Final Project

Objective: -

Based on your work in this course to date, prepare a report for the campaign for whichever candidate you choose.

Data Sets and Resources: -

FX_indicators_2020.csv

FX_indicators_2020_rand_10k.csv

FX_indicators_2020_with_candidate_IDs.csv

About the voterfile.pdf

fx_indicator_data_dictionary 2020.xlsx

Tool Used: - Python (Jupyter notebook)

Models used in this project: -

Decision Tree/Segmentation/Logistic Regression/Linear Regression/KNN/Random Forest/Ensemble Models/Uplift Models used.

1. Overall summary of the statistical approach taken, written in language that is accessible to the campaign management?

- In campaign management project, for each of the specific situation, statistical methods are available for analysis and interpretation of the data. To select the appropriate statistical method, one need to know the assumption and conditions of the statistical methods, so that proper statistical method can be selected for data analysis. Two main statistical methods are used in data analysis: descriptive statistics, which summarizes data using indexes such as mean and median and another is inferential statistics.

- Selection of appropriate statistical method depends on the following three things: Aim and objective of the study, Type and distribution of the data used, and Nature of the observations (paired/unpaired). All type of statistical methods that are used to compare the means are called parametric while statistical methods used to compare other than means (ex-median/mean ranks/proportions) are called nonparametric methods.

- In the present project, we have considered the parametric and non-parametric methods, their assumptions, and how to select appropriate statistical methods for analysis and interpretation of the campaign management project data.

- Selection between Parametric and Nonparametric Methods

Description	**Parametric Methods**	**Nonparametric Methods**
Predict one outcome variable by at least one independent variable	Linear regression model Linear Discriminant Analysis Perceptron Naive Bayes Simple Neural Networks	Nonlinear regression model/Log linear regression model on log normal data/ Decision Tree/ Ensemble Models/Uplift Models k-Nearest Neighbors Decision Trees like CART and C4.5 Support Vector Machines

Descriptive statistics	Mean, Standard deviation	Median, Interquartile range

- Semi-parametric and non-parametric methods

Description	**Statistical methods**	**Data type**
To predict the outcome variable using independent variables	Binary Logistic regression analysis	Outcome variable (two categories), Independent variable (s): Categorical (≥2 categories) or Continuous variables or both
To predict the outcome variable using independent variables	Multinomial Logistic regression analysis	Outcome variable (≥3 categories), Independent variable (s): Categorical (≥2 categories) or continuous variables or both
Area under Curve and cutoff values in the continuous variable	Receiver operating characteristics (ROC) curve	Outcome variable (two categories), Test variable: Continuous
To predict the survival probability of the subjects for the given equal intervals	Life table analysis	Outcome variable (two categories), Follow-up time : Continuous variable

Selection of the appropriate statistical methods is very important for the quality research. It is important that a researcher knows the basic concepts of the statistical methods used to conduct research study that produce a valid and reliable results. There are various statistical methods that can be used in different situations. Each test makes assumptions about the data. These assumptions should be taken into consideration when deciding which the most appropriate test is. Wrong or inappropriate use of statistical methods may lead to defective conclusions, finally would harm the evidence-based practices. Hence, an adequate knowledge of statistics and the appropriate use of statistical tests are important for improving and producing quality of political campaign management.

There are many softwares available online (Python/R) as well as offline for analyzing the data, although it is fact that which set of statistical tests are appropriate for the given data and study objective is still very difficult for the researchers to understand.

2. Specific recommendations for what to do with each voter, messaging-wise

- From assignment 3, built two uplift models predicting how likely it is that a voter will become more likely to support the Democratic candidate based on the test mailings for message A and message B, so does for Republican candidate base on score ratings and messages.

  Such as he/she is more likely to support the Democratic candidate or republican candidate or remains neutral. Weather he/she sticks to his/her party or change his/her partisanship to another party.

- We have combined the two partisanship models (log. Regression and decision tree) made in lesson 2 to create an ensemble model predicting partisanship for democrats with prediction accuracy of 90% and AUC 0.9 that predicts how likely democrat candidate be chosen by voters. Based on Voter Id model score with training and validation data, prediction probability is 0.95 and most of the voters who voted in past as democrats remains democrats.

  The data set has been spilt into full data and small dataset. Quintile lift, Decile lift and gains chart has been plotted for Catboost uplift model that shows 10% got lift and 50 records contains about 80 % of the outcomes.

- We have also built (log. Regression and decision tree) models for predicting candidate support, rather than partisanship. We have chosen 4 supports each for waves.

  wave 1 strong democrat, wave 2 strong democrat, wave 1 strong republican, wave 2 strong democrat. We got prediction accuracy of 85% and AUC 0.8. Based on Voter Id model score with training and validation data, prediction probability is 0.95 and most of the voters who supported in past as democrats/republicans in wave 1 and wave 2 will be most likely to support same candidate in upcoming presidential election as well.

- We have also built model predicting the overall persuadability of voters in FX if voter changed their mind in some way between the first and second waves of IDs. We have chosen 2 persuadability variables" Moved from Republican to Democrat to between wave 1 and wave 2 IDs" and" Moved from Democrat to Republican between wave 1 and wave 2 IDs". We got prediction accuracy of 85% and AUC 0.85. Based on Voter Id model score with training and validation data, prediction probability is 0.9 and most of the voters persuade same party during wave 1 to wave 2 except few voters (5 %) changed their partisanship in wave 2 from democrat to republican.

- We have built Build two uplift models predicting how likely it is that a voter will become more likely to support the Democratic candidate based on the test mailings for message A and message B.

  We have chosen 2 uplift variables" uplift strong democrat messege_A" and "uplift strong democrat messege_B"."

  We got prediction accuracy of 90% and AUC 0.85. Based on Voter Id model score with training and validation data, we got prediction probability distribution for each voter ID and most of the voters persuade Democratic candidate based on the test mailings for message A and message B. with average probability of message A is 95 % and message B is 87%.

3. A technical section to document what was done, covering handling and data-prep, model-building, model assessment and scoring.

- For technical section document what was done, covering handling and data-prep, model-building, model assessment and scoring, see attached reports and Python code files.

