**** 

**PG Data Science and Engineering| Great Lakes Institute of Management June’2020**

**Under Guidance of**

Mr. Shashank Shirude

**Under-Taken By:**

Sudeep Kadyan

Apoorv Jain

Abhishek Prasoon

Harshit Bharadwaj

Urvashi Panwar

**Domain:**

**DATASET SOURCE: -**

The dataset is available on UCI Machine Learning Repository. The various attributes about customer’s online activities have been made available using Google Analytics. The dataset consists of feature vectors belonging to 12,330 sessions.

**DATASET NAME: -**

Online Shopper Purchasing Intention dataset

**PROBLEM STATEMENT: -**

The client is facing the problem of low rate of sales conversion and seeks to increase revenue from its online shopping platform on the basis customer’s online activities.

**DOMAIN/BACKGROUND DETAILS: -**

From clothes to groceries to household items, the possibilities in theretailspace are full of promise. The goal of a predicting customer behavior system is to estimate how buyers will behave in the future based on data of previous behaviors. These systems allow retailers to segment customers and perform personalized marketing actions.

**OBJECTIVE: -**

To predict whether or not a customer is going to generate revenue for the organization on the basis of customer’s online activities using various machine learning algorithms. The goal is to derive insights for determining what drives the consumers to purchase the product in order to increase customer loyalty and retention.

**ABSTRACT: -**

“**When the going gets tough. The tough go Online Shopping”**

The Internet plays an important role in the business world today. Internet has created an enormous volume of business transactions all over the world. Besides, it does not only introduce a new way of doing business, it also changes consumer lifestyle. The retail industry took a 180-degree turn with the emergence of online shopping. With the speed and convenience of online retail, it has become easier for consumers to get what they want when they want it. Further, with the onset of global pandemic, the lifestyle of individuals has changed. From working from home to shopping for everything online, people have taken on doing everything from their homes. Online shopping has become the new normal.

The global pandemic has served as one of the most significant opportunities for the E-commerce companies. Not just did they flourish, some organizations have also achieved extraordinary milestones by catering to all the needs of their customers. But with opportunities comes competition, struggle, and pressure. The pressure to outshine everyone else in understanding customer behavior and delivering curated services. Customers enjoy browsing online catalogs because they present products in an attractive way and provide plenty of information about them.

Based on customer behavior, an organization can make changes to its platform and provide fabricated services to its customers. After all, understanding customer needs is the need of the hour. This would not only help them target the right audience but will also increase their chances of being able to convert their visitors into their customers.

“**See your customers as invited guests to your party. It’s your job to make the customer experience a little bit better**”

