sum  y_{ij} - \rho_{ij} /n$
Misclassification Rate	0.3183	0.3280	0.3353	$\sum (\rho_{ij} \neq \rho_{\text{Max}})/n$
N	94066	94066	64132	n

Confusion Matrix					
		Training	Validation	Test	
Actual	Predicted Count				
		Actual	Predicted Count		
final state	0 1	final state	0 1	final state	0 1
0	30597 16436	0	30266 16767	0	26161 14455
1	13502 33531	1	14085 32948	1	7051 16465

We are getting good accuracy 68.2% for training data set and 66.47% for test data set. This split provides more accuracy (66.47% > 65.83%) for test data set comparing the split “Train 0.4 Validation 0.4 Test 0.2”. But the difference is small, it’s more reasonable to use the data set which delete the live (state). Therefore, we keep the split “Train 0.4 Validation 0.4 Test 0.2” of “delete the live (state)” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	29938	68.17%
Validation	30852	67.20%
Test	21506	66.47%

3. We used the data set “delete live and undefined (state)” to build the model by using the decision tree model

a. Use Stratified split proportion: Training set - 0.4 Validation set - 0.4 Testing set - 0.2

File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

4.ks-projects-201612-after conversion outliers inconsistency and delete live and undefined - JMP Pro

ID	name	category	main_category	currency	currency rate	deadline	goal	lat
1	688564643 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD		1 12/13/2011 4:46 PM	0.01	11/07/2
2	620302213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD		1 12/04/2009 7:07 AM	0.01	11/25/2
3	9572984 Nana	Shorts	Film & Video	USD		1 03/16/2012 6:23 AM	0.15	01/25/2

Stratified Split - JMP Pro

Select Column

- deadline
- goal
- launched
- pledged
- state
- final state**
- backers
- country
- deadline year
- launched year

Specify Data Proportions

Action

OK Cancel Help

Select Focal Group

0 1

Rows

All rows 315,539  
Selected 0  
Excluded 0  
Hidden 0  
Labelled 0

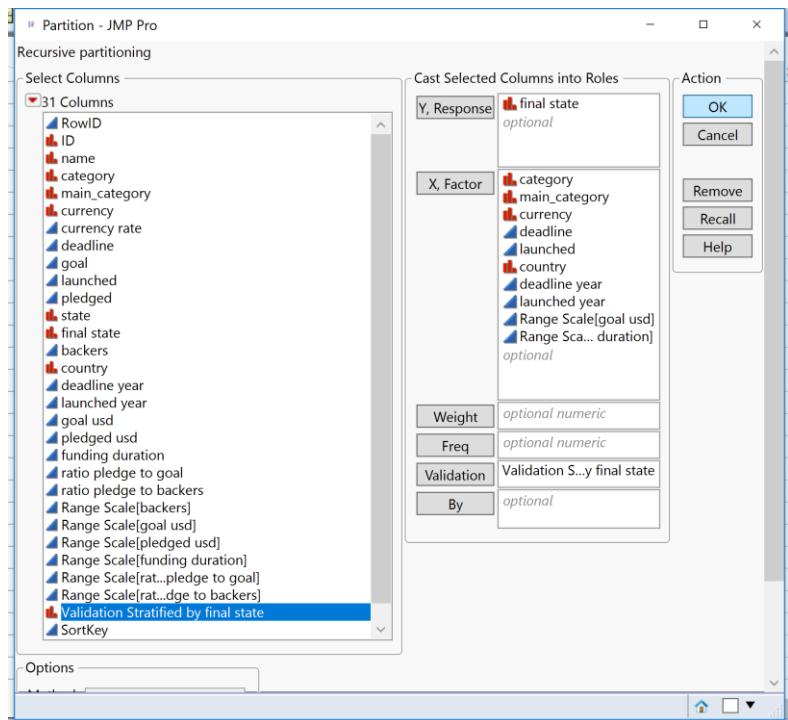
File Edit Tables Rows Cols DOE Analyze Graph Tools Add-Ins View Window Help

4.ks-projects-201612-after conversion outliers inconsistency and delete live and undefined\_1\_PROP\_0.5 - JMP Pro

