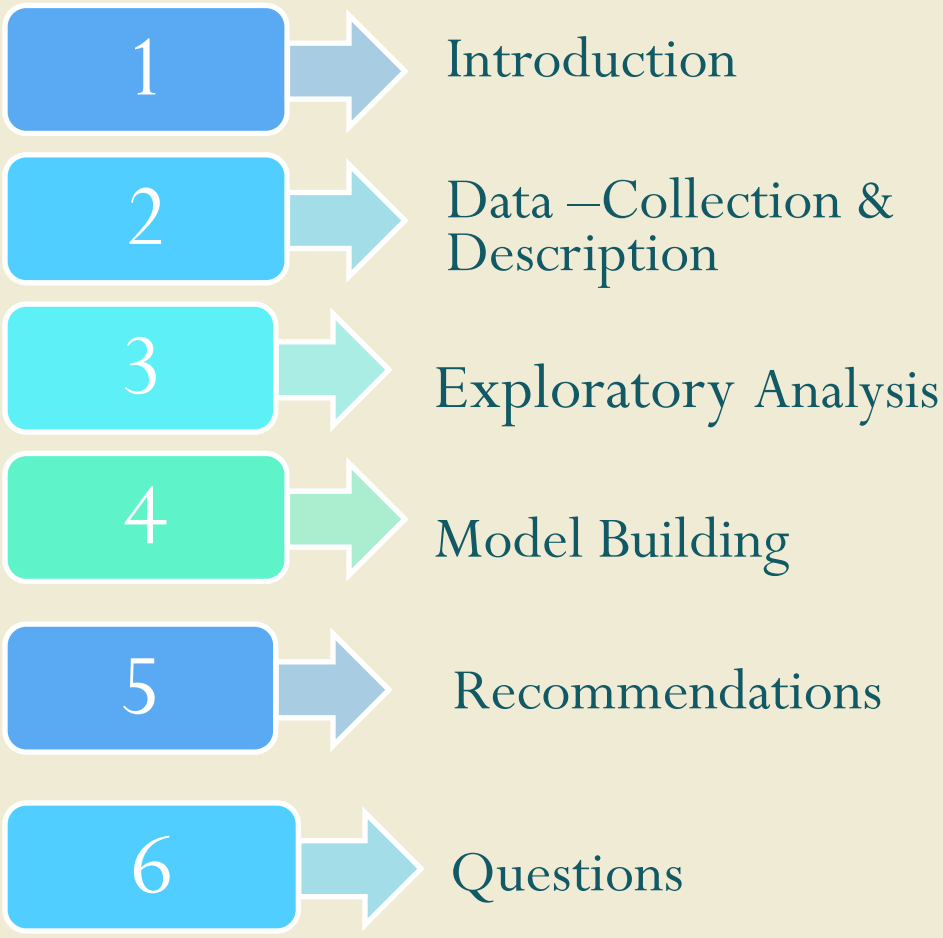
Predicting Cancellations for California Solar Initiative (CSI) Projects

Group – 3

Hakan Ozalp
Nitin Sawant
Peter Protopappas
Prem Swaroop

Agenda


1
Introduction
2
Data –Collection & Description
3
Exploratory Analysis
4
Model Building
5
Recommendations
6
Questions

Introduction

- **The California Solar Initiative (CSI) aims at promoting the use of solar energy by homes, businesses and government buildings**
- **Dataset on all projects undertaken between 2006 and 2010 includes details regarding incentives given, cost, capacity and geography**
- **37606 records and 96 columns in the raw data set**

GOAL: To create a model which can predict potential cancellations in CSI projects based on project attributes

Data Collection & Description

DATA DESCRIPTION

1. Status
2. Program Administrator
3. Incentive Type
4. Incentive Amount
5. Total Cost
6. Incentive to Cost Ratio
7. Nameplate Rating
8. CECPTC Rating
9. Design Factor
10. CSI Rating
11. System Owner Sector

DATA DESCRIPTION

12. Host Customer Physical Address City
13. Host Customer Physical Address County
14. Host Customer Physical Address Zip Code
15. Region
16. PVModule#1 Manufacturer
17. Inverter#1 Manufacturer
18. Incentive Rate
19. Split Incentive
20. Incentive First Step
21. Incentive First Slab
22. Project Initiation Date