* **Jeff Bezos**

**DATA DESCRIPTION: -**

|  |  |  |
| --- | --- | --- |
| Input | Description | Comments |
| Train/Test | Whether the data belongs to the train set or test set |  |
| ID | A unique ID to identify a candidate |  |
| DEPENDENT VARIABLES | | |
| Salary | Annual CTC offered to the candidate (in INR) |  |
| DOJ | Date of joining the company | Candidates may not have accurately entered information. |
| DOL | Date of leaving the company | A value of "present" means the candidate continues to work at the company at the time of collecting this information |
| Designation | Designation offered in the job |  |
| JobCity | City in which the candidate is offered the job |  |
| INDEPENDENT VARIABLES | | |
| Gender | Candidate's gender | m denotes Males and f denotes Females |
| DOB | Date of birth of candidate |  |
| 10percentage | Overall marks obtained in grade 10 examinations | Domain of values: [0,100] |
| 10board | The school board whose curriculum the candidate followed in grade 10 | India has several boards of education which follow their own course curricula. Schools are affiliated to one of these boards. More information here - https://en.wikipedia.org/wiki/Boards\_of\_Education\_in\_India |
| 12graduation | Year of graduation - senior year high school |  |
| 12percentage | Overall marks obtained in grade 12 examinations | Domain of values: [0,100] |
| 12board | The school board whose curriculum the candidate followed |  |
| CollegeID | Unique ID identifying the university/college which the candidate attended for her/his undergraduate | Multiple candidates may belong to the same university/college in this dataset. College names have not been disclosed to maintain privacy. |
| CollegeTier | Each college has been annotated as 1 or 2. The annotations have been computed from the average AMCAT scores obtained by the students in the college/university. Colleges with an average score above a threshold as tagged as 1 and others as 2. |  |
| Degree | Degree obtained/pursued by the candidate | Relevant information here -  \* https://en.wikipedia.org/wiki/Bachelor%27s\_degree#India \* https://en.wikipedia.org/wiki/Bachelor\_of\_Engineering#India \* https://en.wikipedia.org/wiki/Bachelor\_of\_Technology#India |
| Specialization | Specialization pursued by the candidate | Some information here - https://en.wikipedia.org/wiki/Bachelor\_of\_Engineering#India |
| CollegeGPA | Aggregate GPA at graduation | This is the raw information submitted by candidates. Some have submitted percentages while others have posted on a 10-point scale. Some of these GPAs might be relative while others can be absolute. |
| CollegeCityID | A unique ID to identify the city in which the college is located in. |  |
| CollegeCityTier | The tier of the city in which the college is located in. This is annotated based on the population of the cities. |  |
| CollegeState | Name of the state in which the college is located | The provided data set does not accurately capture state-wise distributions. |
| GraduationYear | Year of graduation (Bachelor's degree) |  |
| English | Scores in AMCAT English section | AMCAT or Aspiring Minds Computer Adaptive Test is an employability assessment test to assess the aptitude, reasoning and technical skills of students and new graduates. It is a statistical assessment platform to provide the skill measurements for each student who takes up the test. The assessments focus on three primary skills – cognitive (which include tests like English ability, Logical ability, Quantitative ability, Attention to detail), domain (which includes a test on the specific branch of studies taken up by the student) and personality (based on the big-5 personality traits). The scores are measured and reported on a scale of 100-900 (barring the personality assessments). For more details, see:  Link1: aspiringminds.com  Link2: research.aspiringminds.com |
| Logical | Score in AMCAT Logical ability section |  |
| Quant | Score in AMCAT's Quantitative ability section |  |
| Domain | Scores in AMCAT's domain module | Since different candidates give different domain-specific tests, we report here the percentile of the candidates in their respective tests. The scores are reported on a scale of 0-1.  This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. Aspiring Minds provides an internal map which relates a student's specialization to particular AMCAT tests. In case the score is -1 for this test, you could make use of the scores in the tests mentioned below to get more information. If the scores in the tests below are -1 as well, then it can be concluded that the candidate has not given a module listed here. |
| ComputerProgramming | Score in AMCAT's Computer programming section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| ElectronicsAndSemicon | Score in AMCAT's Electronics & Semiconductor Engineering section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| ComputerScience | Score in AMCAT's Computer Science section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| MechanicalEngg | Score in AMCAT's Mechanical Engineering section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| ElectricalEngg | Score in AMCAT's Electrical Engineering section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| TelecomEngg | Score in AMCAT's Telecommunication Engineering section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| CivilEngg | Score in AMCAT's Civil Engineering section | This is an optional section for the candidates. Those opting out of it get a score of -1. One may consider this as missing data. |
| conscientiousness | Scores in one of the sections of AMCAT's personality test | For a clarity on what the traits mean on which the scores are reported, see https://en.wikipedia.org/wiki/Big\_Five\_personality\_traits  The scores are sampled from a distribution with mean 0 and standard deviation 1 |
| agreeableness | Scores in one of the sections of AMCAT's personality test |
| extraversion | Scores in one of the sections of AMCAT's personality test |
| nueroticism | Scores in one of the sections of AMCAT's personality test |
| openess\_to\_experience | Scores in one of the sections of AMCAT's personality test |

**FUTURE WORK: -**

1. Data Collection
2. Basic Data Exploration
3. Data Visualization
4. Treating Anomalies (like missing values, outliers, bogus data, etc)
5. Evaluating Feature Importance
6. Building Base Models
7. Tuning hyper-parameters / improving model performance/reducing over fit / under fit (if any)
8. Evaluating Performance based on various performance metrics
9. Final Evaluation of the Tuned Models using K-Fold cross validation strategies
10. Final Model Building
11. Drawing Inferences and Providing Recommendations.