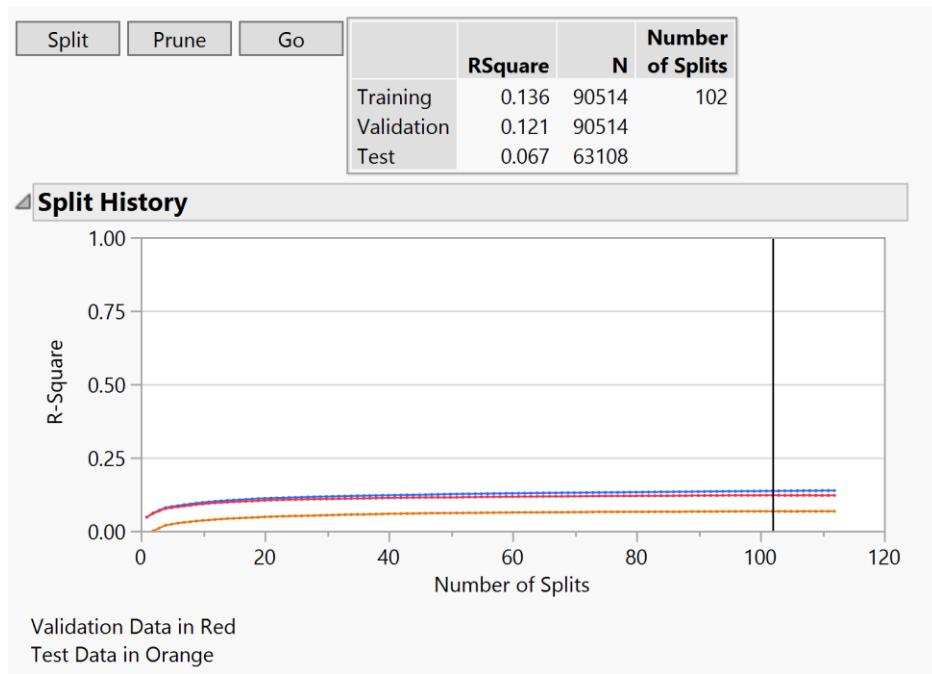
RowID	ID	name	category	main_category	currency	currency rate	deadline	goal	launched	pledged	state	f
1	99730	993547680 Help Brit House ...	Country & Folk	Music	USD	1 05/31/2015 7:53 ...	3000	04/16/2015 7:53 ...	0	failed	0	
2	135295	742745978 Australian ...	Web	Journalism	GBP	1 10/03/2014 10:21...	5000	09/03/2014 10:21...	0	cancelled	0	
3	189831	925428015 A Brooklyn Project	Poetry	Publishing	USD	1 04/10/2014 10:19...	9000	02/17/2014 10:3 ...	66	failed	0	
4	207049	4202417367 Limitless LP ...	Pop	Music	USD	1 05/14/2014 2:00 ...	10000	04/04/2016 10:50...	3440	cancelled	0	
5	243836	1284019559 "THE RISE" ...	Webseries	Film & Video	USD	1 07/27/2011 10:07...	20000	06/27/2011 10:07...	0	cancelled	0	
6	217134	740966921 "The Brethren" ...	Theater	Theater	USD	1 08/08/2010 8:14 ...	12000	06/22/2010 11:55...	0	cancelled	0	
7	129514	801839198 Make It With ...	Documentary	Film & Video	USD	1 07/31/2014 8:30 ...	4200	07/01/2014 8:30 ...	355	failed	0	
8	61483	1362486473 Scruffy: From ...	Children's Books	Publishing	USD	1 02/01/2015 4:36 ...	1500	01/02/2015 4:40 ...	51	failed	0	
9	135566	1219706624 Book Promotion ...	Print	Journalism	USD	1 09/07/2014 11:28...	5000	08/08/2014 11:28...	0	failed	0	
10	89382	1425447797 My Very First ...	Hip-Hop	Music	USD	1 05/16/2015 2:59 ...	2500	03/17/2015 1:59 ...	27	failed	0	
11	312966	1044026253 SISTERS OF THE ...	Horror	Film & Video	USD	1 03/27/2015 5:31 ...	500000	01/28/2015 6:35 ...	0	failed	0	
12	163099	1394860437 "Tucker Miles" ...	Fiction	Publishing	USD	1 04/09/2013 6:20 ...	6000	03/10/2013 6:20 ...	0	failed	0	
13	227071	1553208242 Where the Sun ...	Narrative Film	Film & Video	USD	1 06/30/2016 5:59 ...	15000	05/02/2016 4:41 ...	1	failed	0	
14	237806	994732368 Johanju	Gadgets	Technology	USD	1 08/08/2015 8:30 ...	16000	07/09/2015 8:30 ...	1	failed	0	
15	253993	1140886200 The Dying Room	Publishing	USD	1 11/15/2015 6:00 ...	21000	10/12/2015 11:16...	103	failed	0		
16	301738	935147987 Misaro - Fashion ...	Accessories	Fashion	AUD	0.71 10/29/2015 12:38...	100000	09/29/2015 1:38 ...	0	failed	0	
17	259385	1424119370 "NOT A ...	Theater	Theater	USD	1 06/28/2010 8:35 ...	25000	05/28/2010 9:55 ...	125	failed	0	
18	113188	1870485121 Construction ...	Playing Cards	Games	USD	1 11/16/2014 4:17 ...	3200	10/02/2014 4:17 ...	1223	failed	0	
19	134580	1060501589 I'm wanting ...	Apps	Technology	USD	1 02/14/2015 7:00 ...	5000	12/20/2014 5:47 ...	0	failed	0	
20	298756	115145485 Duluth MN ...	Web	Journalism	USD	1 06/02/2016 5:19 ...	80000	04/03/2016 5:19 ...	1	cancelled	0	
21	125356	60156284 Capturing Hope: ...	Photography	Photography	USD	1 04/07/2014 4:06 ...	4000	03/08/2014 4:06 ...	856	failed	0	
22	268469	1220090730 Le Sud. (Cancelled)	Restaurants	Food	USD	1 05/19/2016 9:49 ...	30000	04/19/2016 9:49 ...	25	cancelled	0	
23	125660	2104369294 SOMEPLACE ...	Shorts	Film & Video	GBP	1.28 03/03/2014 6:42 ...	4000	02/01/2014 6:42 ...	1271.3	failed	0	
24	73518	1647647552 Help Our Small ...	Crafts	Crafts	USD	1 05/04/2015 4:01 ...	2000	04/04/2015 4:01 ...	2	failed	0	
25	200634	1389932916 Fringe Snipper: ...	Accessories	Fashion	USD	1 05/25/2015 2:53 ...	10000	04/25/2015 2:53 ...	110	failed	0	
26	172435	1136792377 Artist Studio Tour	Art	Art	USD	1 07/27/2016 11:33 ...	7000	06/27/2016 11:33 ...	0	failed	0	
27	28294	1410852640 Space Squid ...	Video Games	Games	CAD	0.76 10/13/2014 9:10 ...	600	09/22/2014 9:10 ...	0	cancelled	0	
28	91567	1229586834 The North ...	Photography	Photography	USD	1 06/04/2014 10:46 ...	2500	04/20/2014 10:46 ...	390	failed	0	