  Assignment 2 APA track and TESU Masters.zip

  Assignment 3 APA track and TESU Masters.zip

References: -

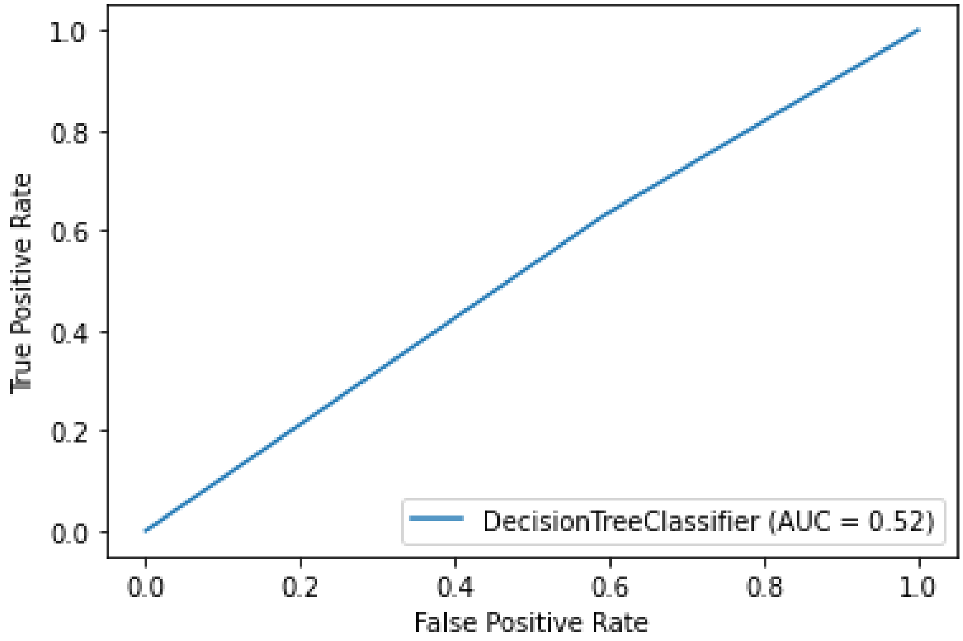
Assignment 2 APA track and TESU Masters

Assignment 3 APA track and TESU Masters

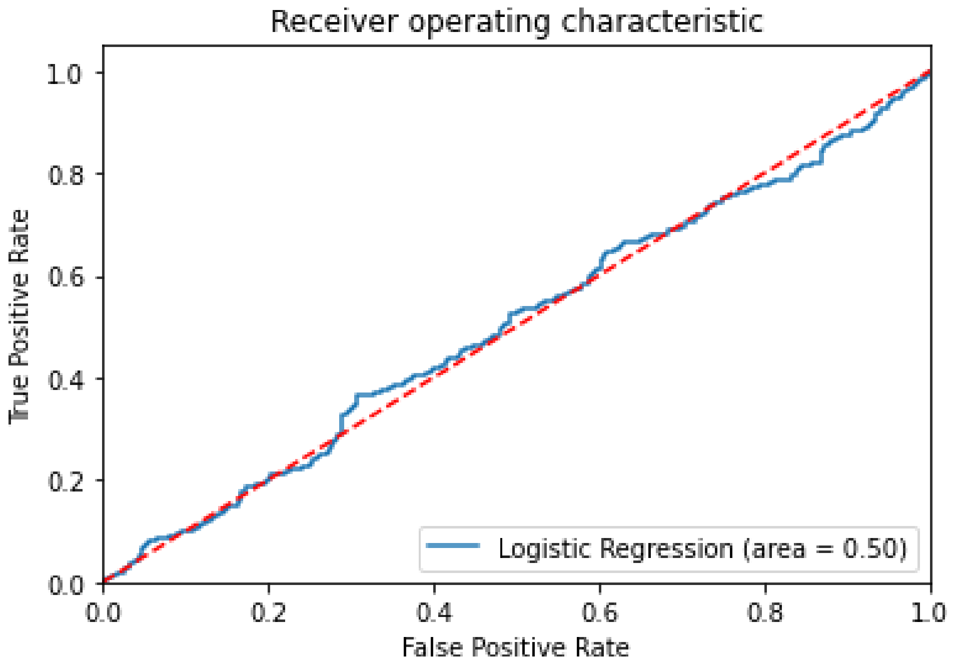
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6639881/

Appendix: -

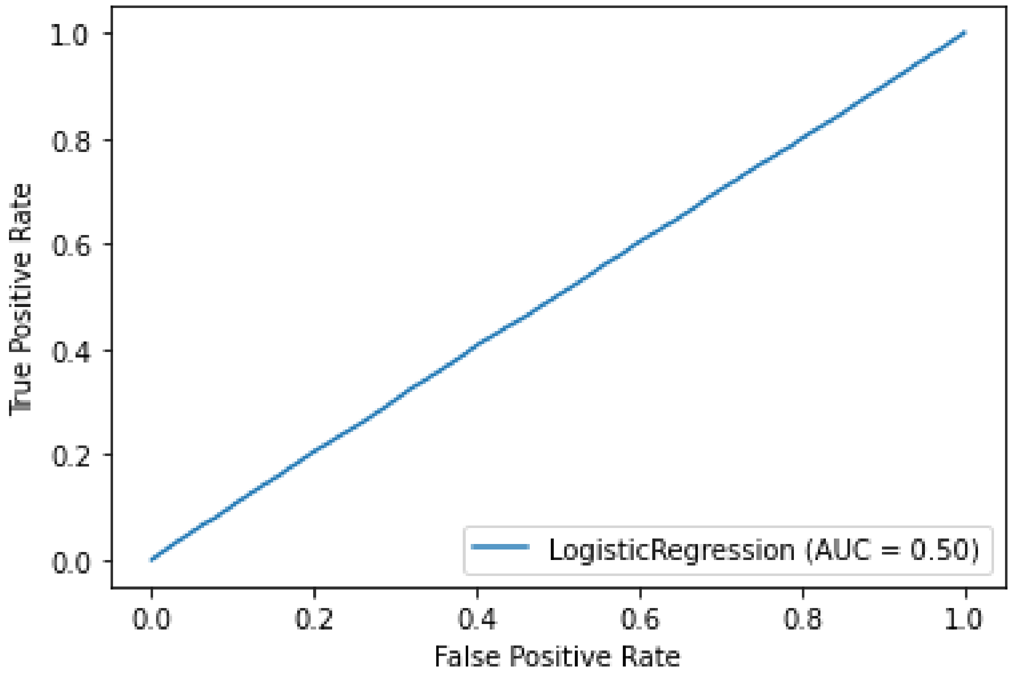
Assignment 2 APA track and TESU Masters – Graphs



