Capstone_Ideas

Ish Gupta

October 25, 2016

Capstone Project Ideas.

Please have a look at below 3 proposals, and suggest your views.

Idea #1:

OutBrain Click Prediction:

Background about Outbrain

Outbrain is an online advertiser specialized in presenting sponsored website links, which recommends articles, blog posts, photos or videos to a reader based on their Behavior. The sites with the recommended articles pay Outbrain for the service, and OUtbrain pays the Site hosting the links.

Outbrain's algorithm module determines which content within the network is relevant to individual users. A larger set of algorithms is run in parallel to determine a set of candidate recommendations. The decision of which recommendations to serve the user is made by machine learning techniques. The algorithmic methods Outbrain uses can be divided into numerous categories, some examples are: contextual, behavioral and personal algorithms. Because Outbrain's algorithms make use of HTTP cookies planted on the local computers of the end users, any clearing of those cookies will affect the recommendations that Outbrain makes.

Agenda of the Analysis

Currently, Outbrain pairs relevant content with curious readers in about 250 billion personalized recommendations every month across many thousands of sites. In this competition, Kagglers are challenged to predict which pieces of content its global base of users are likely to click on. Improving Outbrain's recommendation algorithm will mean more users uncover stories that satisfy their individual tastes.

The data provided contains: * The logs of Users visiting links, and the related documents * Details of the documents which can help relate the User's interests. * Location of the User * Details of available promotional content, which links to the document. * Display details of the advertisement. * Platform being used by the User.

This analysis will give me a chance to deal with a real Business problem, which is also huge (2 billion rows). It will be quite a learning to deal with such a data in R.

Question: How much minimum RAM would I need to work on machine? Current config: RAM- 8 GB, i7, Windows 10.

- For More details, visit Kaggle (https://www.kaggle.com/c/outbrain-click-prediction)

Idea #2:

Allstate Claims Severity

Background about AllState:

The Allstate Corporation is the second largest personal lines insurer in the United States (behind State Farm) and the largest that is publicly held. The company also has personal lines insurance operations in Canada.

Agenda of the Analysis:

Allstate is currently developing automated methods of predicting the cost, and hence severity, of claims. With this analysis, Allstate needs an algorithm which accurately predicts **claims Severity**.

AllState has provided training and test data which needs to be used to predict the loss against each Customer depending upon the relative data points.

Random Forest/ Ranger can be utilized to explore the data, compute importance of the predictors and predict the dependent variable, “loss”.

Predictor variables are encoded, and do not have a desription. There are a total of 116 categorical, and 14 continuous.

- More details are available at Kaggle (https://www.kaggle.com/ishgupta/allstate-claims-severity)

Idea #3:

Santander Product Recommendation

Background about Santander Bank:

Santander Bank, N. A. (formerly Sovereign Bank) is a wholly owned subsidiary of the Spanish Santander Group. Based in Boston, Massachusetts, the bank—whose principal market is in the Northeastern United States—has more than $77 billion in assets, operates 723 retail banking offices, over 2,300 ATMs (including 1,100 in CVS pharmacies throughout the Northeast) and employs approximately 9,000 people.[2] Santander offers an array of financial services and products including retail banking, mortgages, corporate banking, cash management, credit card, capital markets, trust and wealth management, and insurance.

Agenda of the Analysis:

To support needs for a range of financial decisions, Santander Bank offers a lending hand to their customers through personalized product recommendations.

Under their current system, a small number of Santander’s customers receive many recommendations while many others rarely see any resulting in an uneven customer experience. Santander wants to predict which products their existing customers will use in the next month based on their past behavior and that of similar customers.

With a more effective recommendation system in place, Santander can better meet the individual needs of all customers and ensure their satisfaction no matter where they are in life.

Data

The data of Customer Behavior for 1.5 years is available which provides training and test sets of the data to train a Model for prediction. The data provides the product a Customer has already purchased, and model should be able to predict the products, a Customer is expected to purchase.

The data is available at Kaggle (https://www.kaggle.com/c/santander-product-recommendation/data)

Summary:

Summary of the Proposals

Provider	Data Size Very Huge?	Real Problem	Real Data	Training Set Dimension	Test Set Dimension	Max Data Size in a file	Preference
Outbrain	Yes (100 GB)	Yes	Yes	87,141,731 X 3	32,225,162 X 2	2,000,000,000 X 6	1
Santander	No	Yes	Yes	13,647,309 X 48	929,615 X 24	13,647,309 X 48	2
AllState	No	Yes	Yes	188,318 X 132	125,546 X 131	188,318 X 132	3