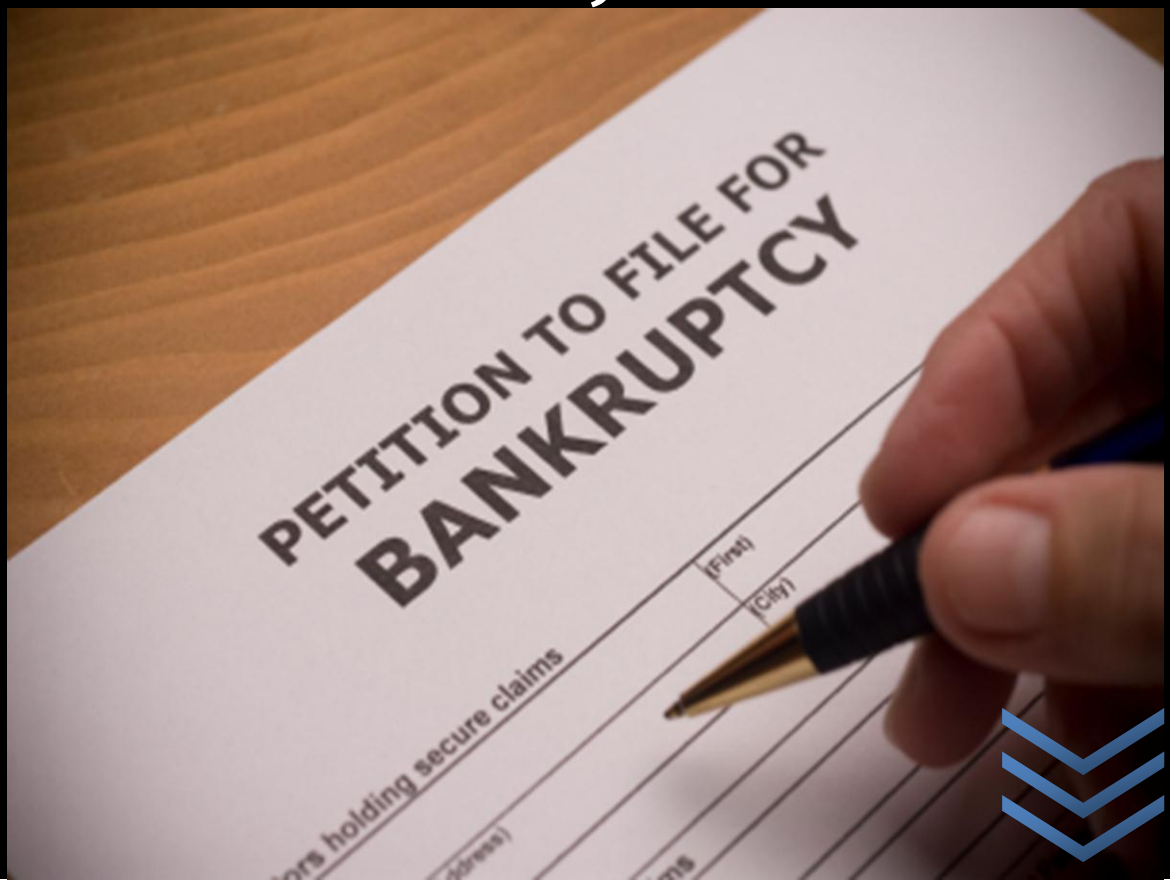
BIDM Project



Group A2

This project intends to help VCs in predicting if a start up will become bankrupt or not

Group Members:

Tressa Joy-61110480

Kapila Monga - 61110400

Rishi Raj Singh- 61110304

Smitha Purohit - 61110633

Srividya Varanasi-61110322

Vivek Vikram Singh- 61110628

Executive summary

"US venture capital investments sank 61% in Q1 2009, dropping to the lowest level in 12 years. VC investments totaled $3bn in Q1 2009 whereas in Q1 2008, investments totaled $7.74bn" - PriceWaterhouseCoopers, the National VC Association and Thomson Reuters.

Keeping in mind this high opportunity cost, it becomes extremely important for the VCs to choose the companies to invest in very cautiously. VCs get their cash out only when your business is acquired by another company or "goes public," that is, when its shares can be publicly traded on a stock exchange.

Through this project we strive to develop a model based on past data for the VCs to **predict whether a particular company seeking VC funding would file for a bankruptcy/become defunct in future or not** in order to assist the VCs in their decision making.