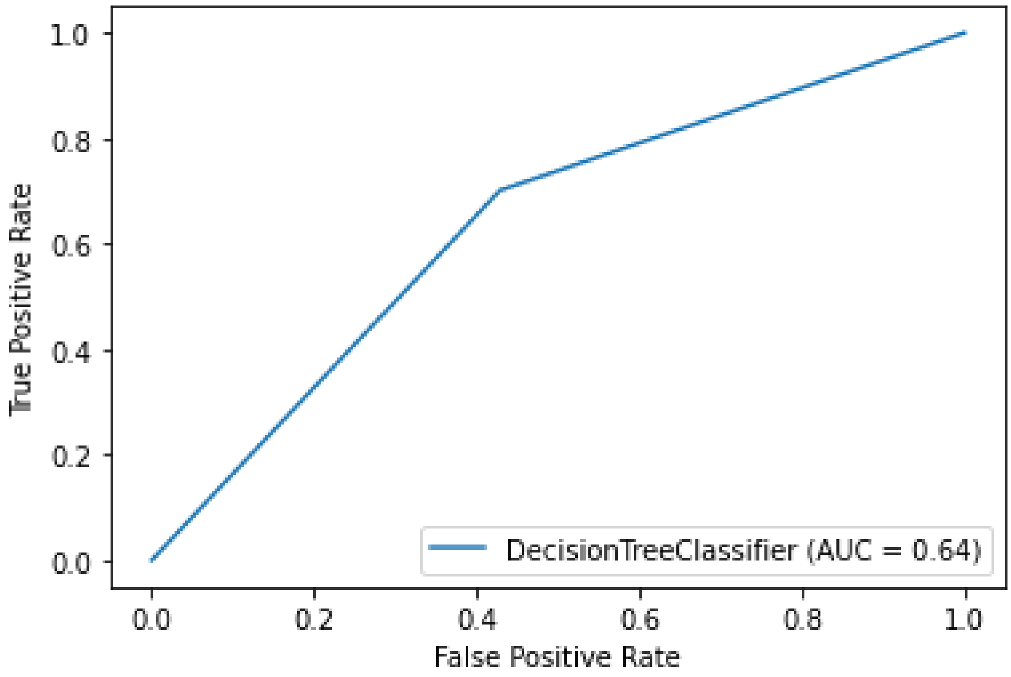
roc_curve for Decision Tree classifier for large and small data. (Predicting partisanship)



roc_curve for logistic regression for large and small data. (Predicting partisanship)



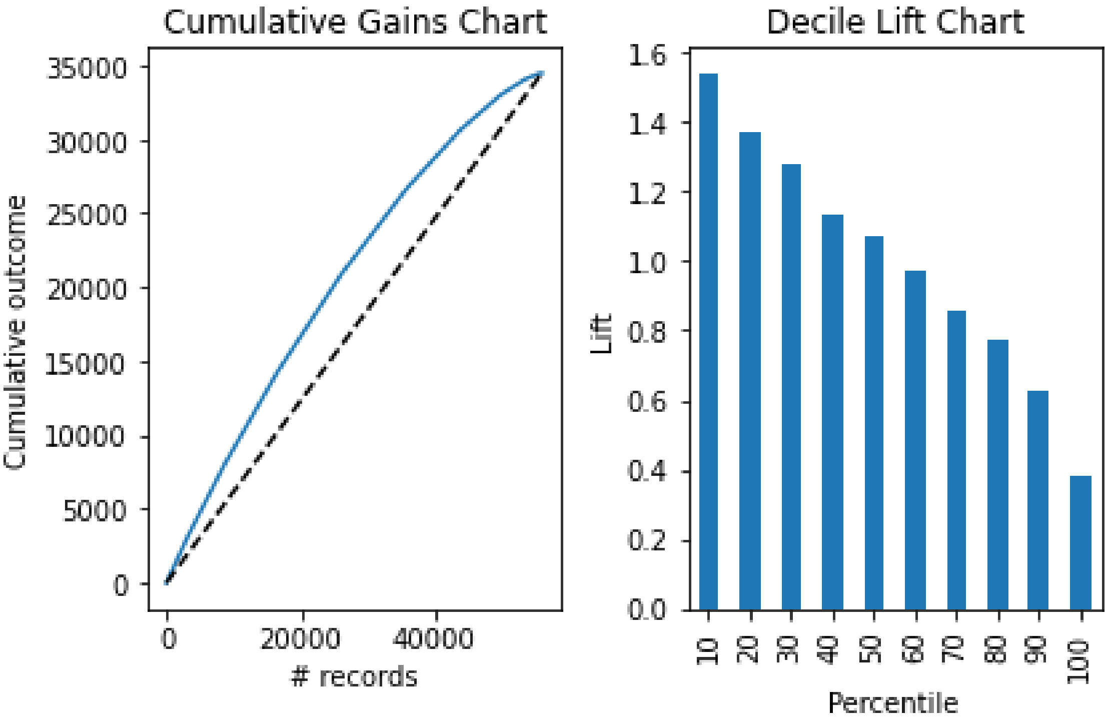
roc_curve for logistic regression for large and small data. (Predicting turnout)



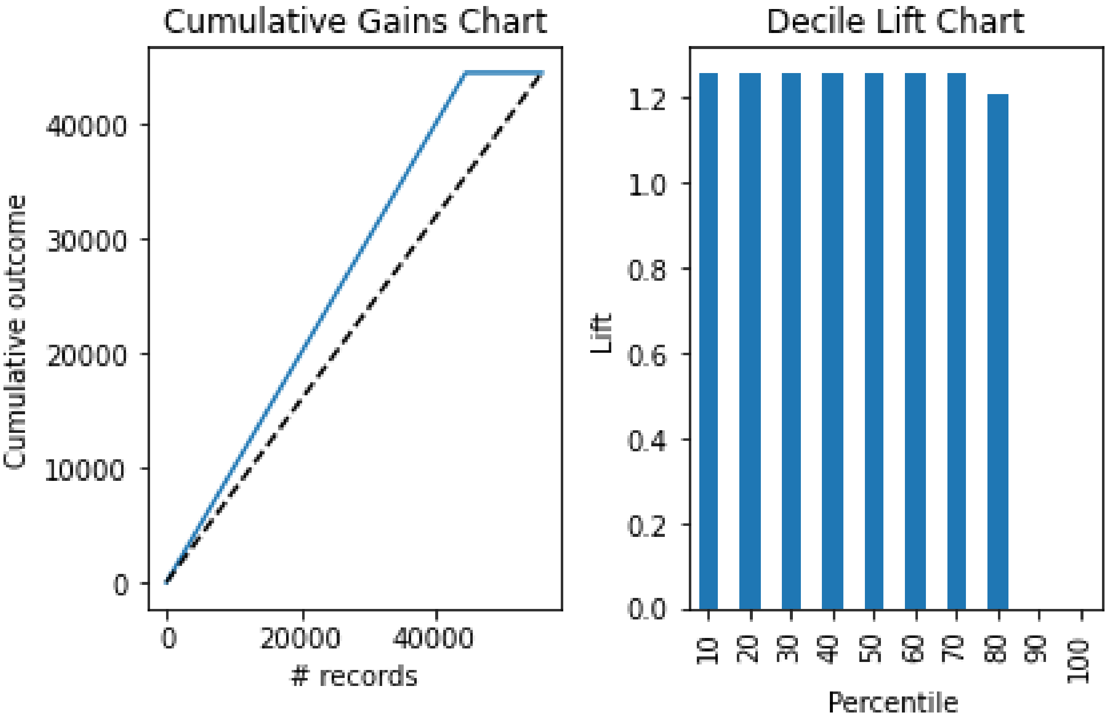
roc_curve for Decision Tree Classifier for large and small data. (Predicting turnout)