Data Collection & Description

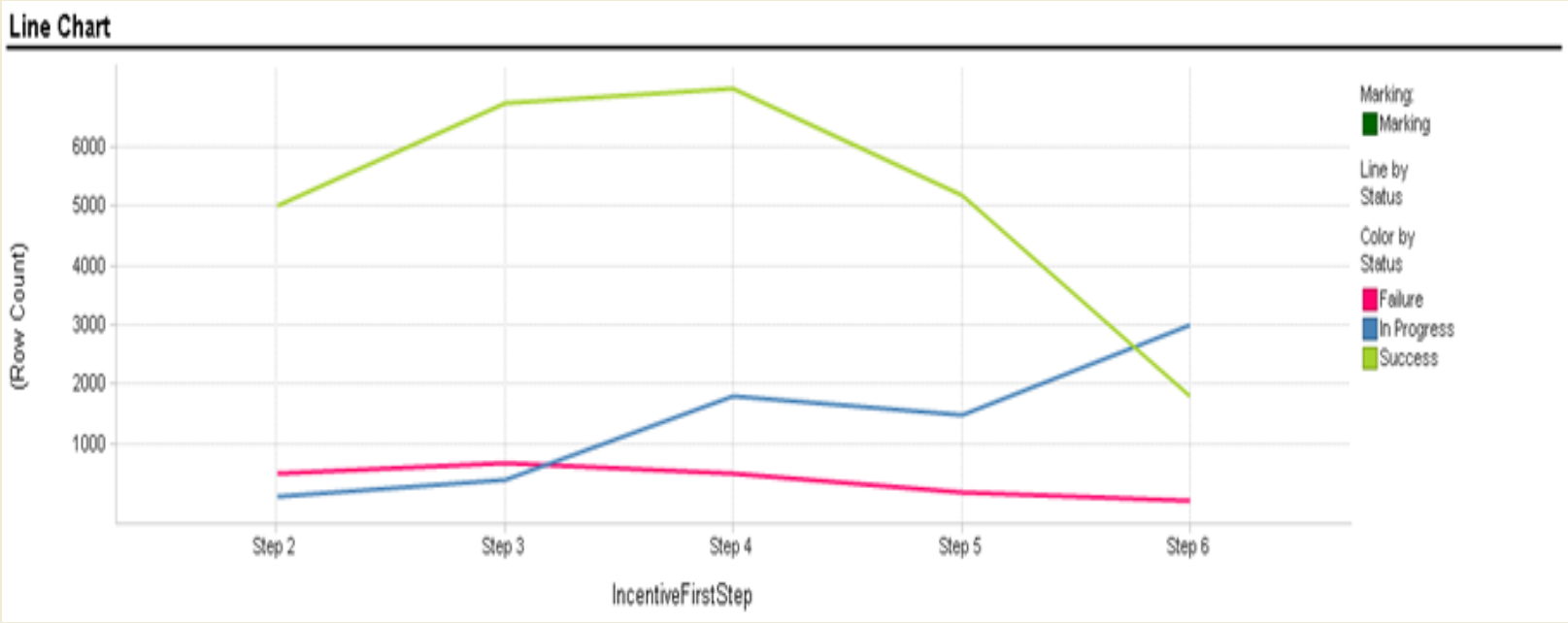
DATA CLEANUP CHALLENGES

- Y variable - Project Status (Success or Failure)
- Handling missing values
- Extracting numeric information coded in descriptive fields
- Reducing the number of fields
- Duplicate records
- Categorizing geographical data

PARTITIONING & OVERSAMPLING

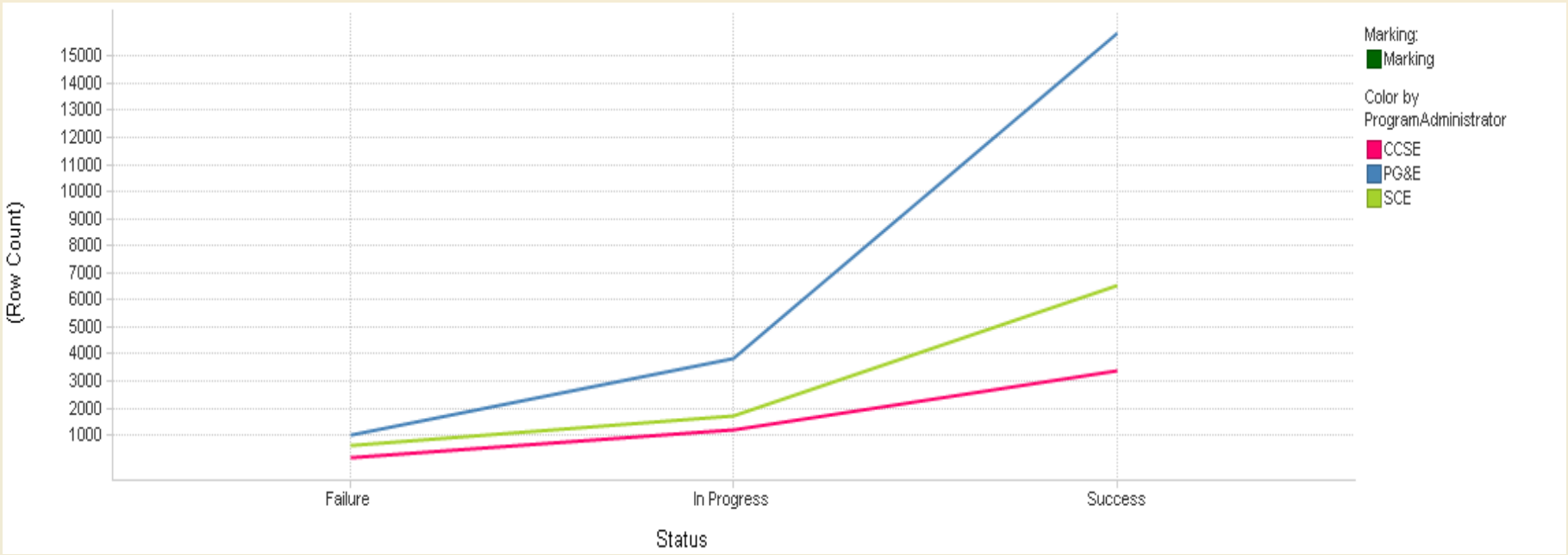
- Only 6% records with status "Failure"
- Oversampling to provide enough samples for model training
- Creating partitions with the same class proportion as the population
- Three partitions created – Train, Validate and Test