Problem description

Problem statement: ***Predict if a company seeking investment from VC would become bankrupt/defunct in the future.***

The problem addressed here is to device a model for the VCs to help them decide which companies to fund. For this we predict whether a particular company seeking funding would file for bankruptcy/become defunct in the future.

Traditionally the decision of the VCs were based on a random combination of factors like Referrals, Revenues- present and forecasted, Channel partners, Business development deals, Management team, High growth market, Short term investment ,Good return on investment, A clearly defined exist strategy etc. However the survival rate of these companies is only 33%. Through this project, we aim to make this decision based on data acquired over the past 20 years by identifying the key predictors that affect the final success/failure of the company funded.

The databank is an exhaustive list of all companies seeking funding in various rounds over the past 25 years. It includes companies and VCs all over the world. We decided to focus on companies in the US funded by VCs all over the world.

Organization of the data is that it lists for each company and for each of its funding rounds the details of the VCs collaborating to raise the capital. It gives information about the company such as the operating city, product type, operating industry etc and information about the VCs such as firm country and city, industry preference, location preference, stage preference etc.

Data Preprocessing

1. The original data had 162 columns. Exhibit 1 shows a snapshot of few of the variables and their description. There were a lot of redundant/repetitive columns. Some of the columns were not relevant for the prediction objective. Hence those columns were eliminated.
2. Missing Values: There were some data rows with missing values. Exhibit 2 shows is how these values were replaced.
3. Dummy variables and bins were created based on the data mining method being used.

Column Name	Data Handling
Company_city	Replaced Blanks by 'No Information'
Company_nation_desc	Replaced Blanks by 'No Information'
Company_veic	Replaced Blanks by 'No Information'
Firm_city	Replaced Blanks by 'No Information'
Firm_founded_year	Replaced Blanks by '1234'
Firm_nation_desc	Replaced Blanks by 'No Information'
Firm_geography_pref	Replaced Blanks by 'No Information'
Firm_stage_pref	Replaced Blanks by 'Any'
Firm_pref_max	Replaced Blanks by '9999999'
Firm_industry_pref	Replaced Blanks by 'No Information'
Firm_pref_min	Replaced Blanks by '0'
Fund_nation_desc	Replaced Blanks by 'No Information'
Fund_size	Replaced Blanks by '9999999'

Data Analysis and Results

1. **Classification Trees:** The different predictors and the target variable for the classification tree are listed in exhibit 3. Most of the predictors (Xs) were either dummy variables or binary variables, all of which assumed the value of either 0 or 1.

   Results: The score results for the training, validation and Test data are given below:

Training Data

Classification Confusion Matrix		
	Predicted Class	
Actual Class	No	Yes
No	5254	1
Yes	14	33

Error Report			
Class	# Cases	# Errors	% Error
No	5255	1	0.02
Yes	47	14	29.79
Overall	5302	15	0.28

Validation Data

Classification Confusion Matrix		
	Predicted Class	
Actual Class	No	Yes
No	3148	2
Yes	22	10

Error Report			
Class	# Cases	# Errors	% Error
No	3150	2	0.06
Yes	32	22	68.75
Overall	3182	24	0.75

Test Data

Classification Confusion Matrix		
	Predicted Class	
Actual Class	No	Yes
No	2107	4
Yes	6	4

Error Report			
Class	# Cases	# Errors	% Error
No	2111	4	0.19
Yes	10	6	60.00
Overall	2121	10	0.47

The tree was selected to be best pruned tree and all inputs variables were normalized. However, we see that the %Error in predicting Bankruptcy (Yes) in validation and Test data is very large (68.75% and 60% respectively).

Conclusion: From the results we can conclude that the classification tree model does not help in making accurate percentages. The reason for failure of this model can be attributed to the input variables being binary and not continuous. In this case the classification tree will not be able to fit an accurate model.

2. **Naïve Bayes Analysis:** To carry out this analysis further, the number of columns was reduced by eliminating the relevance of the data towards classifying VC as unsuccessful. The data set originally contained the situation of the company after the Private Equity funding. Of the available situations we selected 'Chapter 7, Chapter 11 and Defunct' as 'Status' of company depicting bankruptcy. The others situations were selected as representing 'No' bankruptcy.

   On the above data set a naïve bayes classification was run using the first 10,000 rows. A cut off probability of 0.5 for success (bankruptcy) was used. On running the model an error of about 22% was obtained. To improve the performance of the classification the cut off was increased to 0.85, and the error was found to have reduced to 14%. It was important to have the error reduced significantly as the 'Type-II" error in this case would prove to be quite possible and we would like to avoid a situation of error as far as possible.