Sample size: 244,136

Input variables:



### Results:



Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1359	0.1213	0.0665	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.2289	0.2063	0.1141	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.5990	0.6091	0.6092	$\sum -\log(p_{ij})/n$
RMSE	0.4547	0.4595	0.4601	$\sqrt{\sum(y_{ij}-\hat{y}_{ij})^2/n}$
Mean Abs Dev	0.4135	0.4181	0.4175	$\sum  y_{ij}-\hat{y}_{ij} /n$
Misclassification Rate	0.3262	0.3359	0.3383	$\sum (\hat{p}_{ij} \neq p_{Max})/n$
N	90514	90514	63108	n

Confusion Matrix				
		Training		Validation
		Actual		Predicted Count
Actual		0	1	
final state		0	1	
0		29689	15568	29370 15887
1		13962	31295	14513 30744

		Test	
		Actual	
		0	1
Actual		0	1
final state		0	1
0		26412	14067
1		7284	15345

We are getting good accuracy 67.4% for training data set and 66.17% for test data set. This split provides more accuracy (66.17% > 65.83%) for test data set comparing the split “Train 0.4 Validation 0.4 Test 0.2”. But the difference is small, it’s more reasonable to use the data set which delete the live (state). Therefore, we keep the split “Train 0.4 Validation 0.4 Test 0.2” of “delete the live (state)” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	29530	67.38%
Validation	30400	66.41%
Test	21351	66.17%

4. We used the data set “delete live, undefined and suspended (state)” to build the model by using the decision tree model

a. Use Stratified proportion: Training set - 0.4 Validation set - 0.4 Testing set - 0.2

5.xls-projects-201612-after conversion outliers inconsistency and delete live and undefined and suspend - JMP Pro

Stratified Split - JMP Pro

Select Column: final state

Specify Data Proportions:

- Training Set: 0.4
- Validation Set: 0.4
- Test Set: 0.2

Action: OK, Cancel, Help

Select Focal Group: 0, 1

ID	name	category	main_category	currency	currency rate	deadline	goal	laur	
1	68856463 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/20	
2	62030213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/20	
				USD	1	03/16/2012 6:23 AM	0.15	01/25/20	
				USD	1	07/19/2011 3:59 PM	0.5	07/12/20	
				USD	1	04/01/2010 6:00 PM	1	01/09/20	
				GBP	1.28	11/16/2014 12:21 AM	1	09/11/20	
				USD	1	11/11/2016 9:17 PM	1	09/12/20	
				USD	1	07/16/2012 8:10 PM	1	05/17/20	
				USD	1	06/11/2015 1:25 AM	1	04/12/20	
				USD	1	10/22/2016 4:56 PM	1	08/23/20	
				USD	1	02/02/2015 12:11 AM	1	12/04/20	
				USD	1	04/14/2012 1:26 AM	1	02/14/20	
				USD	1	05/02/2015 6:25 PM	1	03/03/20	
				USD	1	08/22/2011 10:28 PM	1	06/28/20	
				USD	1	09/02/2011 6:00 AM	1	07/10/20	
				USD	1	01/21/2016 11:20 PM	1	12/02/20	
				EUR	1.13	01/29/2016 4:23 PM	1	12/15/20	
				USD	1	04/03/2015 12:36 AM	1	02/18/20	
				USD	1	10/30/2014 6:25 AM	1	09/19/20	
				USD	1	12/31/2015 9:54 PM	1	11/21/20	
				USD	1	01/31/2016 1:55 PM	1	12/26/20	
				USD	1	08/29/2016 7:36 PM	1	07/21/20	
				Zines	1.28	02/29/2016 1:00 AM	1	01/28/20	
				GBP	1.28	02/29/2016 1:00 AM	1	01/28/20	
				Horror	1	11/01/2014 7:59 AM	1	09/30/20	
				Film & Video	USD	1	06/12/2015 8:38 PM	1	05/11/20
				Comedy	USD	1	04/08/2015 11:30 PM	1	03/08/20
				Video Games	USD	1	04/17/2016 4:57 AM	1	03/17/20
				Drama	USD	1	10/19/2016 5:15 PM	1	09/19/20
				Film & Video	GBP	1.28	10/19/2016 5:15 PM	1	09/19/20