Assignment 3 APA track and TESU Masters – Graphs

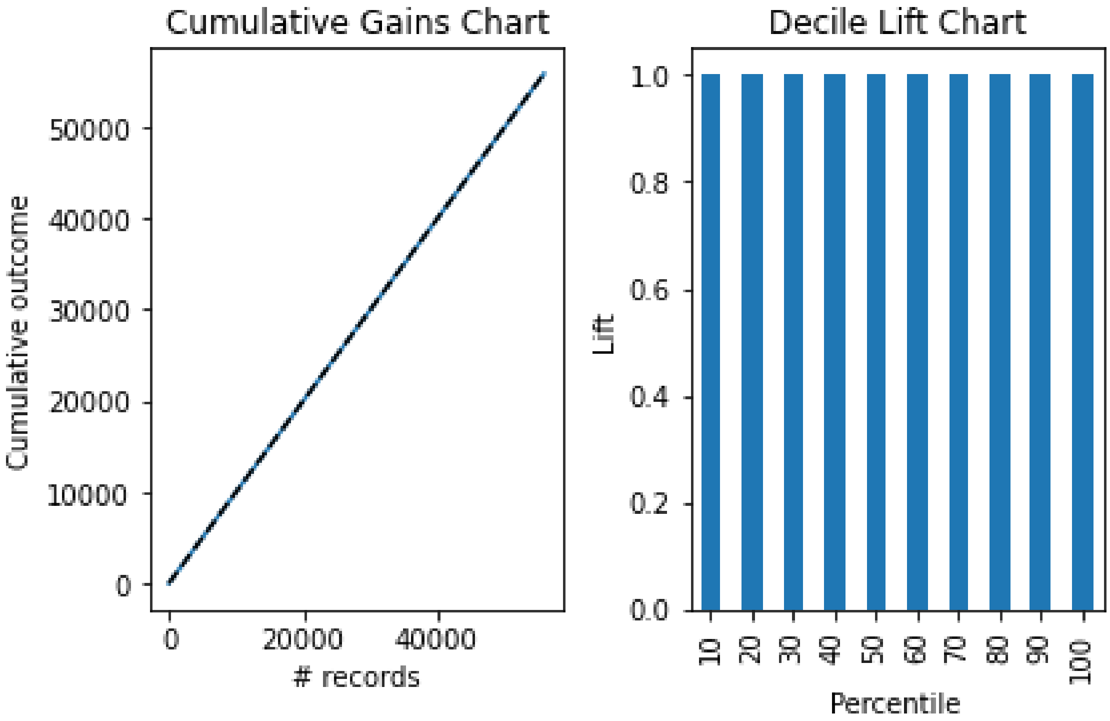
ensemble model predicting partisanship: -



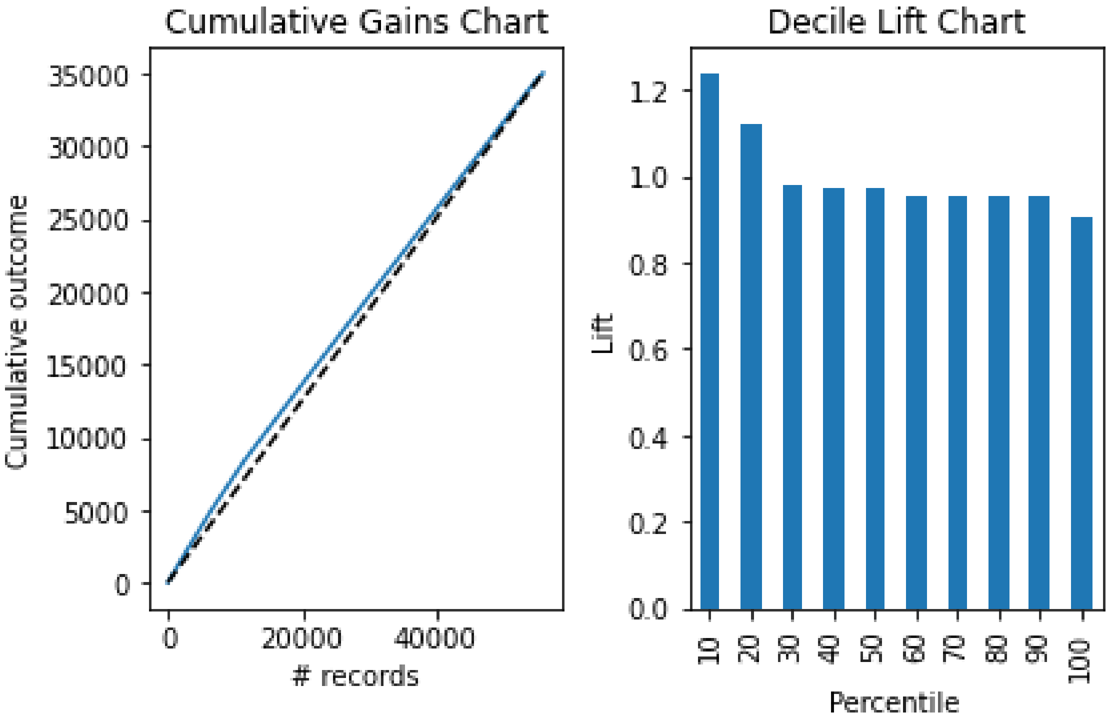
generating cumulative gains and decile for bagged tree



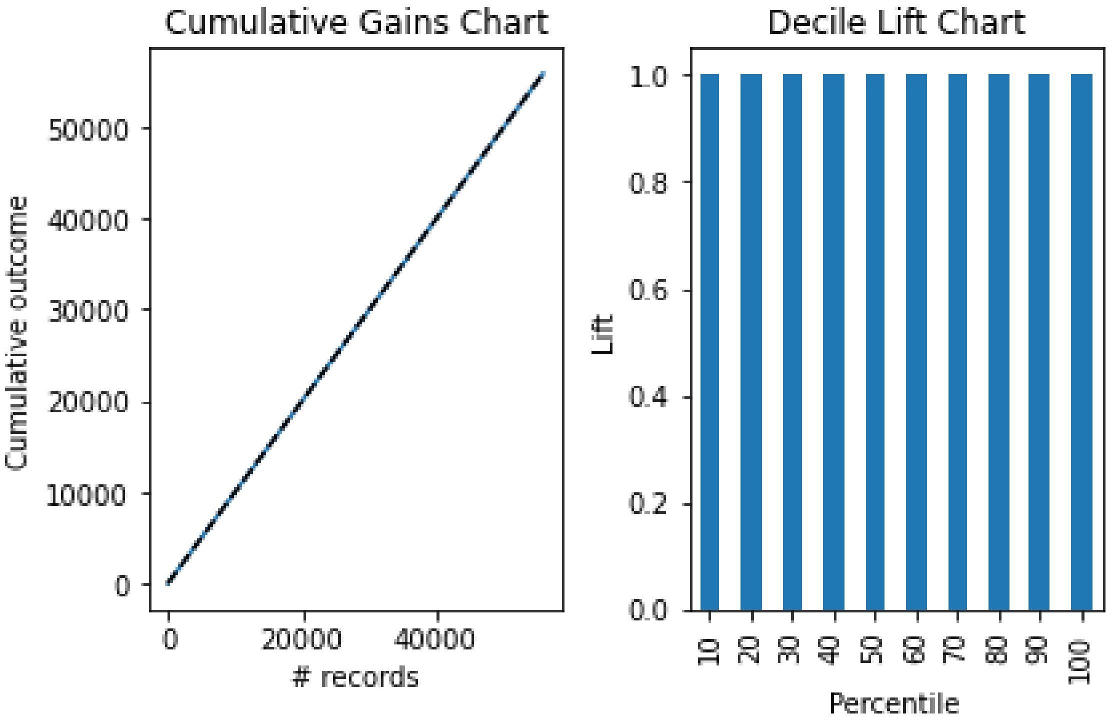
generating cumulative gains and decile for random forest



