Predicting Project Failures for California Solar Initiative Projects

Team 3

Hakan Ozalp
Peter Protopappas
Nitin Sawant
Prem Swaroop

BUDT733 Data Mining

Spring 2010

May 10th, 2010

Executive Summary

The California Solar Initiative (CSI) provides incentives to customers in territories administered by Pacific Gas and Electric Company (PG&E), Southern California Edison (SCE), and San Diego Gas & Electric (SDG&E). These three utilities represent about 68 percent of California's electric load. The CSI provides cash back for solar energy systems for existing homes, as well as existing and new commercial, industrial, government, non-profit, and agricultural properties – within the service territories of the three above-listed utilities. The CSI has a budget of $2,167 million over 10 years, and the goal is to reach 1,940 megawatts (MW) of installed solar capacity by 2016.

In an effort to track data on the projects initiated through the CSI, the CSI has provided data on all projects initiated between September 2006 and February 2010. Each record in the dataset provides information on the incentive structure, cost, total kWatt capacity and location of a single project. As of January approximately 2500 projects in the dataset were either cancelled or withdrawn. This study is aimed at investigating the data to learn from these instances and to understand the cause behind these failures with the goal of producing a model to predict the likelihood of a project failure. Specifically four classification test models for were used, Classification Trees, Logistic Regression, Discriminant Analysis, and Naïve Bayes. After running each, the best model, Classification Trees with an overall error rate of 26.83%, was selected and recommended for use by CSI in predicting project failures.

Technical Summary

Data Preparation

Background

The original data-set had 96 fields, which can be categorized into six types:

1. Program Related Details
2. Status
3. Ownership
4. Geography
5. Processing Dates
6. Installation Related Details

We provide a snapshot of the data in Exhibit A. The dataset contained 37,606 records, consisting of solar projects since inception of the program, spanning three states and four sectors. We focused our work on Residential projects in state of California (CA), which comprised of 34,479 records.

Set Creation, Transformation and Handling missing values

Count	Set			
Status	**Test**	**Training**	**Validation**	**Row Total**
Failure	366	916	549	**1831**
Success	5154	12886	7732	**25772**
Column Total	**5520**	**13802**	**8281**	**27603**

% age to Column Total	Test	Training	Validation	Row Total
Failure	6.63	6.64	6.63	**6.63**
Success	93.37	93.36	93.37	**93.37**
Column Total	**20**	**50**	**30**	**100**

We then created separate Training, Validation, and Test sets in 50-30-20 ratio using Spotfire Miner. We were not able to use XLMiner for this purpose due to its record limitations. A fourth Predict set was also created which consisted of "In Progress" projects. This set was not used in the modeling process. We ensured that all datasets had the same proportion of Failure vs Success projects.

The early decision to split into three distinct sets was taken, so that each of the team members could work on individual modeling tasks, but always reported the performance measures on the same Test Set. This way the results across the team members and various models would be exactly comparable, and not impacted by a random split. We suspected the random split could produce significantly different results due to the class imbalance.

In addition to the modifications described above here is a list of some of the other pre-processing operations applied to the data before any models were run.

1. Retained only PV Manufacturer#1, Inverter Manufacturer#1 ; removed the rest of

manufacturers, models, and quantity fields since they were empty

2. Added imputed field Incentive Rate, it has weighted average of the values in Incentive Design field.
3. Added imputed field Split Incentive: 0 indicates Incentive Design had no split (i.e., had only one Step), 1 if it was split.
4. Added imputed field Incentive First Step, with the first step in Incentive Step if split, else whichever Step it had.
5. Added imputed field Incentive First Slab, with the first %age amount in the Incentive Step in case of split design, else 100%.
6. Added imputed field Project Initiation Date, with the minimum of the initial three date fields, viz., First New Reservation Request Date, First Online Reservation Request Submitted Date, and First Reservation Request Review Date.
7. 60 counties are broken into 8 regions (new column Region) based on data in this site (http://www.ca.gov/About/Government/Local/counties.html). These regions seem to be a good indicator of terrain (coast, mountains, valley, desert)
8. Missing values were imputed using Regression and Classification Trees

Exhibit B gives details on the final dataset used for modeling.