12.xls-projects-201612-after conversion outliers inconsistency and delete live and undefined and suspend\_04&04&02 - JMP Pro

Stratified Split - JMP Pro

Select Column: final state

Specify Data Proportions:

- Training Set: 0.4
- Validation Set: 0.4
- Test Set: 0.2

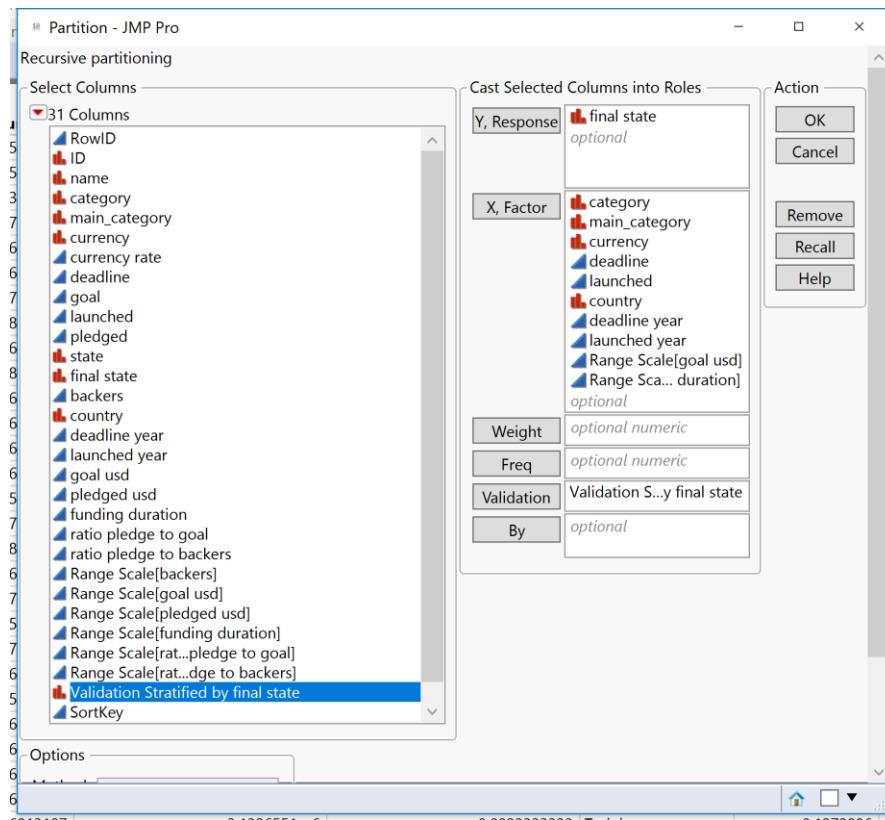
Action: OK, Cancel, Help

Select Focal Group: 0, 1

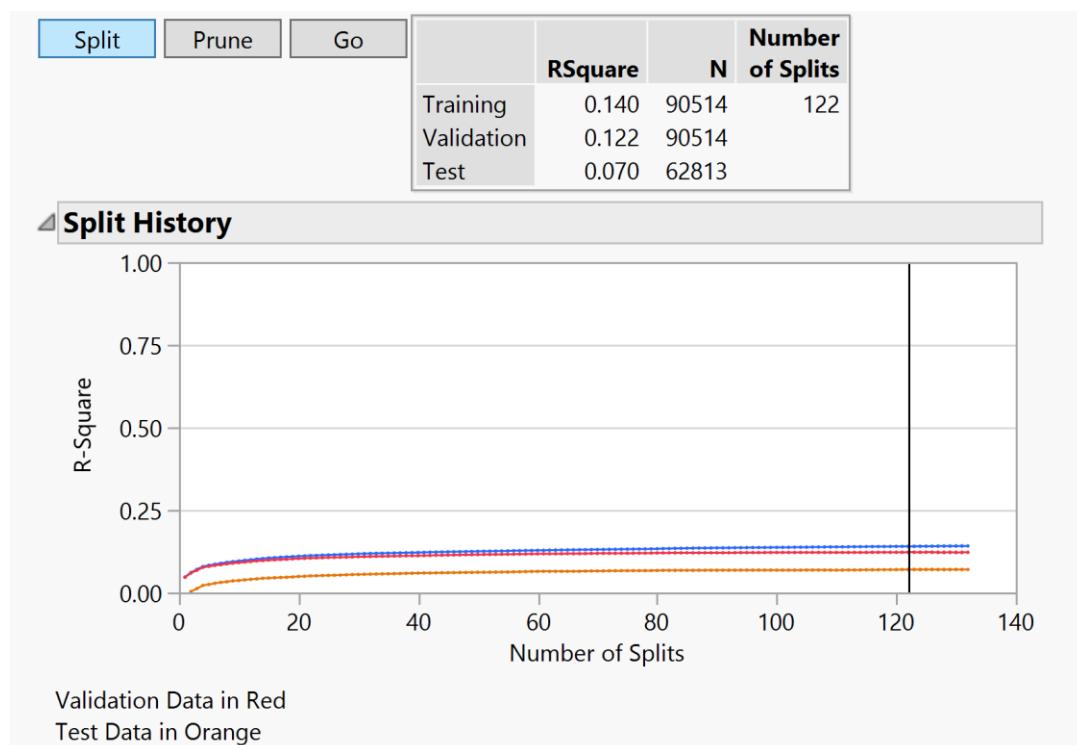
Range Scale[funding duration]	Range Scale[rat...pledge to goal]	Range Scale[rat...dge to backers]	Validation Stratified by ...	SortKey	RowNumber
1	0.3186813187	0	0 Training	0.86492936	35112
2	0.3186813187	0	0 Training	0.02169268	35113
3	0.4835164835	2.3319862e-6	0.0051066667 Training	0.47123209	35114
4	0.3186813187	1.1720818e-8	0.00055 Training	0.27066201	35115
5	0.4285714286	1.450575e-7	0.0073692308 Training	0.69133637	35116
6	0.3186813187	4.6166706e-6	0.0186354839 Training	0.46676744	35117
7	0.3956043956	3.7280194e-6	0.0020733333 Training	0.63354848	35118
8	0.6483516484	3.2605186e-8	0.0028815 Training	0.16213681	35119
9	0.1428571429	4.1347851e-6	0.0051533865 Training	0.19599447	35120
10	0.2197802198	3.5878491e-6	0.0333766316 Training	0.29057625	35121
11	0.6153846154	3.963767e-9	0.0062 Training	0.59119347	35122
12	0.3076923077	9.6377094e-7	0.00201 Training	0.97306451	35123
13	0.3186813187	0	0 Training	0.11085677	35124
14	0.3186813187	1.3655819e-6	0.0032363636 Training	0.52887268	35125
15	0.2637362637	1.2791265e-6	0.0066692308 Training	0.39015702	35126
16	0.3186813187	0	0 Training	0.20349824	35127
17	0.6483516484	1.3425665e-8	0.00035 Training	0.87733563	35128
18	0.3186813187	0	0 Training	0.87054656	35129
19	0.6373626374	9.7895473e-7	0.00765625 Training	0.74428064	35130
20	0.3626373626	2.3636249e-7	0.0020857143 Training	0.61223675	35131
21	0.6483516484	0	0 Training	0.7927692	35132
22	0.3186813187	5.9936004e-8	0.005 Training	0.31332113	35133
23	0.3186813187	0	0 Training	0.032244	35134