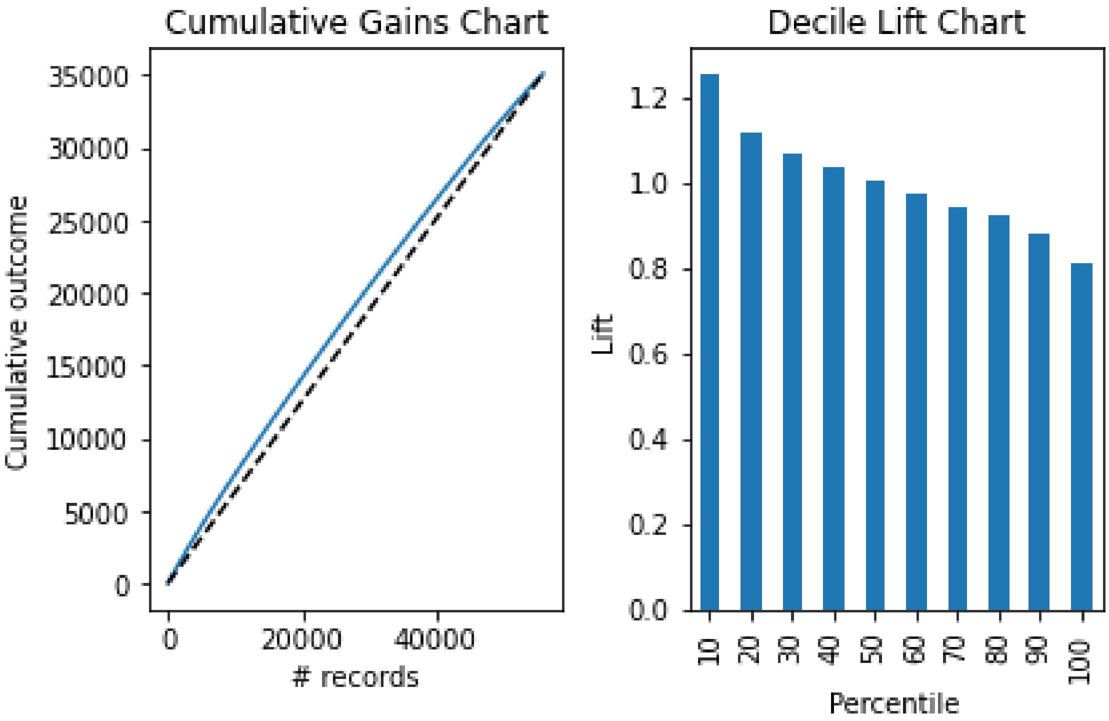
generating cumulative gains and decile AdaBoost with classifier



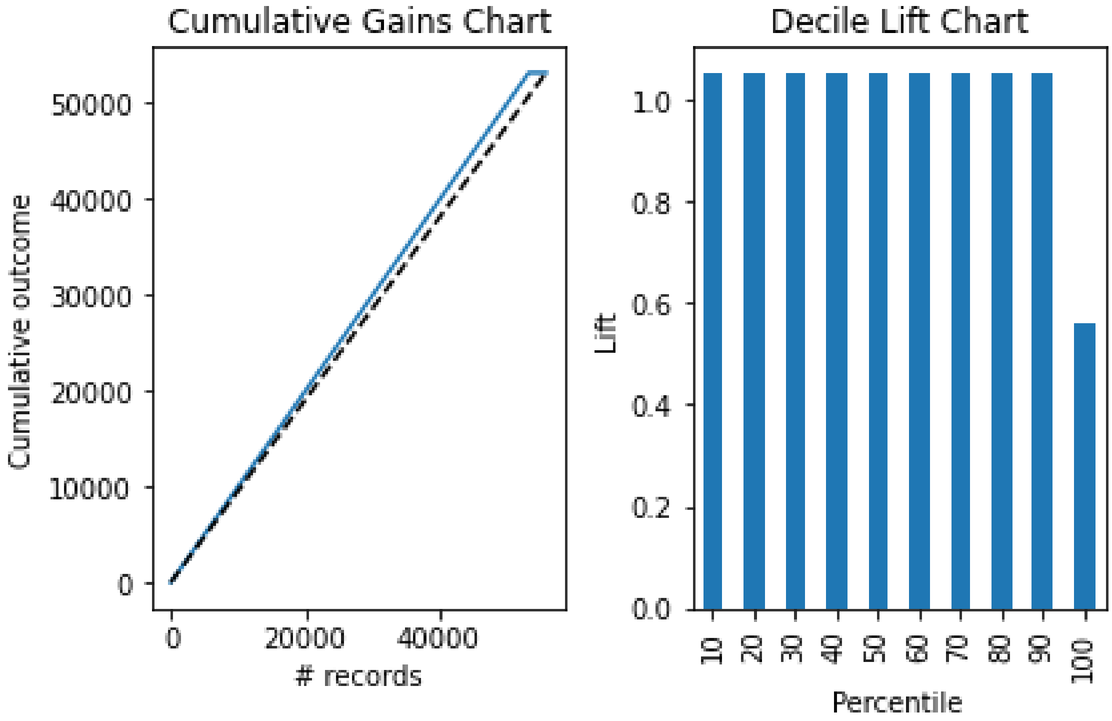
generating cumulative gains and decile AdaBoost with regressor