Exploratory Analysis



IncentiveFirstStep and Status not so related

Exploratory Analysis



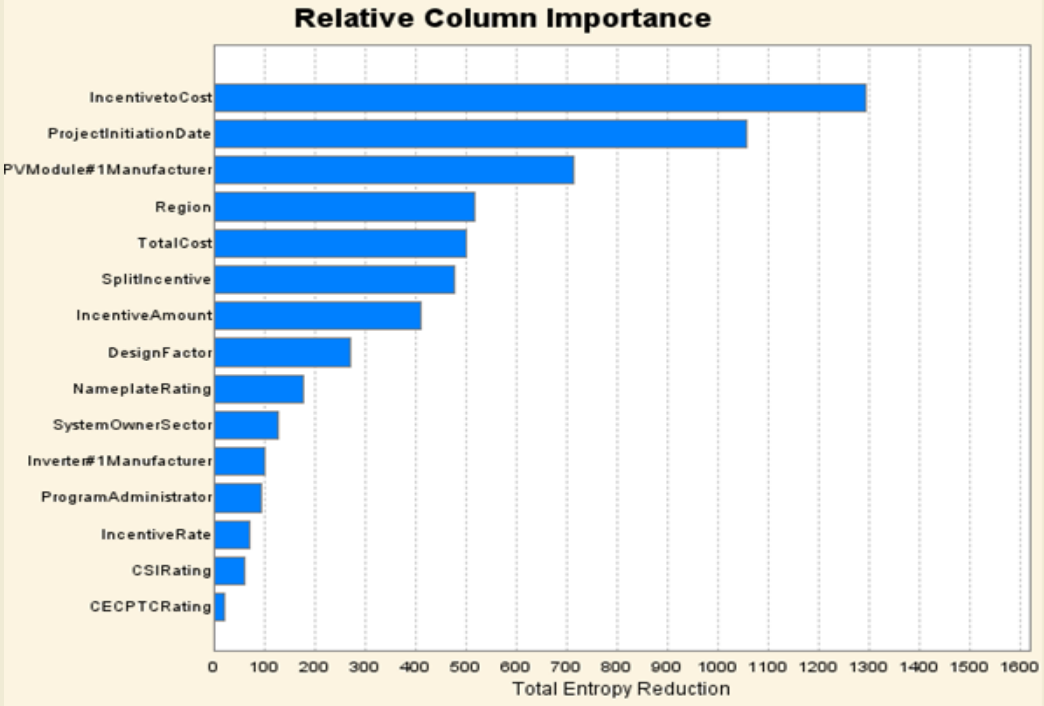
- Different rates of success for different program administrators.

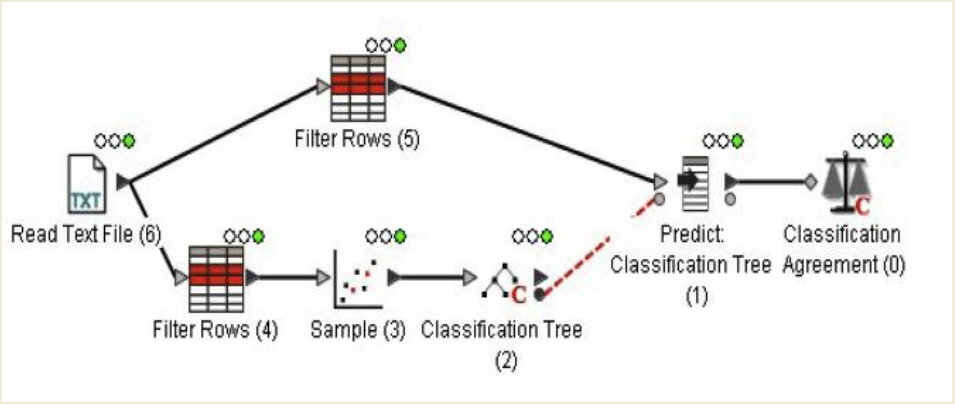
- Program administrators are defined by their respective regions. Region could be a variable of interest for prediction.

Model Development – Classification Trees

The top four variables were identified as

- Incentive To Cost
- Project Initiation Date
- PVModule#1 Manufacturer
- Region





Confusion Matrix

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

Model Development – Discriminant Analysis and Logistic Regression

Discriminant Analysis

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

Logistic Regression

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

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

Recommendations

- Models were compared based on misclassification costs and after adjusting for the difference in class proportion in the sample versus the population
- Classification Tree based model was selected
- Model can be applied to current data in ongoing projects and model can be further fine tuned based on the results
- Challenges – Changes in the CSI project environment (incentive structures)