24	0.6483516484	8.4252897e-7	0.0055909091 Training	0.8499269	35135
25	0.6483516484	4.7948803e-9	0.0001 Training	0.20544235	35136
26	0.3186813187	5.5141123e-7	0.0063085714 Training	0.67081064	35137
27	0.3186813187	0	0 Training	0.8316216	35138
28	0.3186813187	0	0 Training	0.5688312	35139

Sample size: 243,841

Input variables:



### Results:



#### Fit Details

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1398	0.1221	0.0698	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2349	0.2076	0.1196	$(1 - L(0)/L(\text{model}))^{(2/n)}/(1 - L(0)^{(2/n)})$
Mean -Log p	0.5963	0.6085	0.6079	$\sum -\text{Log}(p[j])/n$
RMSE	0.4537	0.4590	0.4598	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4118	0.4168	0.4163	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3254	0.3350	0.3364	$\sum (p[j] \neq p_{\text{Max}})/n$
N	90514	90514	62813	n

#### Confusion Matrix

Training		Validation		Test		
Actual		Predicted Count		Actual		
final state		0	1	final state	0	1
0	30355	14902	0	29941	15316	
1	14555	30702	1	15005	30252	

Actual		Predicted Count		Actual		
final state		0	1	final state	0	1
0	26588	13596	0	7532	15097	
1						

We are getting good accuracy 67.5% for training data set and 66.36% for test data set. This split provides more accuracy (66.36% > 65.83%) for test data set comparing the split “Train 0.4 Validation 0.4 Test 0.2”. But the difference is small, we keep the split “Train 0.4 Validation 0.4 Test 0.2” of “delete the live (state)” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	29,457	67.46%
Validation	30,321	66.50%
Test	21,128	66.36%

5. We used the data set “all (state)” and “undefined is successful” to build the model by using the decision tree model

a. Use Stratified split proportion: Training set - 0.4 Validation set - 0.4 Testing set - 0.2

Screenshot of JMP Pro software interface showing two windows for data analysis.