Data Exploration

Exploring the dataset with Spotfire, we looked for obvious patterns within the dataset. We created box-plots and line charts for visualizing relationships between the variables with respect to project status.

Intuitively, the first set of variables to look for was related to incentives. However, we found no visual association between incentives and successful projects (Exhibit C). The chart below shows that successful projects decrease in number as incentive rates reduce but this decrease ignored the projects which were currently in progress.

Since the dataset was across years we also looked for time series based correlations. We did not find any obvious trends. Some other variables of interest turned out to be average incentive amount, which changed from region to region, incentive type schemes. However, there were no clearly visible relationships.

We found out that program administrators have different success rates in overall for their projects (Exhibit D). Since program administrators are defined by the region they operate, it suggested that the region could be a variable of interest for our predictions as well.

Model Development

Overall Modeling Approach

We decided to develop four models, naïve bayes, classification trees, logistic regression and discriminant analysis. For each model several iterations tried to improve upon the error rates and simplify required models parameters. Spotfire Miner was used for classification trees, XLMiner was used for discriminant analysis and Weka was used for naïve bayes and logistic regression.

For developing the models we combined the records marked as training and validation into the "training partition" and the test set was used as the "validation partition".

Oversampling was done on this training partition to produce a stratified sample having equal proportion of both classes. Sampling procedure produced as many records as input, drawing samples repeatedly with replacement from the data. Due to this stochasticity in the procedure, repeated runs with exactly same settings produced different quality of results; we produce here the best we could obtain after multiple runs

Naïve Bayes

The naïve Bayes model generated using Weka provided the following confusion matrix. The validation set used for scoring the results has a different proportion than the population, which needs to be corrected when comparing results.

Classified as -->	a	b	Error Rate
a = Failure	2758	5642	**67.17%**
b = Success	677	8107	**7.71%**
			36.77%

Classification Tree

Classification trees were then run on the output from sampling procedure, with mostly default settings provided by Spotfire Miner. We ran the procedures with different settings multiple times. The best results on test data were produced with following default settings: Single tree, Min Node Size before split: 10 and after split: 5, Splitting Criteria: Entropy.

We introduced Pruning: Minimum Complexity for best results. It may be noted that Spotfire Miner by default conducts 5-fold cross-validation – and we left this setting as-is. We then produced the preditor node, and ran test data through it, followed by producing the classification agreement. Exhibit F shows a graph generated using Spotfire Miner which shows the various predictors sorted by their contribution to entropy reduction. IncentiveCost, ProjectInitiationDate, PVModule#1Manufacturer and Region were identified as the top four variables. The table below gives the confusion matrix provided by the model.

Validation set used for scoring here used the same proportion of both classes as the population.

Classified as -->	a	b	Error Rates
a = Failure	258	107	**29.32%**
b = Success	1374	3780	**26.66%**
			26.83%

Logistic Regression

Exhibit G lists the odds ratios generated by Weka for the model with the best combination of error rate and number of predictors (ProgramAdministrator, IncentiveType, IncentiveToCost, CSIRating, SystemOwnerSector and Region). The predictor that stood out was IncentiveToCost with a very high odds ratio (4047). Partially this was due to the units associated with the ratio. The typical increments in this predictor were of the order of 0.01. We also discovered that Failed projects had significantly lower ratios than Successful projects since the cost data associated with failed projects was incomplete. The confusion matrix provided by the model is given below.

Similar to naïve Bayes, the validation set used for scoring the results has a different proportion than the population, which needs to be corrected when comparing results.

