Design of Type II Diabetes Intervention Program using Predictive Modeling

Business Data Analytics with Data Mining

Presented By:

Harshit Mehta (HM23388)

Sanjit Paliwal (SP42626)

Contents

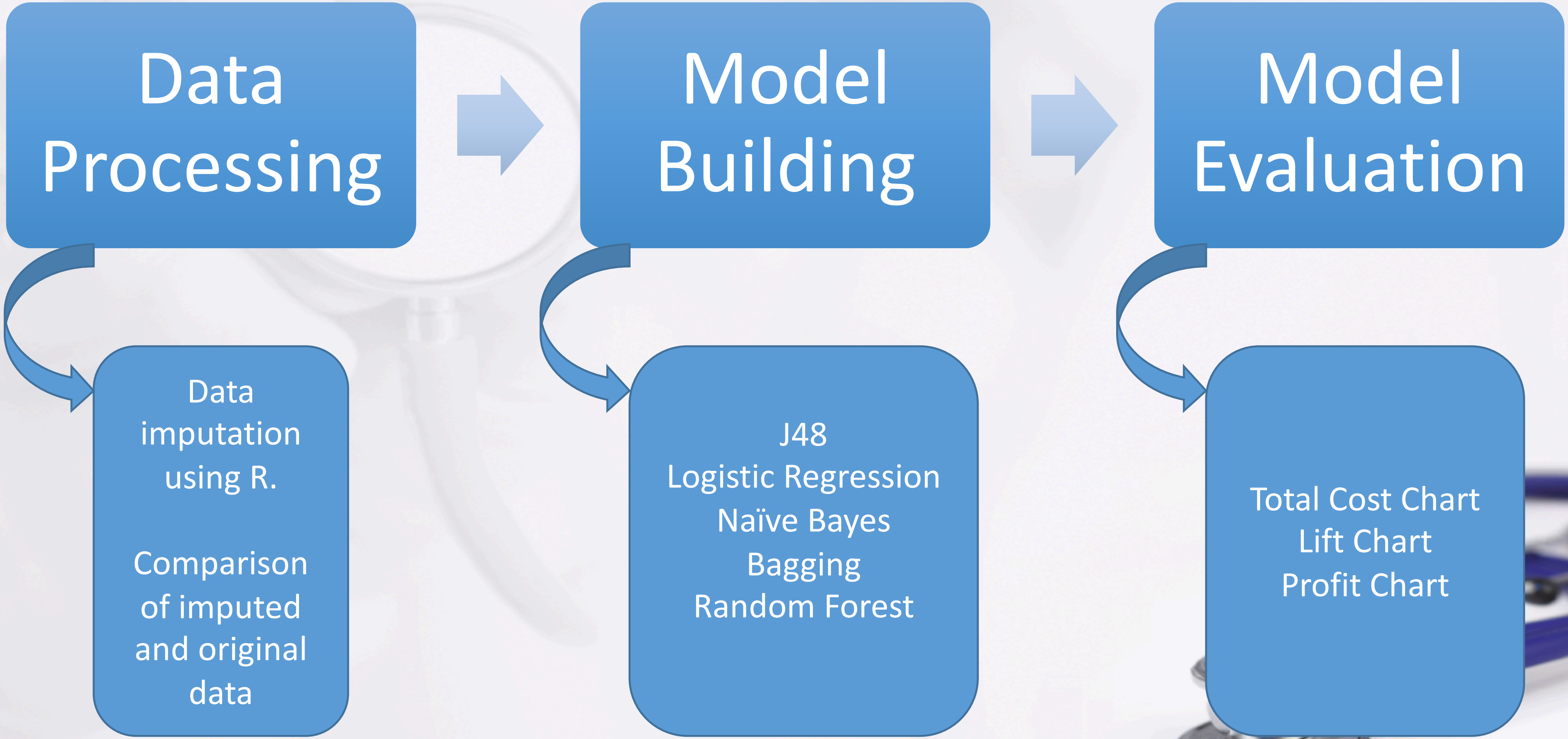
- Objective
- Process
- Data Processing
- Model Building
- Model Evaluation
- Project Results & Insights

Objective

Design of Type II Diabetes Intervention Program using Predictive Modeling

- Intervention program helps reduce chance of developing diabetes by 60%
- Aim is to design a program that minimizes total cost for all prediabetics using various predictive modeling techniques and comparing it with current practice in which a patient having Fasting Blood Glucose>110 is classified as “Diabetic”
- Predictive Model predicts the class of a person as “Diabetic” or “Non-Diabetic” and helps in ranking patients according to their likelihood of being “Diabetic” or “Non-Diabetic”

Process



Data Processing

Data imputation using R.