   Conclusion: The analysis worked well on the data and was tried on new data. The implications of the model are profound and will have significant influence on the way the PE

funds look at new proposals. With an error of about 14 %, a PE fund will be able to identify if the investment could turn defunct.

3. **K-Nearest Neighbors:** The KNN algorithm was applied to the data set after partitioning it into training, validation and test data set. The algorithm training is done on the training data set and the best performance for the validation and test data set is obtained with k=4 and cut-off probability as 0.03. The error reports are as follows. The results for the analysis are shown in exhibit 4.

   Conclusion: From the error report of the test data set we may say that for any new data that the model encounters it would be able to predict the class i.e. whether the company would go bankrupt or not with an overall error of 12.07%.

4. **Logistic Regression:** The initial data of 10,000 rows was partitioned in Training, Test & Validation Data. Realizing that the dependent variables in our data set were categorical, Logistic regression was run. The model hence obtained was able to predict whether a firm which VC intends to fund will go bankrupt or not with approx 13% error rate. The model also showed that some of the critical factors in predicting whether the company which the VC intends to fund will go bankrupt or not are Company VE Primary Industry Class, Firm Preferred Investment Stage & Age of Venture Capital firms. Please refer to Exhibit 5 on Logistic regression for details on the regression model.

   Conclusion: If he VC has following information about the company which it intends to fund i.e. Company's current public Status, Company VE Primary Industry Class, Firm Preferred Investment Stage & Age of Venture Capital firms, the Logistic regression model will be able to predict whether the company will go bankrupt or not with 87% accuracy.

5. **Linear Discriminant Analysis:** Since our data contained multiple dummy variables, we performed Logistic regression. However, we also performed Linear Discriminant Analysis in order to profile our data, rank predictor importance and to calculate 'classification scores' in order to predict the status of a new data entry. The results are shown in exhibit 6.

Output Variable: Bankruptcy Status(Yes/No)

Cut off Probability: 0.98

Conclusion: So, we can conclude that given any new data the model would be able to predict its class with approx 93% accuracy. The regression model is shown in exhibit 6 with the most significant predictors. Hence, the major factors determining whether a particular investment will result in Bankruptcy or not is dependent on mainly Firm_Stage_Pref and Firm_Type, Company VE Primary Industry Class.

Conclusion: By examining the results obtained from different data mining techniques we can draw the following conclusions:

1. The Key factors that help in predicting whether the company that the VC firm decides to fund will go bankrupt are:
   Company VE Primary Industry Class
   A. Firm Preferred Investment Stage
   B. Age of Venture Capital firms
   C. Venture capital firm type
   D. Primary Industry Class of the company

Few general insights from the overall data are:

- United States Companies in medical/health industry/life sciences industry have less chance of going bankrupt.
- Older VC firms because of the experience effect generally pick the correct firms to invest in.

Exhibits

Exhibit 1: Data set description

Data Variables	Description	Data Variables	Description
round_date	Date on which the funding was given	company_ipo_flag	Company IPO Y/N
fund_close_date	Close data for the fund	company_region	Geographic region
ipo_date	date on which IPO was offered	company_primcus	Company Primary Customer Type
company_sit_date	Date on which the company situation was declared	company_sic	Company Primary SIC Code
std_buyout	1/0 for company buyout status	company_product	Company Product
cusip	Company id- used to map company to the company names document	stage1	Company stage Level 1 at Round date
company_area_code	Area code	stage2	Company stage Level 1 at Round date
ve_auditor	Company Auditor	stage3	Company stage Level 1 at Round date
ve_banker	Company Banker	company_state_desc	Company State
busdesc	Company Business description	company_state	Company State
company_city	City name	company_exchange	Company Stock Exchange
company_sit	Status of the company (acquired/Bankrupt etc)	company_ticker	Company Ticker (e.g. NSCP)

Exhibit 2:

Column Name	Data Handling
Company_city	Replaced Blanks by 'No Information'
Company_nation_desc	Replaced Blanks by 'No Information'
Company_veic	Replaced Blanks by 'No Information'
Firm_city	Replaced Blanks by 'No Information'
Firm_founded_year	Replaced Blanks by '1234'
Firm_nation_desc	Replaced Blanks by 'No Information'
Firm_geography_pref	Replaced Blanks by 'No Information'
Firm_stage_pref	Replaced Blanks by 'Any'
Firm_pref_max	Replaced Blanks by '9999999'
Firm_industry_pref	Replaced Blanks by 'No Information'
Firm_pref_min	Replaced Blanks by '0'
Fund_nation_desc	Replaced Blanks by 'No Information'
Fund_size	Replaced Blanks by '9999999'