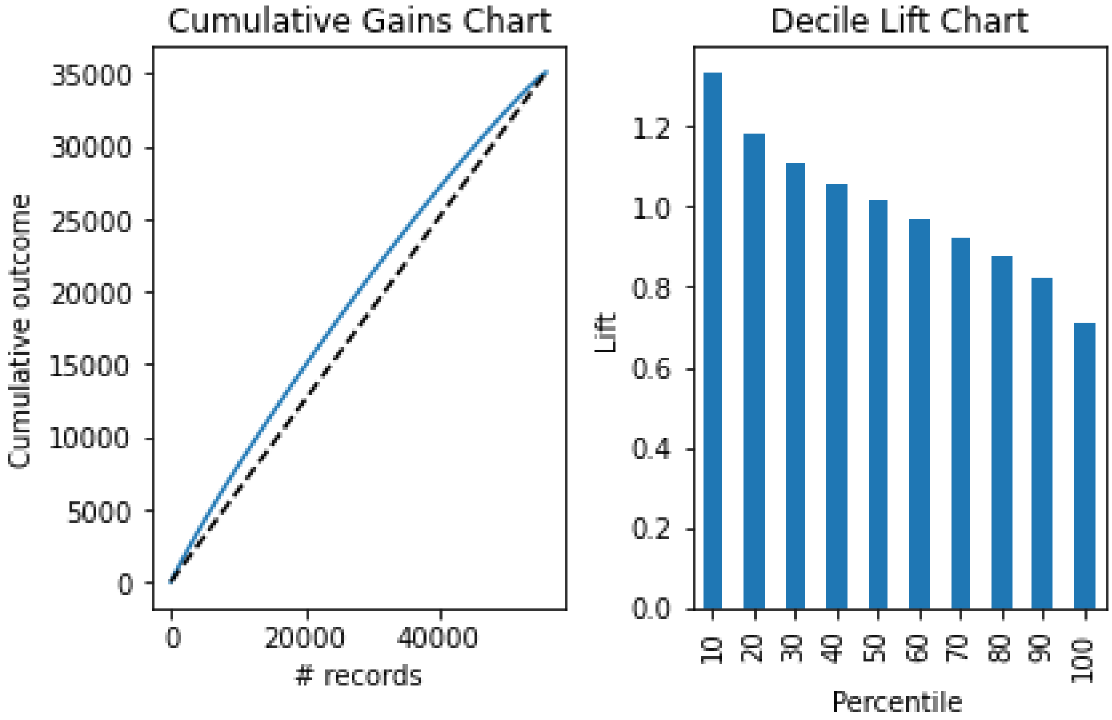
generating cumulative gains and decile # Gradient Boosting (GBM) with Classifier



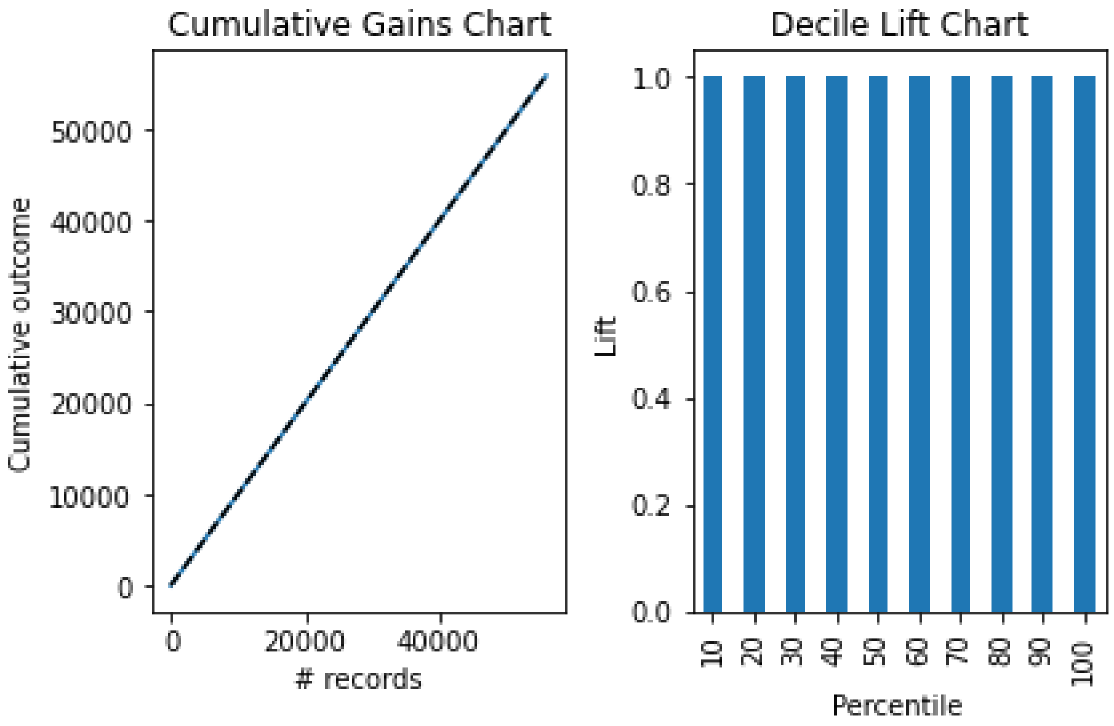
generating cumulative gains and decile # Gradient Boosting (GBM) with regressor



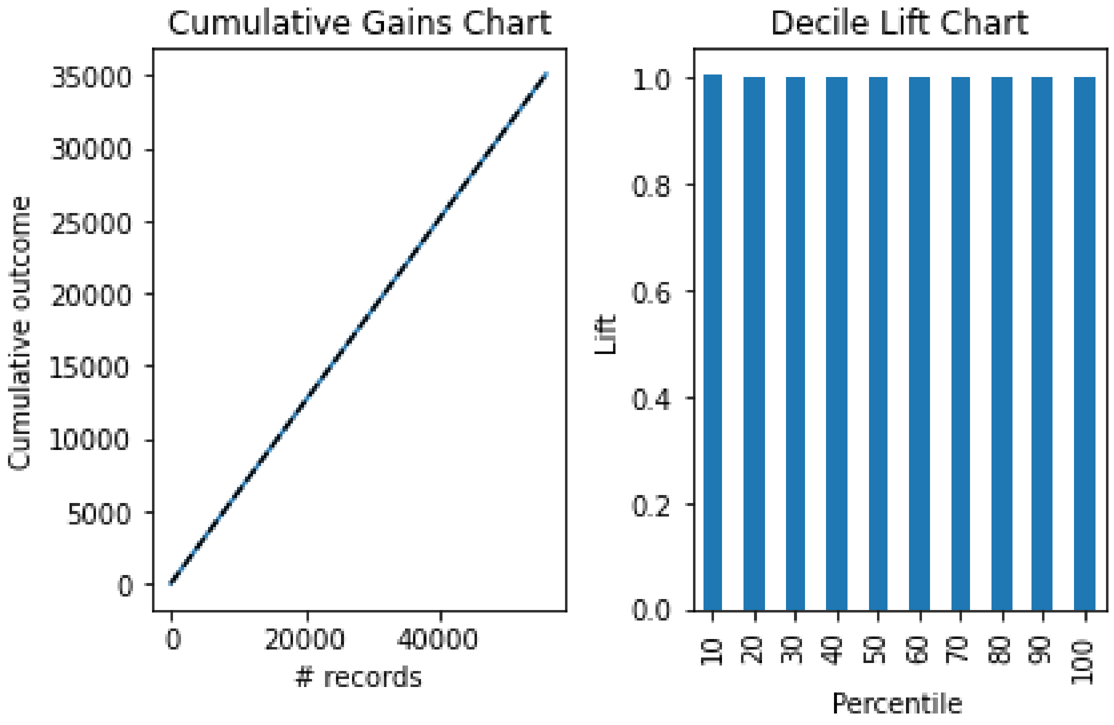
generating cumulative gains and decile # XGBoost with Classifier