**Window 1: Recode - state - JMP Pro**

This window shows a recode dialog for the "state" column. The "New Column" dropdown is set to "final state". The "Count Old Values (6)" section lists: failed (16852), canceled (3298), suspended (1476), successful (113143), live (4439), and undefined (683). The "New Values (2)" dropdown is set to 0 and 1. A "Group controls" section contains a checked checkbox for "View Groups". The "Latest change: undefined ->" field is empty. Buttons at the bottom include "Recode", "Close", and "Help".

deadline year	launched year	goal usd	pledged usd	funding duration	ratio pledge to goal	ratio pledged to backer
2011	2011	0.01	0	36	0	16
2009	2009	0.01	100	9	10000	0
2012	2012	0.15	0	51	0	0
2011	2011	0.5	0	7	0	0
2010	2010	1	0	82	0	0
2014	2014	1.28	0	60	0	0
2016	2016	1	0	60	0	0
2012	2012	1	0	60	0	0
2015	2015	1	0	60	0	0
2016	2016	1	0	60	0	0
2017	2016	1	0	60	0	0
2015	2014	1	0	60	0	0
2012	2012	1	0	60	0	0
2015	2015	1	0	60	0	0
2011	2011	1	0	55	0	0
2011	2011	1	0	54	0	0
2016	2015	1	0	50	0	0
2016	2015	1.13	0	45	0	0
2015	2015	1	0	44	0	0
2014	2014	1	0	41	0	0
2015	2015	1	0	40	0	0
2016	2015	1	0	36	0	0
2016	2016	1	0	33	0	0
2015	2015	1	0	32	0	0
2015	2015	1	0	32	0	0

**Window 2: Stratified Split - JMP Pro**

This window shows a stratified split dialog for the "state" column. The "Select Column" dropdown is set to "state". The "Specify Data Proportions" section has "Training Set" at 0.4, "Validation Set" at 0.4, and "Test Set" at 0.2. The "Action" buttons are "OK", "Cancel", and "Help". The "Select Focal Group" dropdown is set to 0 and 1. Buttons at the bottom include "evaluations done".

ID	name	category	main_category	currency	currency rate	deadline	goal	lat
1	688564643 Word-of-mouth publishing: get "Corruptions" out ...	Fiction	Publishing	USD	1	12/13/2011 4:46 PM	0.01	11/07/2012
2	620302213 LOVELAND Round 6: A Force More Powerful	Conceptual Art	Art	USD	1	12/04/2009 7:07 AM	0.01	11/25/2012
3	9572984 Nana	Shorts	Film & Video	USD	1	03/16/2012 6:23 AM	0.15	01/25/2012
19	2064457704 Fare thee well GRATEFUL DEAD documentary	Photobooks	Photography	USD	1	07/19/2011 3:59 PM	0.5	07/12/2012
20	1054378026 Reece Ran's WINTER	Fiction	Publishing	USD	1	04/01/2010 6:00 PM	1	01/09/2012
21	1882394705 JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...	Fiction	Publishing	USD	1	11/16/2014 12:21 AM	1	09/17/2012
22	1357482430 .cifl (Cancelled)	Tabletop Games	Games	USD	1	11/11/2016 9:17 PM	1	09/12/2012
23	310936687 SMILE #UWokeUp	Nonfiction	Publishing	USD	1	07/16/2012 8:10 PM	1	05/17/2012
24	378468728 Things to do in London (Cancelled)	Zines	Publishing	GBP	1	06/11/2015 1:25 AM	1	04/12/2012
25	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...	Horror	Film & Video	USD	1	10/22/2016 4:56 PM	1	08/23/2012
26	944578907 Spore Animated (Cancelled)	Comedy	Film & Video	USD	1	02/03/2017 11:34 PM	1	12/05/2012
19	2064457704 Fare thee well GRATEFUL DEAD documentary	Photobooks	Photography	USD	1	02/02/2015 12:11 AM	1	12/04/2012
20	1054378026 Reece Ran's WINTER	Fiction	Publishing	USD	1	04/14/2012 1:26 AM	1	02/14/2012
21	1882394705 JOE'S ERNEST HEMINGWAY STORY , BACK AGAIN ...	Fiction	Publishing	USD	1	05/02/2015 6:25 PM	1	03/03/2012
22	1357482430 .cifl (Cancelled)	Tabletop Games	Games	USD	1	08/22/2011 10:28 PM	1	06/28/2012
23	310936687 SMILE #UWokeUp	Nonfiction	Publishing	USD	1	09/02/2011 6:00 AM	1	07/10/2012
24	378468728 Things to do in London (Cancelled)	Zines	Publishing	GBP	1	01/21/2016 11:20 PM	1	12/02/2012
25	515995141 THE SOPHIA MOVIE - A Suspense/Horror Film ...	Horror	Film & Video	USD	1	01/29/2016 4:23 PM	1	12/15/2012
26	944578907 Spore Animated (Cancelled)	Comedy	Film & Video	USD	1	04/03/2015 12:36 AM	1	02/18/2012

JMP Pro window showing a data table titled "14.ks-projects-201612-after conversion outliers inconsistency include all state (undefined is succesful)\_0.48&0.4&0.2 - JMP Pro".