Classified as -->	a	b	Error Rates
a = Failure	5082	3318	**39.50%**
b = Success	3064	5720	**34.88%**
			37.14%

Discriminant Analysis

Exhibit H lists the classification function scores generated by XLMiner for the model with the following variables: Program Administrator – Dummies, Incentive Type – Dummies, IncentiveAmount, TotalCost, IncentivetoCost, CSIRating, Region – Dummies, IncentiveRate. The most impactful variables in terms of classification scores were IncentivetoCost, Incentive Type and Program Administrator variables. The confusion matrix provided by the model is given below.

Misclassification cost of 10 to 1 is used for misclassifying a "Failure" project as we discuss in our recommendation as well. The validation set used for scoring here has the slightly different proportion as of the population, which needs to be accounted for during comparison models.

Classified as -->	a	b	Error Rates
a = Failure	**364**	**210**	**57.69%**
b = Success	**5154**	**876**	**17.00%**
			19.68%

Evaluation

Since the misclassification cost associated with incorrectly classifying a "Failure" project was much higher than the cost associated with misclassifying a "Success" project, we decided to incorporate this imbalance by assigning a higher weight of 10 to misclassified "Failure" records. This metric was used instead of the overall error rate for comparing the various models.

Cost metric was computed as below for each of the model results:

Cost Metric (CM) = Misclassification cost ratio * Population Proportion of "Failure" Class * Sensitivity (CM1) + Population Proportion of "Success" Class * Specificity (CM2)

Model	a	b	Scoring set proportion for “Failure”=(a+b)/n	Sensitivity=b/(a+b)	Population proportion for “Failure”	CM1	CM
	c	d	Scoring set proportion for “Success”=(c+d)/n	Specificity=c/(c+d)	Population proportion for “Success”	CM2	
Naïve Bayes	2758	5642	0.49	0.67	0.07	0.44	**0.52**
	677	8107	0.51	0.08	0.93	0.07	
Classification Tree	258	107	0.07	0.29	0.07	0.19	**0.44**
	1374	3780	0.93	0.27	0.93	0.25	
Logistic Regression	5082	3318	0.49	0.4	0.07	0.26	**0.59**
	3064	5720	0.51	0.35	0.93	0.33	
Discriminant Analysis	364	210	0.09	0.37	0.07	0.24	**1.04**
	5154	876	0.91	0.85	0.93	0.8	

Recommendation

The mis-classification cost metric of 0.44 provided by Classification Tree was significantly better than those than provided by Naïve Bayes (0.52), Logistic Regression (0.59 or Discriminant Analysis (1.04).

The Classification Tree model can be used for predicting outcome of the “Predict” set which represents ongoing projects. The model can be further refined based on the results.

Furthermore, an ensemble of Classification Tree along with Logistic Regression and Naïve Bayes could also be tried to improve results.

Exhibit A – Sample of the original data set

Program Admin	Program	Incentive Design	Incentive Type	Incentive Amount	Total Cost	Nameplate Rating	CEC PTC Rating	Design Factor	CSI Rating	Current Incentive Application Status
PG&E	Large Commercial (>= 10 kW)	$2.30 per Watt EPBB	EPBB	91052	250569	34.34	29.038	0.96477	28.015	Completed
PG&E	Large Commercial (>= 10 kW)	$0.39 per kWh FiveYearPBI	FiveYearPBI	2331209	12183004	865.52	746.469	1.0156825	758.174	PBI - In Payment
PG&E	Large Commercial (>= 10 kW)	$0.39 per kWh FiveYearPBI	FiveYearPBI	201995	618689	76.8	64.983	1.01094	65.694	PBI - In Payment
PG&E	Large Commercial (>= 10 kW)	$0.39 per kWh FiveYearPBI	FiveYearPBI	231935	704214	86.4	73.106	1.03182	75.432	PBI - In Payment
PG&E	Large Commercial (>= 10 kW)	$0.39 per kWh FiveYearPBI	FiveYearPBI	235197	692664	86.4	73.106	1.04633	76.493	PBI - In Payment
PG&E	Large Commercial (>= 10 kW)	$0.46 per kWh FiveYearPBI	FiveYearPBI	642832	1540000	201.6	168.086	0.9491	159.53	Pending Payment
PG&E	Large Commercial (>= 10 kW)	$0.37 per kWh FiveYearPBI	FiveYearPBI	481154	1540000	200.464	172.862	0.8587679	148.449	Pending Payment
PG&E	Large Commercial (>= 10 kW)	$0.34 per kWh FiveYearPBI	FiveYearPBI	3732228	8937500	1205.624	1004.607	0.974	978.487	PBI - In Payment
PG&E	Large Commercial (>= 10 kW)	$0.39 per kWh FiveYearPBI	FiveYearPBI	3758040	8718365.5	1284.48	1076.94	0.94	1012.324	Cancelled
PG&E	Small Commercial (< 10 kW) and All Residential	$2.50 per Watt EPBB	EPBB	6248	24200	3.06	2.55	0.98	2.499	Cancelled