Comparison of imputed and original data

Model Building

J48
Logistic Regression
Naïve Bayes
Bagging
Random Forest

Model Evaluation

Total Cost Chart
Lift Chart
Profit Chart

Data Processing

- Missing value imputation was done using R Studio’s MICE package

Model	Accuracy		Area under ROC	
	Original Data	Imputed Data	Original Data	Imputed Data
J48	94.92%	94.965%	0.616	0.624
Logistic Regression	94.91%	94.914%	0.667	0.686
Naïve Bayes	91.58%	91.980%	0.667	0.672
Bagging with 70 iterations	94.53%	95.024%	0.752	0.760
Random Forest with 100 iterations	94.78%	95.032%	0.748	0.754

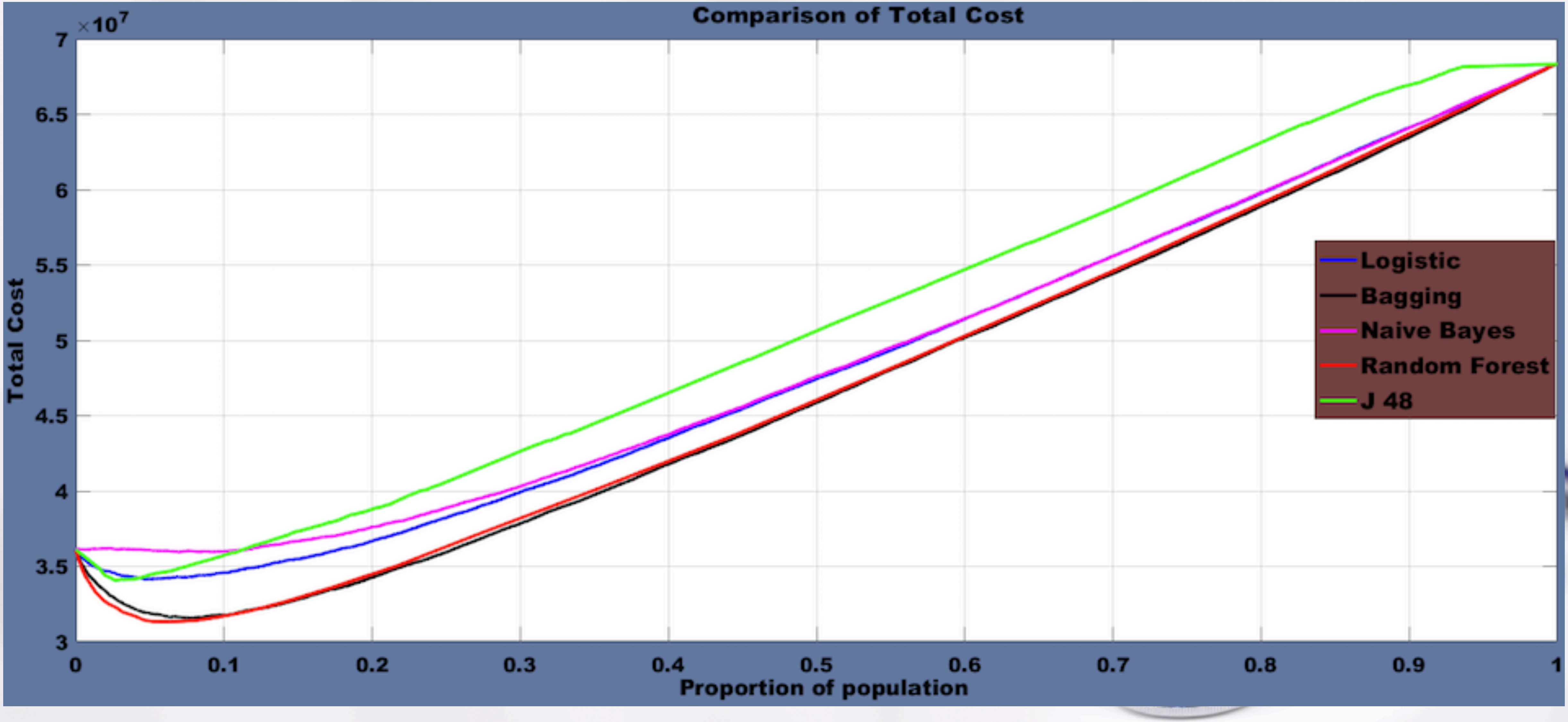
Above results are using 10 fold CV which show improvement with imputed data

