Capstone Project Proposal

Avazu Click Through Rate Prediction

(https://www.kaggle.com/c/avazu-ctr-prediction)

Domain Background

In domain of web advertisement Click Through Rate (CTR) is defined as the ratio of total number of people clicking on an ad and total number of people visiting the webpage. Similarly, we can compute click through rate of any component of our webpage.

Click through rate is one of the important metrics for companies that have online products. It helps them to analyze their products which leads to better design, increased customer satisfaction and revenue growth.

Problem Statement

A very important challenge for an online company is to estimate the click through rate of the web component (an ad or a link). This estimate can help them evaluate the effectiveness of the component and a model for estimation can also be used to design future components in such a way that chances of their effectiveness increases.

Take example of an ad space on a web page. If you can effectively model that a user will click or show interest in one of the ads displayed on your webpage, you can estimate how much money that ad will make for you. From there you can design recommendation systems to display ads that maximizes your advertisement revenue.

For this project I would try to create a machine learning model (supervised classification) which can predict if a user will click on an advertisement campaign run by Avazu an Advertisement Platform. The data consist of 11 days of user activity (approximately 40 million impressions/page views). The data was uploaded on Kaggle by Avazu for an online competition 2 years ago.

Datasets and Input

The dataset provided is already divided in train and test set. Train set contains approximately 35 million rows (data for 10 days) and test set contains approximately 5 million rows (data for 1 day). The set of fields are as follows:

- **id: ad identifier**
- **click: 0/1 for non-click/click**

- **hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.**
- **C1 -- anonymized categorical variable**
- **banner_pos**
- **site_id**
- **site_domain**
- **site_category**
- **app_id**
- **app_domain**
- **app_category**
- **device_id**
- **device_ip**
- **device_model**
- **device_type**
- **device_conn_type**
- **C14-C21 -- anonymized categorical variables**

In the dataset, the field click is our target variable which is a binary variable determining if the user clicked on the ad or not. Rest of the fields are features that will be used by model to predict the target variable. Some of the fields are anonymized categorical variable representing some feature of user or web-platform.

Solution Statement

As the size of dataset is large (approximately 7 GB) the challenge would be to select a model which can perform online learning (learn by examples iteratively instead of batch learning). I intend to explore various online learning algorithms and determine their efficiency in modelling the user behavior. If time would permit I would also like to explore if deep neural networks would be efficient in modelling the user click behavior

Benchmark Model

I plan to use a simple logistic regression classifier as my benchmark model.

Evaluation Metric

I am planning to use log loss and F1-score as metrics to evaluate the performance of the model.

Project Design

I am planning to use python2, pandas, ipython (jupyter), scikit-learn, numpy and matplotlib to perform various tasks required for this project.

I intend to divide my project in following sections:

1. Exploratory Data Analysis
2. Data Processing
    a. Outlier detection
    b. Missing data
3. Feature selection, feature scaling, feature transformation.
4. Model selection
5. Implementation and validation
6. Model Evaluation
7. Performance tuning
8. Conclusions