Exhibit 3: Classification Tree Variables

Variable	Name of the variable	Data Type	Values
Y	Bankruptcy status	Categorical	Yes/No
X1	Same City	Binary	0/1
X2	Same Nation	Binary	0/1
X3	Public status of the organization	Dummy Variables	Public Private Registered Subsidiary
X4	Primary Customer Type	Dummy Variables	Business Consumer etc
X5	Company Industry	Dummy Variables	Information Technology Medical/Healthcare High Technology
X6	VC age	Continuous	Numerical
X7	Firm Stage Preference	Dummy Variables	27 categories
X8	Firm Max Funding	Continuous	Numerical
X9	Firm type	Dummy Variables	7 Categories

Exhibit 4: K-Nearest Neighbors

Validation Data scoring - Summary Report (for k=4)

Cut off Prob.Val. for Success (Updatable)	0.03

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	21	11
0	428	2722

Error Report			
Class	# Cases	# Errors	% Error
1	32	11	34.38
0	3150	428	13.59
Overall	3182	439	13.80

Test Data scoring - Summary Report (for k=4)

Cut off Prob.Val. for Success (Updatable)	0.03

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	9	1
0	255	1856

Error Report			
Class	# Cases	# Errors	% Error
1	10	1	10.00
0	2111	255	12.08
Overall	2121	256	12.07

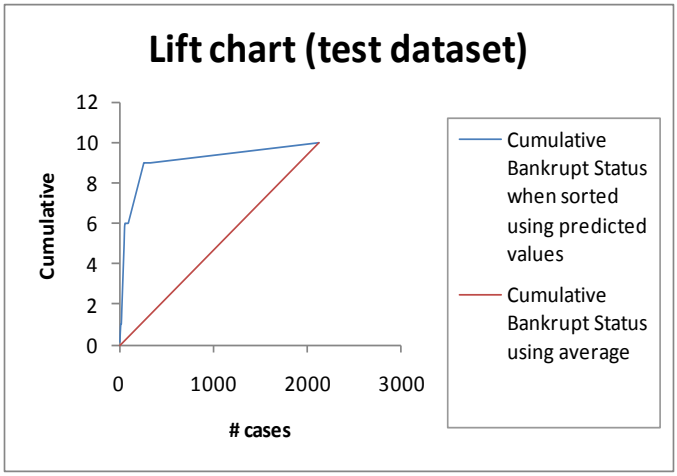


Exhibit 5: Logistic Regression

Training Data Scoring:

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
No	5259	696	13.23
Yes	43	4	9.30
Overall	5302	700	13.20

Validation Data Scoring:

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
No	3152	429	13.61
Yes	30	3	10.00
Overall	3182	432	13.58

Test Data Scoring:

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
No	2105	282	13.40
Yes	16	2	12.50
Overall	2121	284	13.39

The full regression model is as under:

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	12.62384892	1.5338316	0	*
SameCityContinuous	0.74530333	1.04183745	0.47437805	2.10708046
pubstatus_Private	-4.39932203	0.85664064	0.00000028	0.01228567
pubstatus_Public	-2.75588894	0.89388663	0.00204897	0.0635525
pubstatus_Registration	11.85260773	4740.899902	0.99800521	140450.1406
CompanyNationNonUS	16.39825439	1550.99231	0.99156433	13233390