Model Building

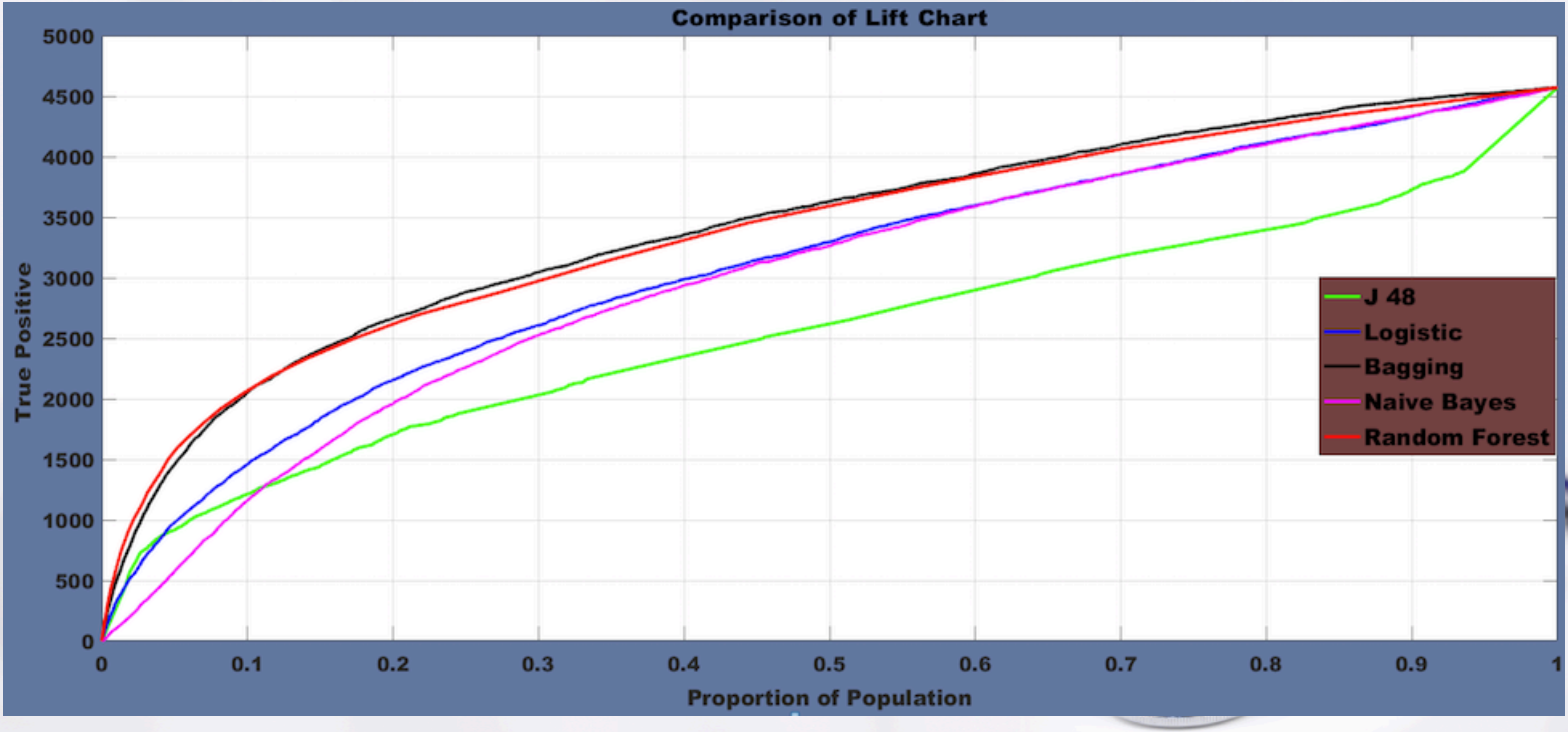
Predictive Models were built using J48, Logistic Regression, Naïve Bayes, Bagging and Random Forest. Below is the comparison using 10 fold Cross Validation

Model	True Positive	False Positive	False Negative	True Negative	Threshold Probability	% of Pop targeted	Total Cost
J48	737	1699	3834	83633	0.125	2.7	$34,079,120
Logistic Regression	965	3362	3606	81970	0.129	4.8	$34,133,000
Naïve Bayes	826	5420	3745	79912	0.2397	6.94	$35,943,260
Bagging with 70 iterations	1833	5053	2738	80279	0.1127	7.66	$31,554,080
Random Forest with 100 iterations	1684	3649	2887	81683	0.15	5.93	$31,328,540