The table has columns: RowID, ID, name, category, main\_category, currency, currency rate, deadline, goal, launched, pledged, state, final state, backers, country, deadline year, launched year, Range Scale[rat...pledge to goal], Range Scale[rat...dge to backers], Validation Stratified by ..., SortKey, RowNumber, Prob(final state==0), and Prob(final state==1).

Sample size: 253,356

All Rows

Sample size: 253,356

Input variables:

Partition - JMP Pro

Recursive partitioning

Select Columns

31 Columns

- RowID
- ID
- name
- category
- main\_category
- currency
- currency rate
- deadline
- goal
- launched
- pledged
- state
- final state
- backers
- country
- deadline year
- launched year
- goal usd
- pledged usd
- funding duration
- ratio pledge to goal
- ratio pledge to backers
- Range Scale[backers]
- Range Scale[goal usd]
- Range Scale[pledged usd]
- Range Scale[funding duration]
- Range Scale[rat...pledge to goal]
- Range Scale[rat...dge to backers]
- Validation Stratified by final state
- SortKey

Cast Selected Columns into Roles

Y, Response: final state  
optional

X, Factor: category, main\_category, currency, deadline, launched, pledged, state, country, deadline year, launched year, Range Scale[goal usd], Range Scale[duration]  
optional

Weight: optional numeric

Freq: optional numeric

Validation: Validation S...y final state

By: optional

Action

OK

Cancel

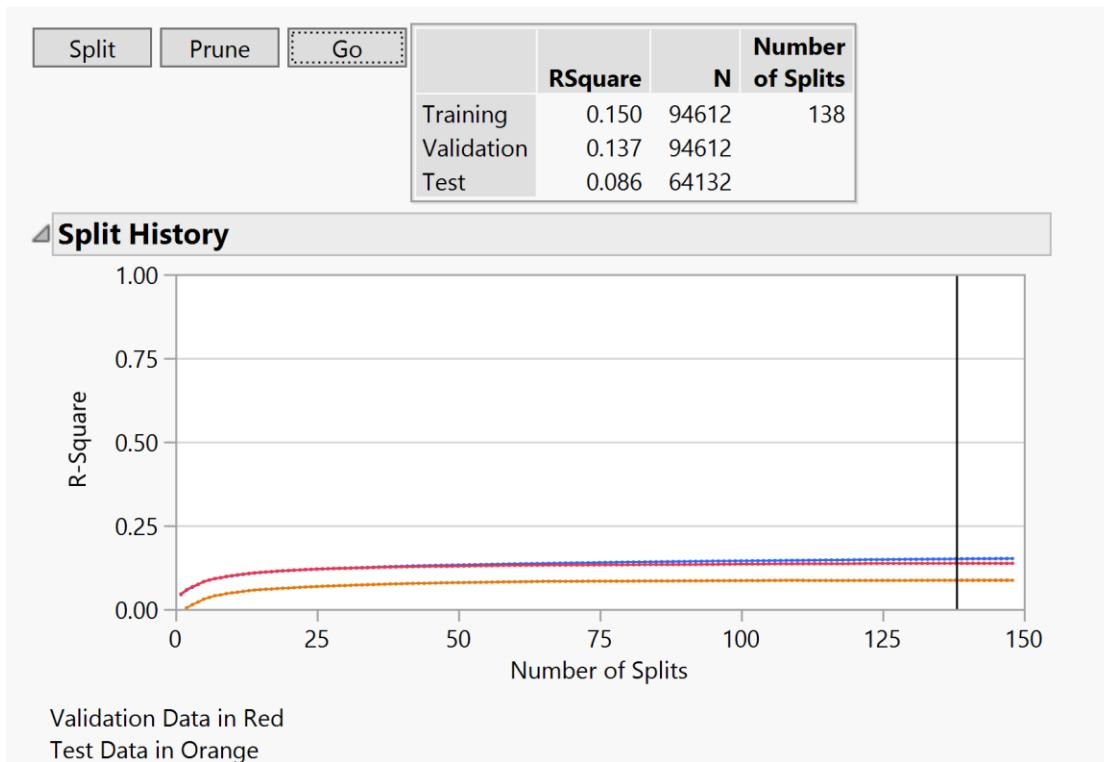
Remove

Recall

Help

Options

Results:



**Fit Details**

Measure	Training	Validation	Test	Definition
Entropy RSquare	0.1496	0.1365	0.0859	$1 - \text{Loglike}(\text{model})/\text{Loglike}(0)$
Generalized RSquare	0.2498	0.2299	0.1460	$(1 - (L(0)/L(\text{model}))^{(2/n)})/(1 - L(0)^{(2/n)})$
Mean -Log p	0.5894	0.5985	0.6018	$\sum -\text{Log}(p[j])/n$
RMSE	0.4508	0.4545	0.4566	$\sqrt{\sum (y[j] - p[j])^2/n}$
Mean Abs Dev	0.4067	0.4103	0.4111	$\sum  y[j] - p[j] /n$
Misclassification Rate	0.3219	0.3263	0.3284	$\sum (p[j] \neq p_{\text{Max}})/n$
N	94612	94612	64132	n