generating cumulative gains and decile # XGBoost with XGBRegressor

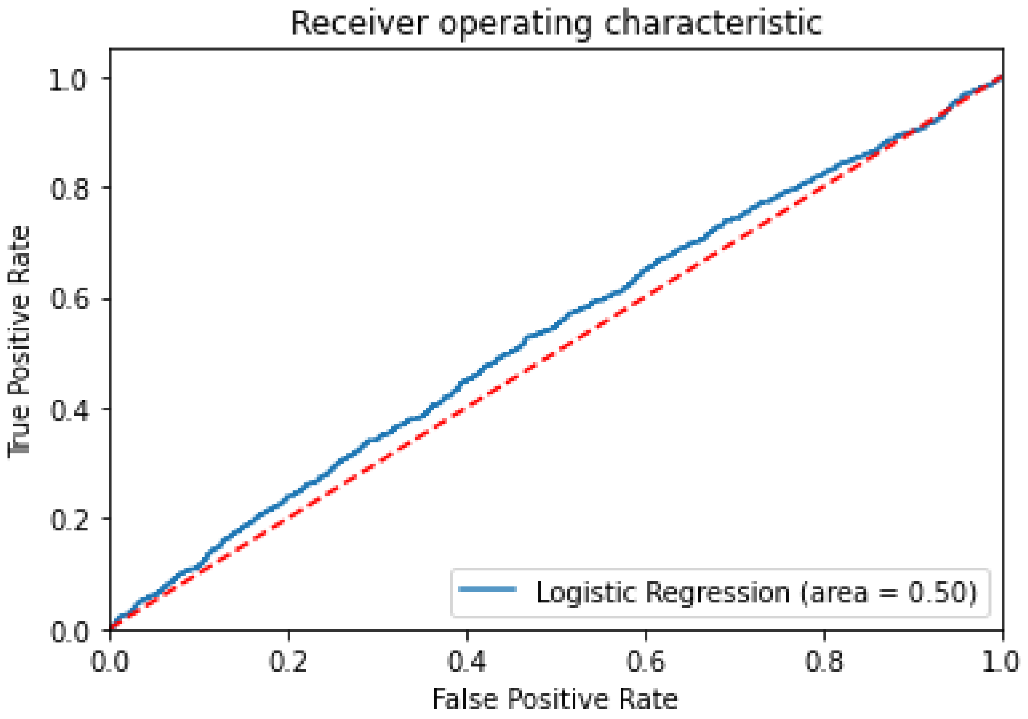


generating cumulative gains and decile catboost with classifier

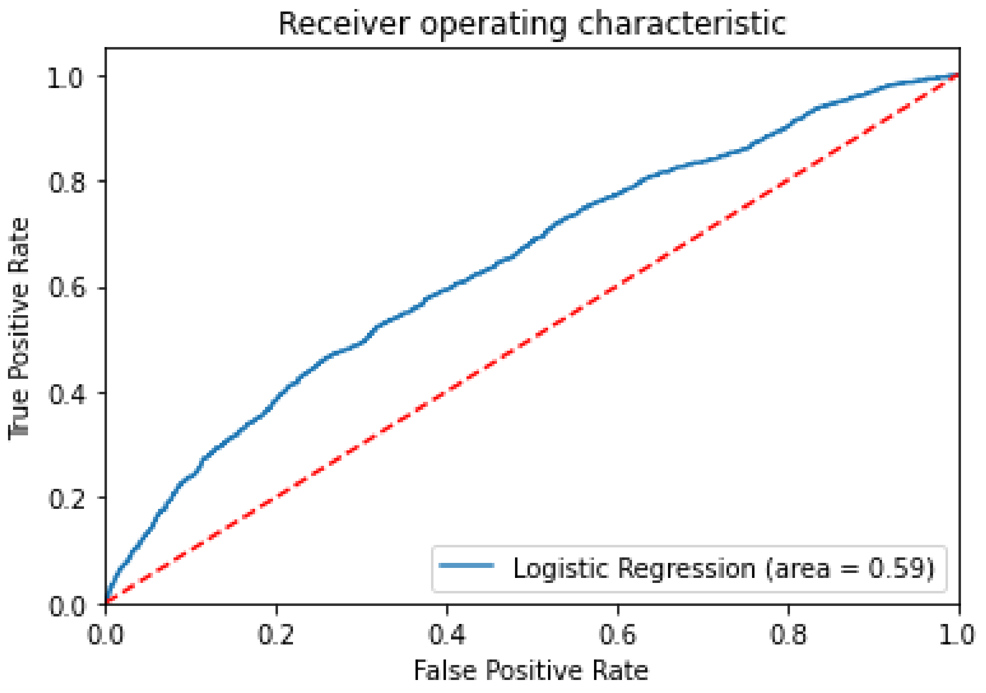


generating cumulative gains and decile catboost with regression

Build log regression models predicting candidate support (y1 = CAND1_SDA_Y, wave 1 strong all way democrat)

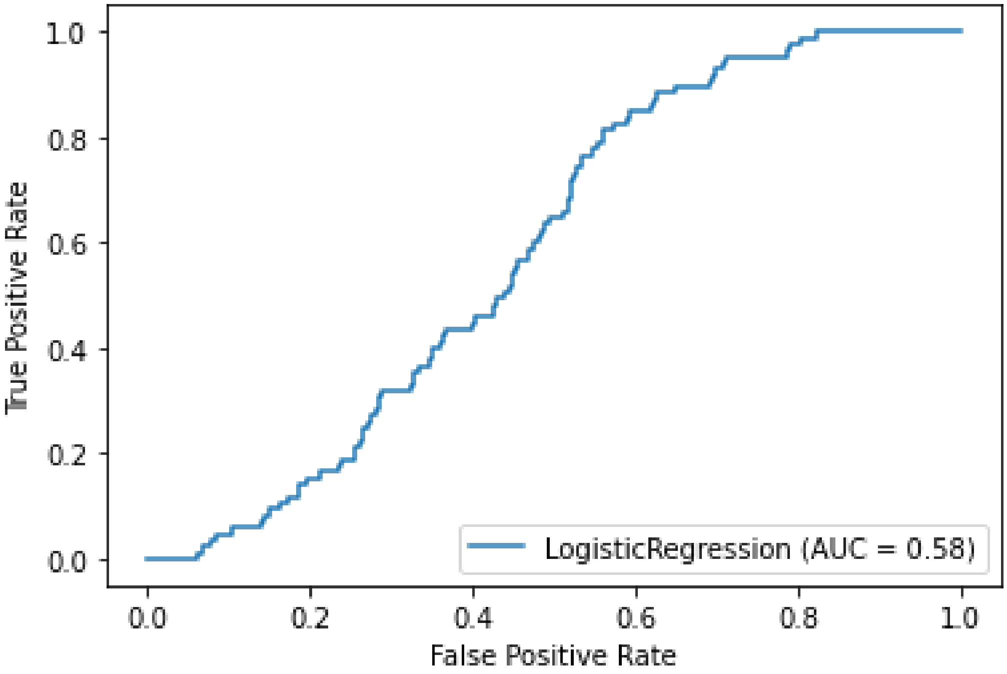


Build log regression models predicting candidate support (y4 = CAND2_SRA_Y, wave 2 strong all way republican)