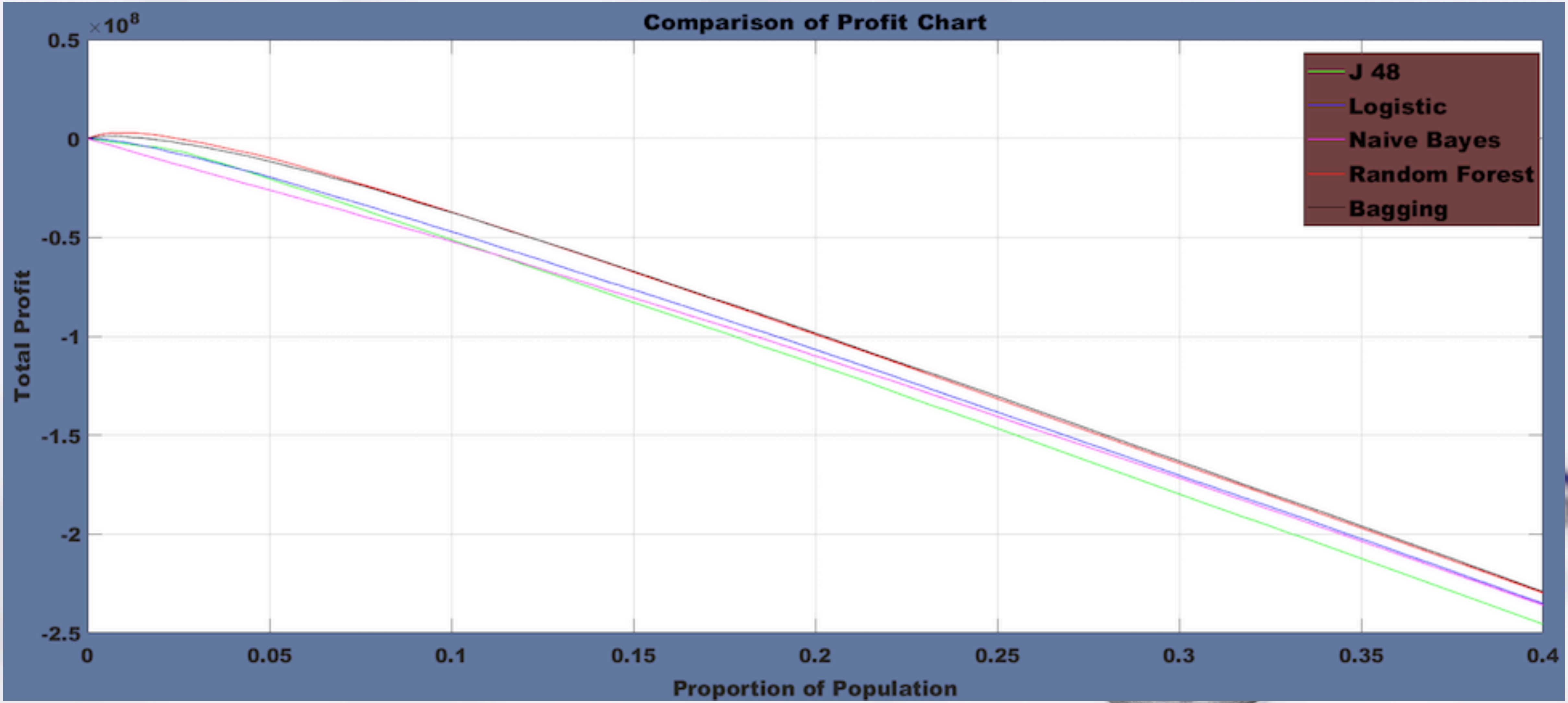
Model Evaluation (Total Cost Chart)



Model Evaluation (Lift Chart)



Model Evaluation (Profit Chart)



Project Results and Insights

- **Random Forest with 100 iterations** has the lowest estimated total cost.
- The total cost for all the prediabetics according to the current practice of using Fasting Blood Glucose > 110 for classifying a patient as “Diabetic” is $**35,639,660.**
- By implementing the Random Forest with 100 iterations model we can save cost by **$4,311,120** which is equivalent to saving **$ 943.15** per prediabetic.
- In US, no. of prediabetics is 86 Million. Using the same model, we can save **$81 Million** over the current practice for 86 Million prediabetics.

Thank You