Exhibit B – Statistics on final dataset used for modeling

Count	Status			
Set	**Failure**	**In Progress**	**Success**	**Total**
Predict		6775		**6775**
Test	365		5154	**5519**
Training	914		12886	**13800**
Validation	548		7732	**8280**
Total Result	**1827**	**6775**	**25772**	**34374**

Proportion of Failure to Total (without In Progress)	
Test	6.61
Training	6.62
Validation	6.62
Total Result	**6.62**

Missing Values Imputed				
Variable	**Failure**	**In Progress**	**Success**	**Total**
IncentiveAmount	16	74		**90**
IncentivetoCost	282	268	15	**565**
NameplateRating	25	18		**43**
CECPTCRating	7	18		**25**
DesignFactor	19	8		**27**
SystemOwnerSector	10	6	298	**314**
Total	**359**	**392**	**313**	

Exhibit C – Visualization 1(Counted Status vs. IncentiveFirstStep)

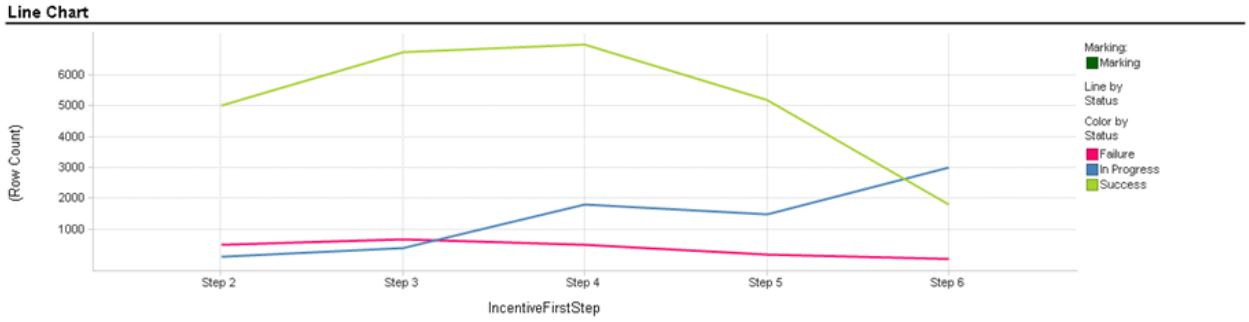

Exhibit D – Visualization 2(Counted Status vs. Program Administrator)