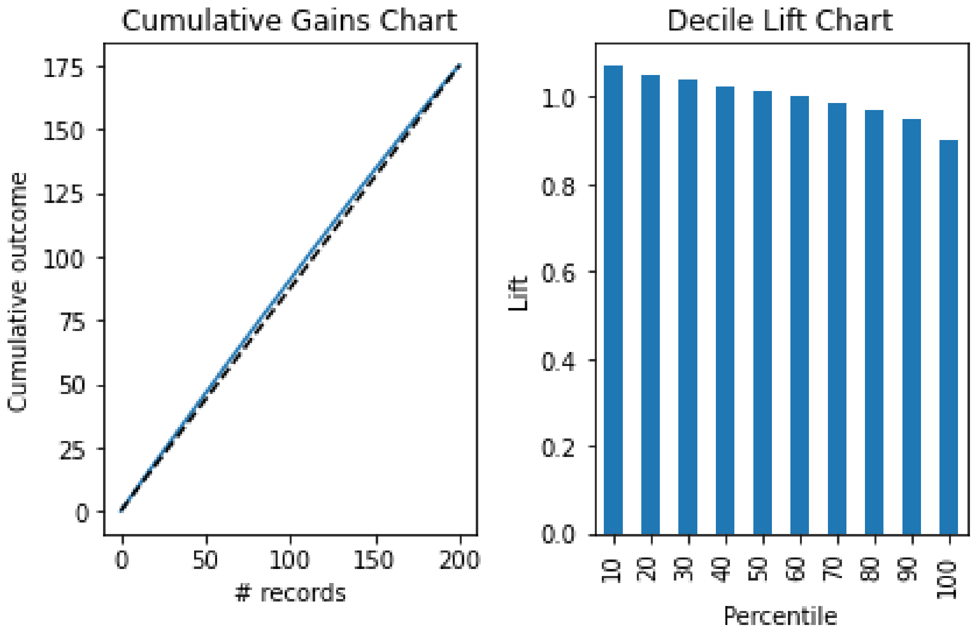


Q5. [3 points] Build a model predicting the overall persuadability of voters in FX (wave 1 to wave 2)

Build log regression model predicting persuadability of voters in FX for candidate support (y1 = CAND1_SDA_Y, wave 1 strong all way democrat) to (y3 = CAND2_SDA_Y, wave 2 strong all way democrat)



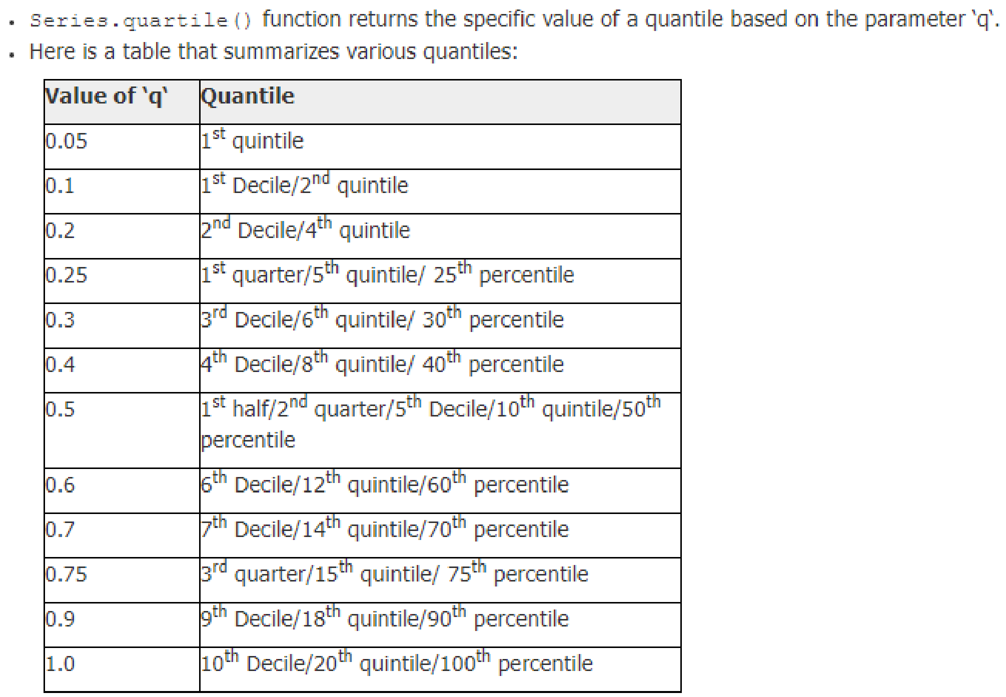
Q6. [4 points] Build two uplift models



generating cumulative gains and decile for Uplift_Catboost

q7 df voter ID modelscore log reg quintile (scores for Messge A & Message B)

```
First Quartile:0.85
Second Quartile:0.88
Third Quartile:0.91
100th Percentile:0.95
1st Percentile:0.82
```



- `Series.quartile()` function returns the specific value of a quantile based on the parameter ‘q‘.
- Here is a table that summarizes various quantiles:

Value of ‘q‘	Quantile
0.05	1st quintile
0.1	1st Decile/2nd quintile
0.2	2nd Decile/4th quintile
0.25	1st quarter/5th quintile/ 25th percentile
0.3	3rd Decile/6th quintile/ 30th percentile
0.4	4th Decile/8th quintile/ 40th percentile
0.5	1st half/2nd quarter/5th Decile/10th quintile/50th percentile
0.6	6th Decile/12th quintile/60th percentile
0.7	7th Decile/14th quintile/70th percentile
0.75	3rd quarter/15th quintile/ 75th percentile
0.9	9th Decile/18th quintile/90th percentile
1.0	10th Decile/20th quintile/100th percentile

quintile for partisanship model log reg

```
First Quartile:1.00
Second Quartile:1.00
Third Quartile:1.00
100th Percentile:1.00
1st Percentile:1.00
```

quintile for partisanship model decison tree

```
First Quartile:0.00
Second Quartile:1.00
Third Quartile:1.00
100th Percentile:1.00
1st Percentile:0.00
```