**Confusion Matrix**

Training		Validation		Test	
Actual	Predicted Count	Actual	Predicted Count	Actual	Predicted Count
final state	0 1	final state	0 1	final state	0 1
0	32016 15290	0	31778 15528	0	27153 13326
1	15161 32145	1	15348 31958	1	7737 15916

We are getting good accuracy 67.8% for training data set and 67.16% for test data set. This split provides more accuracy (67.16% > 65.83%) for test data set comparing the split “Train 0.4 Validation 0.4 Test 0.2”. But the difference is small, we keep the split “Train 0.4 Validation 0.4 Test 0.2” of “delete the live (state)” to build the model.

Data set	Number of False positive and False Negative Errors	Accuracy
Training	30,451	67.81%
Validation	30,876	67.37%
Test	21,063	67.16%

### Comparison of Decision Tree model for every project status

Data Set	Training-Validation-Test proportion	Sample Size	Number of splits	Training				
				Predict 1, Actual 0	Predict 0, Actual 1	Total Error	Total	Ratio
delete observations with status as live	0.3-0.3-0.4	262,261	120	12,241	9,611	21,852	67,886	67.81%
delete observations with status as live	0.2-0.2-0.6	280,247	97	7,979	6,735	14,714	45,256	67.49%
delete observations with status as live	0.4-0.4-0.2	244,273	135	16,071	13,386	29,457	90,514	67.46%
Consider status undefined as unsuccesful	0.2-0.2-0.6	286,462	91	8,075	6,873	14,948	47,034	68.22%
Consider status undefined as unsuccesful	0.4-0.4-0.2	252,264	128	16,436	13,502	29,938	94,066	68.17%
delete status - live and undefined	0.4-0.4-0.2	244,136	102	15,568	13,962	29,530	90,514	67.38%
delete status - live, undefined and suspended	0.4-0.4-0.2	243,841	122	14,902	14,555	29,457	90,514	67.46%
Change status undefined as succesful	0.4-0.4-0.2	253,356	138	15,290	15,161	30,451	94,612	67.81%

Data Set	Training-Validation-Test proportion	Sample Size	Number of splits	Validation				
				Predict 1, Actual 0	Predict 0, Actual 1	Total Error	Total	Ratio
delete observations with status as live	0.3-0.3-0.4	262,261	120	12,808	9,895	22,703	67,886	66.56%
delete observations with status as live	0.2-0.2-0.6	280,247	97	8,435	6,894	15,329	45,258	66.13%
delete observations with status as live	0.4-0.4-0.2	244,273	135	16,521	13,708	30,229	90,514	66.60%
Consider status undefined as unsuccesful	0.2-0.2-0.6	286,462	91	8,402	7,416	15,818	47,032	66.37%
Consider status undefined as unsuccesful	0.4-0.4-0.2	252,264	128	16,767	14,085	30,852	94,066	67.20%
delete status - live and undefined	0.4-0.4-0.2	244,136	102	15,887	14,513	30,400	90,514	66.41%
delete status - live, undefined and suspended	0.4-0.4-0.2	243,841	122	15,316	15,005	30,321	90,514	66.50%
Change status undefined as succesful	0.4-0.4-0.2	253,356	138	15,528	15,348	30,876	94,612	67.37%

Data Set	Training-Validation-Test proportion	Sample Size	Number of splits	Test				
				Predict 1, Actual 0	Predict 0, Actual 1	Total Error	Total	Ratio
delete observations with status as live	0.3-0.3-0.4	262,261	120	30,505	13,579	44,084	126,489	65.15%
delete observations with status as live	0.2-0.2-0.6	280,247	97	45,119	21,032	66,151	189,733	65.13%
delete observations with status as live	0.4-0.4-0.2	244,273	135	14,693	6,915	21,608	63,245	65.83%
Consider status undefined as unsuccesful	0.2-0.2-0.6	286,462	91	43,987	21,746	65,733	192,396	65.83%
Consider status undefined as unsuccesful	0.4-0.4-0.2	252,264	128	14,455	7,051	21,506	64,132	66.47%
delete status - live and undefined	0.4-0.4-0.2	244,136	102	14,067	7,284	21,351	63,108	66.17%
delete status - live, undefined and suspended	0.4-0.4-0.2	243,841	122	13,596	7,532	21,128	62,813	66.36%
Change status undefined as succesful	0.4-0.4-0.2	253,356	138	13,326	7,737	21,063	64,132	67.16%

**RESULT** - Thus, from the test observation we can conclude that the modeling results we get when considering status live, undefined, suspended as successful is approximately the same as results obtained when dropping or deleting observations with status live, undefined, suspended. Therefore, there is no significant differences in accuracy and number of errors obtained from each of these cases.