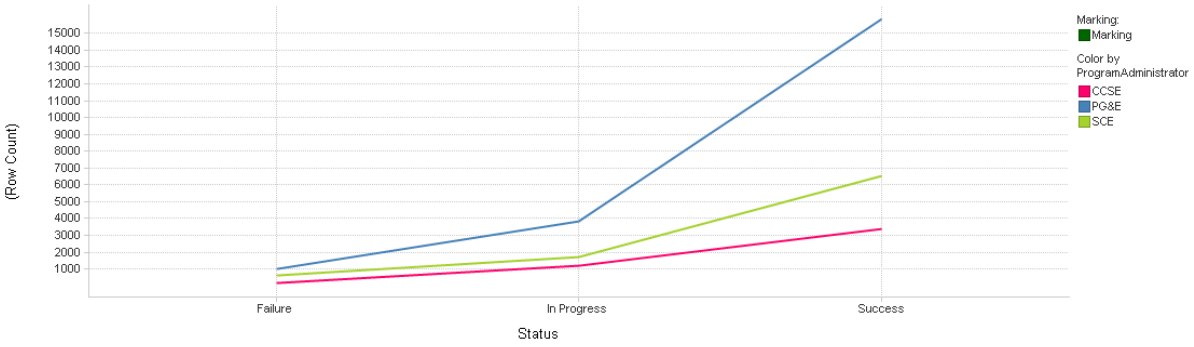

Exhibit E – Spotfire Miner workflow for Classification tree models

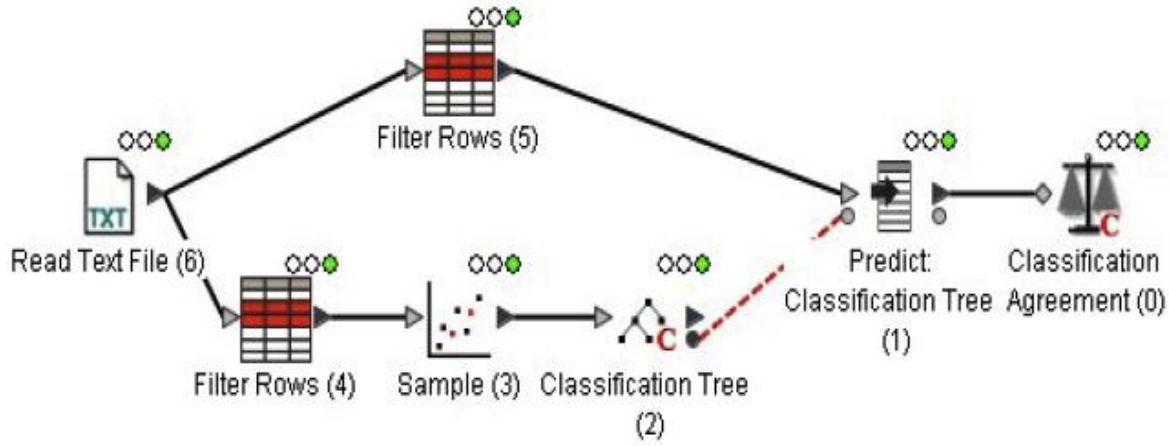

Exhibit F – Variable Importance graph for Classification Trees

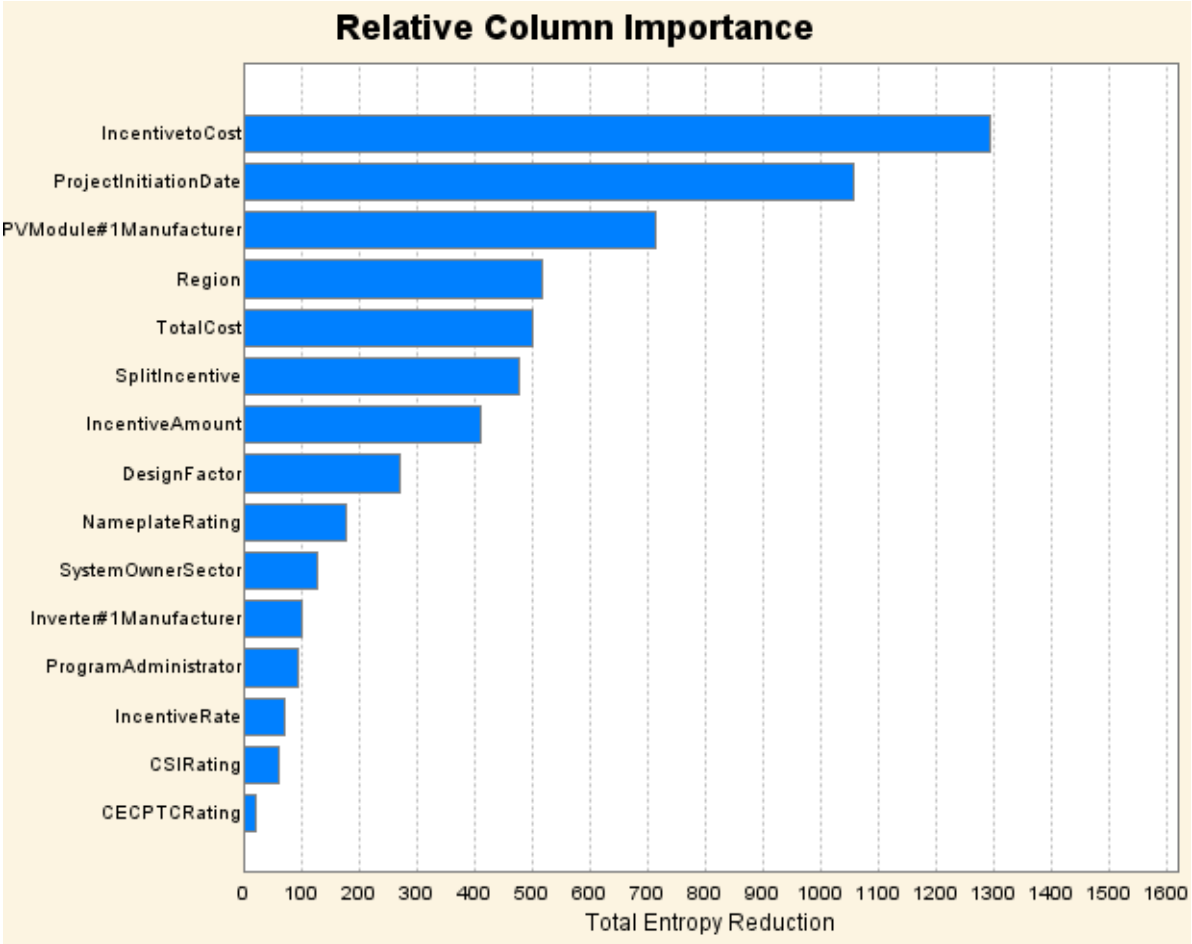

Exhibit G –Logistic Regression Coefficients

Odds Ratios...

Variable	Class Failure
ProgramAdministrator=PG&E	0.7422
ProgramAdministrator=CCSE	1.4464
IncentiveType=EPBB	0.5587
IncentivetoCost	4646.9616
CSIRating	1.0596
SystemOwnerSector=Residential	1.69E+14
SystemOwnerSector=Commercial	4.56E+14
SystemOwnerSector=Government	0
Region=BayArea	0.8664
Region=CentralCoast	0.4609
Region=CentralSierra	1.622
Region=GreaterSacramento	0.8276
Region=NorthernSacramento	0.4106
Region=SanJoaquin	1.0711
Region=SouthernBorder	0.2965
Region=SouthernCalifornia	0.9335

Exhibit H –Discriminant Analysis Classification Function

Variables	Classification Function	
	Success	Failure
Constant	-36,45683289	-37,7853508
ProgramAdministrator_CCSE	2,21627975	2,90660572
ProgramAdministrator_PG&E	0,75585097	0,21433637
IncentiveType_EPBB	75,32743835	74,07193756
IncentiveAmount	0,0012184	0,0012157
TotalCost	0,00005718	0,00007279
IncentivetoCost	6,26172686	9,59914303
CSIRating	-2,7268703	-2,80722189
Region_Bay Area	2,20955157	2,24403071
Region_Central Coast	2,36498022	1,84272933
Region_Central Sierra	1,8431493	2,06552887
Region_Greater Sacramento	2,68524766	2,53960919
Region_Northern	2,02042103	2,05131769
Region_Northern Sacramento	1,77613175	1,09082878
Region_San Joaquin	2,09094429	2,3711853
Region_Southern Border	2,47616982	1,13995349
IncentiveRate	-3,81777787	-3,11462426