Three ideas for Capstone project2

Idea 1. Election campaign data

1.    What is the problem you want to solve?

Can campaign finance data predict election outcome?

2.    Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

This problem is faced by politicians – incumbents as well as new candidates.

They can take proactive steps to win elections

3.    What data are you going to use for this? How will you acquire this data?

I would like to use Kaggle dataset for election finance.

<https://www.kaggle.com/danerbland/electionfinance/data>

It has 1814 data points with 51 columns.

4.    In brief, outline your approach to solving this problem (knowing that this might change later).

I will start with data visualization using matplotlib.

I would like create a logistic regression model for prediction of election outcome and check the health of the model. Jupyter notebook will be used.

I would also like to use decision tree model to figure out which columns are significant . This model will also suggest the possible ways of winning the election.

Lastly, the artificial neural network will be used to check whether that provides a better model than the logistic regression to predict the outcome.

5.    What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The proposal will be part of a github repository for my project. All code and further documentation I write will be added to this repository.

I will also create a power point presentation for the entire project work.

Idea2. Predicting effectiveness of Marketing campaign

Predicting the effectiveness of a marketing campaign

1.    What is the problem you want to solve?

Banks and financial institutions have huge data and information about their customers. This data can be used to create and keep clear relationship and connection with the customers in order to target them individually for definite products or banking offers. Usually,the selected customers are contacted directly through:personal contact, telephone cellular, mail, and email or any other contacts to advertise the new product/service or give an offer, this kind of marketing is called direct marketing. In fact,direct marketing is the main strategy of many of the banks, financial institutions and insurance companies for interacting with their customers.

But, the challenge is to improve the effectiveness of these direct marketing campaigns. They are often ‘hit and miss’. However, if they are effective, they have tremendous impact on the topline.

2.    Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

This problem is faced by banks, financial institutions and insurance companies during their targeted product marketing efforts.

I run a Financial services company I do engage in direct product sales and marketing. I face this problem everyday and I would like to solve this problem.

The cost and effectiveness of the campaigns are two imporatnt parameters for me. More directed and customized the campaign, the more is the effectiveness.

3.    What data are you going to use for this? How will you acquire this data?

I would like to use ‘Bank Marketing Data Set ’ from UCI data reporsitory.

It has 41188 records. I plan to do a train-test split.

<https://archive.ics.uci.edu/ml/datasets/bank+marketing>

4.    In brief, outline your approach to solving this problem (knowing that this might change later).

I will start with data clean up and data wrangling.

Then I will do data visualization using matplotlib. The outcome will be a good data story.

I would like to create a logistic regression model for prediction of effectiveness of the campaigns. I will build the confusion matrix and ROC curve. I will do Hyper parameter tuning for the logistic regression model.

I would also like to use decision tree model, preferably Random forest for predicting effectiveness of the campaigns. This model should also substantiate that the feature set used in logistic regression were correct. If not, I will re- create the model for logistic regression.

Lastly, I will try to the artificial neural network will be used to check whether that provides a better model than the logistic regression to predict the default.

I will try to fine tune my model by varying the nodes and hidden layers.

This is a nice to have. I am not very sure whether I will be able to achieve it or not.

5.    What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The proposal will be part of a github repository for my project. All code and further documentation I write will be added to this repository.

I will also create a power point presentation for the entire project work.

Idea 3. FBI NICS Firearm Background Check Data

1.    What is the problem you want to solve?

# Is there a correlation between FBI NICS (National Instant Criminal Background Check System) background checks and actual firearm sales?

2.    Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise?

This analysis can be used by firearm manufacturers as well as FBI and city police departments.

3.    What data are you going to use for this? How will you acquire this data?

I would like to use dataset from ‘Data is Plural’ .

<https://github.com/BuzzFeedNews/nics-firearm-background-checks>

The FBI provides data on the number of firearm checks by month, state, and type

- currently covers November 1998 – April 2018.

4.    In brief, outline your approach to solving this problem (knowing that this might change later).

I will start with data visualization using matplotlib.

This dataset is ripe for unsupervised learning. There will be several clusters of background checks.

There will be several imterpretations of clusters. For example, does the gun type(Hand Gun, Long Gun etc.) and its background check numbers tell any story?

Does the type of use tell any story? From example the inquiry came from pre-pawn, Return, or Rental place etc.

Do the sheer number based on the state tell any story? And many other.

I would also like to use decision tree model to figure out which columns are significant . Random Forest would be one that I will experiment with in this case.

5.    What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The proposal will be part of a github repository for my project. All code and further documentation I write will be added to this repository.

I will also create a power point presentation for the entire project work.