company_veic6a_Medical/Health/Life Science	2.76052547	0.79014641	0.0004764	15.80814743
company_veic6a_Non-High Technology	0.75738037	0.4560923	0.09679668	2.13268209
firm_stage_pref_Acquisition	-2.18178558	1.23661017	0.07767685	0.11283987
firm_stage_pref_Balanced	-2.31089711	1.14453542	0.04347993	0.09917223
firm_stage_pref_Control-block Purchases	-4.98024082	1.36376536	0.00026038	0.00687241
firm_stage_pref_Distressed Debt	-4.20629787	1.59738553	0.00845748	0.01490143
firm_stage_pref_Early Stage	-1.60578775	1.08103251	0.13743247	0.20073137
firm_stage_pref_Expansion	13.64739513	1947.642578	0.99440914	845256.125
firm_stage_pref_First Stage Financing	13.4328804	2327.829346	0.99539578	682065.375
firm_stage_pref_Fund of Funds	11.98837471	4741.724121	0.99798274	160873.6875
firm_stage_pref_Fund of Funds of Second	13.1481142	18306.99609	0.99942696	513042.8125
firm_stage_pref_Generalist PE	12.33412838	6333.88623	0.99844629	227323.1875
firm_stage_pref_Industry Rollups	12.54209518	5708.794922	0.99824709	279874.1875
firm_stage_pref_Joint Ventures	12.91556168	8424.390625	0.99877673	406590.4375
firm_stage_pref_Later Stage	-3.12045026	1.22285759	0.01071775	0.04413729
firm_stage_pref_Leveraged Buyout	-1.82181931	1.27983963	0.15459859	0.16173124
firm_stage_pref_Management Buyouts	13.57645416	10693.71973	0.99898702	787370.625
firm_stage_pref_Mezzanine	13.34694862	1983.120117	0.99463004	625901.625
firm_stage_pref_Open Market	12.56974602	35855.76172	0.99972028	287720.7188
firm_stage_pref_Other	14.59449863	19017.48633	0.99938768	2179266
firm_stage_pref_Private Placement	12.03882122	5519.442871	0.99825966	169197.375
firm_stage_pref_Public Companies	12.67458439	7491.530762	0.99865007	319522.9688
firm_stage_pref_Recapitalizations	12.69898129	4160.945801	0.99756491	327414.3125
firm_stage_pref_Research and Development	18.87444687	2022.111328	0.99255264	157423000
firm_stage_pref_Second Stage Financing	13.59771824	2121.09082	0.99488503	804292.25
firm_stage_pref_Seed	-3.1980567	1.10300982	0.00373888	0.04084149
firm_stage_pref_Special Situation	13.29226112	7335.561523	0.99855423	592591.625
firm_pref_max	-0.00000004	0.00000047	0.93101144	0.99999994
firm_type_Affiliate/Subsidary of Oth. Financial. Instit.	17.10965729	2440.987793	0.99440742	26954420
firm_type_Business Development Fund	-1.36972058	1.10654736	0.21577807	0.25417796
firm_type_Commercial Bank Affiliate or Subsidiary	15.30824089	2377.242676	0.99486208	4449229
firm_type_Corporate Subsidiary or Affiliate	14.18335819	2147.074707	0.99472928	1444622
firm_type_Corporate Venture Program	12.59046936	6649.715332	0.99848932	293745.4688
firm_type_Federal Govt Affiliated Program	14.83759594	23871.55664	0.99950409	2778983
firm_type_Incubators	15.20588303	19698.55078	0.99938411	4016346
Firmage	-0.10547637	0.01521755	0	0.89989573

The significant predictors along with the p-values are as under:

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	12.623849	1.5338316	0	*
pubstatus_Private	-4.399322	0.8566406	2.8E-07	0.0122857
pubstatus_Public	-2.7558889	0.8938866	0.002049	0.0635525
company_veic6a_Medical/Health/Life	2.7605255	0.7901464	0.0004764	15.808147

Science				
firm_stage_pref_Balanced	-2.3108971	1.1445354	0.0434799	0.0991722
firm_stage_pref_Control-block Purchases	-4.9802408	1.3637654	0.0002604	0.0068724
firm_stage_pref_Distressed Debt	-4.2062979	1.5973855	0.0084575	0.0149014
firm_stage_pref_Later Stage	-3.1204503	1.2228576	0.0107178	0.0441373
firm_stage_pref_Seed	-3.1980567	1.1030098	0.0037389	0.0408415
Firmage	-0.1054764	0.0152176	0	0.8998957

Exhibit 6: Linear Discriminant Analysis

Validation Data Scoring:

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
No	3151	202	6.41
Yes	30	4	13.33
Overall	3181	206	6.48

Test Data Scoring:

Error Report			
Class	**# Cases**	**# Errors**	**% Error**
No	2104	138	6.56
Yes	16	3	18.75
Overall	2120	141	6.65

Variables	No	Yes	Difference
company_veic6a_Medical/Health/Life Science	**8.64153576**	**7.67212152**	**0.96941424**
firm_stage_pref_Joint Ventures	**38.19779205**	**35.20859528**	**2.98919677**
firm_stage_pref_Open Market	**33.78020859**	**32.65989304**	**1.12031555**
firm_stage_pref_Other	**39.04252243**	**37.24837875**	**1.79414368**
firm_stage_pref_Research and Development	**36.59062195**	**34.68931961**	**1.90130234**
firm_type_Affiliate/Subsidary of Oth. Financial. Instit.	**-0.06781529**	**-1.75951064**	**1.69169535**