# Title: AD Click Prediction

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Ankit | Ankit | aankit@hawk.iit.edu |
| Mayank | Goswami | [mgoswami1@hawk.iit.edu](mailto:mgoswami1@hawk.iit.edu) |
| Venkata Narasimha Sri Dattu | Manapragada | vmanapragada@hawk.iit.edu |

1. **Introduction**This project is aimed to train a prediction model to predict the click-through rate for advertisements based on features of the users. Nowadays, internet is full of different kind of advertisements through banners, ads in form of text, images, flash, video, and audio. Ads are used by companies to increase brand awareness, connect and engage with potential customers and hence it is important for the companies to know where to advertise and how successful the advertising is going on a platform. A prediction and recommendation system will be implemented to predict the click through rate of potential users and will suggest target users based on user preferences. Our prediction and recommendation system based on the click probability will make it possible for advertiser to know the probability of the user clicking the advertisement. It’ll support in deciding of putting the advertisement on the certain website so that advertisements only target at the users who might be interested in the ads, thus maximizing the profit of the advertisers and saving the excessive resource which is wasted on useless advertising. This project utilizes the various libraries available in Python to build the model.

1. **Data Sets**Dataset is of American Express data collected from Kaggle.  
   Source: <https://www.kaggle.com/huzefalw/amex-hackathon>  
   Size of the train data is 463290  
   Size of test data is 128857

*Description of Data Variables*

**Train Data:**

|  |  |
| --- | --- |
| Feature | Definition |
| session\_id (Nominal) | Unique ID for every session |
| DateTime (DateTime) | TimeStamp |
| user\_id (Nominal) | User ID for each user |
| product (Nominal) | Product ID of various products |
| campaign\_id (Nominal) | Unique ID for ad campaign hosted |
| webpage\_id (Nominal) | Webpage ID at which the ad is displayed |
| product\_category\_1 (Nominal) | Product category 1 (Ordered) |
| product\_category\_2 (Nominal) | Product category 2 |
| user\_group\_id (Nominal) | Customer segmentation ID |
| gender (Nominal : Binary, Symmetric) | Gender of the user |
| age\_level (Nominal) | Age level of the user |
| user\_depth (Nominal : Ordinal) | Interaction level of user with the web platform (1 - low, 2 - medium, 3 - High) |
| city\_development\_index (Nominal) | Scaled development index of the residence city |
| var\_1 (Nominal : Binary) | Anonymous variable |
| is\_click (Nominal : Binary, Symmetric) | User will click on the advertisement or not (0 or 1) |

**Research Problems***Problem we want to solve:*  
On internet while browsing there are various advertisements and companies in the various services domain leverage huge internet traffic arriving at their websites by strategically placing ads/promotions for cross selling of various products on web pages. Some of them are display ads which incorporates text, logos and pictures or images positioned on a website when a user browses other are search ads which show up when an internet user types in specific key words or phrases in search of a product or service.  
  
Some advertisements are useful and attractive to us, while other are useless for us and we will never click them. In the cycle of advertisement Websites provide advertising spaces and Advertisers pay to provide advertisement. Advertising center allocate space to advertisement and collect fees and Users browse and click the ad. Hence advertising plays the greatest role and it needs to be effective. We need to understand and placate the psychological needs of users to effectively inspire their click. For example, some users are fond of offers provided by sellers, some users are attracted by good after-sales services and others by high quality and high quantity. Therefore, for different users, the same advertising strategy and website clearly does not apply. So, it is important to understand psychological behavior of users and design modeling to place ads that best meet the user’s psychological needs and their presentation.  
  
Companies believe that a predictive model that forecasts whether a user will click on the ad/promotion would help them extract the maximum value of the advertisement they have posted. We as a team are building a model to predict whether a user will click on an ad or not based on various given features.

We will build an “Ad Click Prediction” model that will predict whether the user will click on ad or not. Our label is “**is\_click**” which is a Binary Feature with values 0 or 1

Since our dataset is more than 1 lakh records, we will use Hold-Out validation.

For making prediction, we will be using Classification model since our label is Nominal. We will make prediction using Logistic Regression, SVM, KNN, Decision Trees and Naïve Bayes models.

Since our data is not very well defined, we will be using various Data Mining techniques to convert it into usable form:  
- Data Cleaning (will need to take care of noisy data and missing values)  
- Data Transformation (since our features possess nominal features hence, we will have to first convert them into Numerical variables)  
- Data Reduction (we need to check for features which are not at all important for our prediction and remove them)

1. **Potential Solutions**To build prediction models, we need to go for required preprocessing as per the model:  
   1.) Data Cleaning: We need to take care of the noisy data and missing values. For missing values in Nominal features, we can use below mentioned techniques:  
   - Use the most frequent value to fill in missing values of that attribute  
   - Use the most frequent value belonging to the same class to fill in the missing values  
   - Build a prediction model (Classification – in case of Nominal feature) to predict missing values (if the feature is extremely important).  
     
   2.) Data Transformation: Since our features possess nominal features hence, we will need to convert them into numerical variables.

3.) Data Reduction: We need to take care of the unnecessary columns in the data set which have very less correlation with the label and eventually drop them.  
  
There will be various other issues while creating our prediction models. For example, corresponding to every prediction model we need to incorporate different preprocessing techniques. Naïve Bayes will consider Nominal values whereas KNN will consider Numerical features, hence corresponding to every model, we need to do preprocessing and finally create prediction model.

1. **Evaluations**We will be using Hold out evaluation since our data set is very large. We will use RMSE, Accuracy for evaluating our model and AIC/BIC during feature selection.
2. **Expected Outcomes**

Expectedly, we will build a model which will predict whether the user will click on the advertisement or not.