Report of Internship undertaken at Vedantu Innovations Pvt. Ltd.

NILESH KHATRI

(201712099)

Supervisor

Prof. Manish Khare

On-site Supervisor

Mr. Pranav Mallar

Vedantu Innovations Pvt. Ltd.

Bangalore

Introduction

- Vedantu is India's leading Online tutoring company which enables students to learn LIVE with some of India's best-curated teachers. Vedantu`s USP is its quality of teachers. It has some 500+ teachers who have taught more than 1 Million hours to 40,000+ students spread across 1000+ cities from 30+ countries. Vedantu is founded by IITian friends who have been teachers themselves with over 13 years of teaching experience and having taught over 10,000 students.

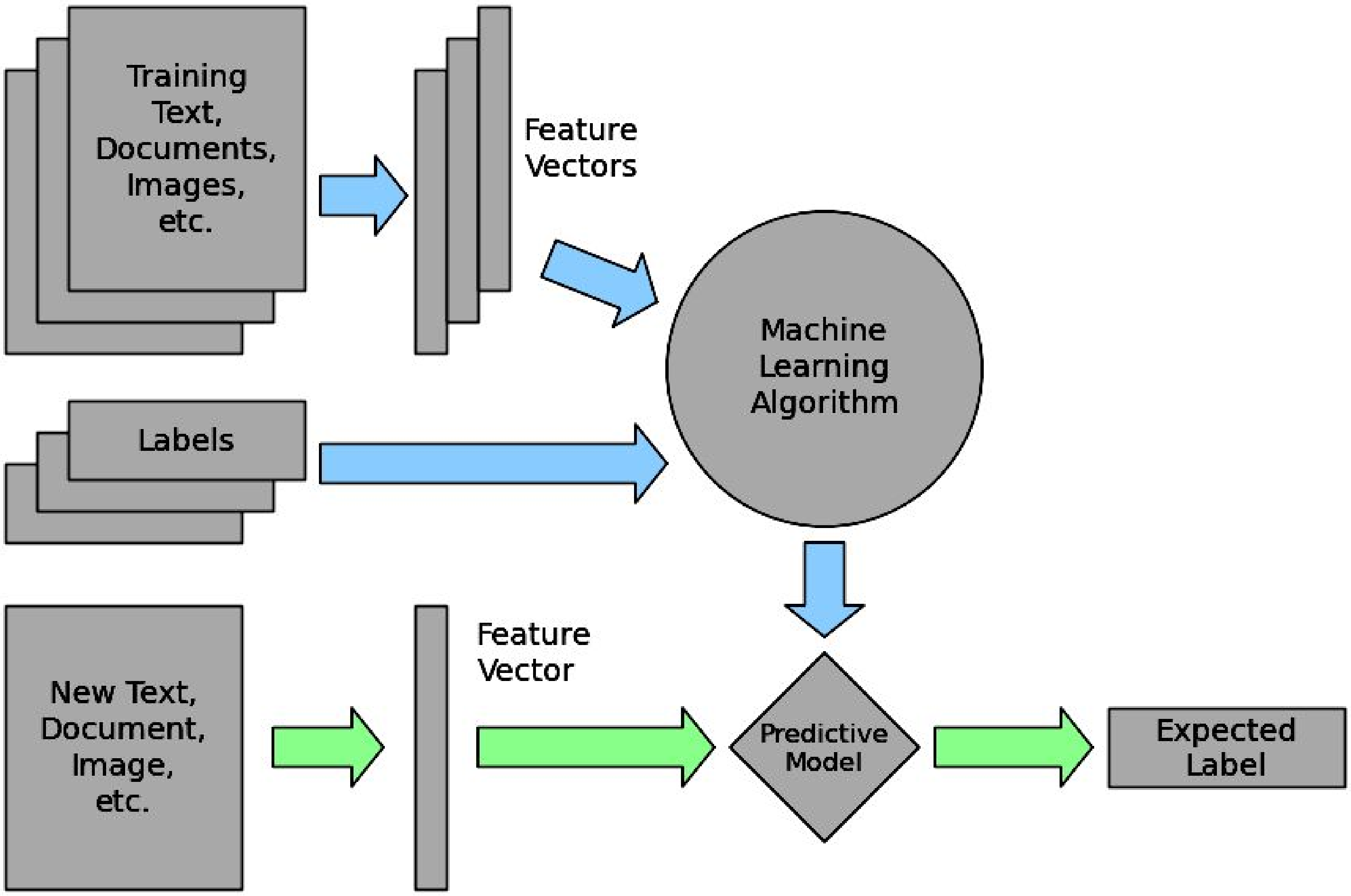
- Vedantu's online tutoring platform enables LIVE interactive learning between a teacher and a student. It offers individual and group classes. On Vedantu a teacher is able to give personalized teaching using two-way audio, video and whiteboarding tools where both teacher and student are able to see, hear, write and interact in real-time. Imagine it like 'Skype' custom made for education.

Scope

- As a part of the Internship, I was in the Data Science Team working with Database and Data Analysis Techniques to improve the quality and services that Vedantu provides in Vedantu's online tutoring platform. Unearthing the very meaning of the efficiencies and inefficiencies, good and bad, useful and useless, profit or losses are dependents on unraveling these combinations that have surrounded us. And drawing insights from this data is known as *data analytics*.

- As a part of the Data Science Team, it is very important to Analyse the Data and help Vedantu to improve the quality and Standards to continuously grow and achieve numerous Milestones in very less time.

Context Diagram


Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Detailed Use Cases that have been implemented

- Clickstream analysis :

  Clickstream data is the trail of digital breadcrumbs left by users as they click their way through a website, and it's loaded with valuable customer information for businesses.

- Sentiment analysis :

  Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. When a company releases a new product, monitoring and analyzing social media content can play a large role in quickly remediating bugs and errors.

- Predictive analytics :

  Predictive analytics is the branch of the advanced analytics which is used to make predictions about unknown future events. Predictive analytics uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about the future.

- Technical Issue Segmentation :

  During Online Classes many students may face some of the technical issues which makes a bad impact on the Platform and those issues need to be solved immediately to help students take advantage of the platform without any issues, for which I was supposed to make Technical issues detection model using machine learning.

- Data Cleaning Techniques :

  Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognizing unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodeling, or removing the dirty or crude data.

Programming Contribution

- Clickstream:
  - Get details about the Users Activities such as how many times User visits particular study folder, subjects, chapter as well as its latest count and time details.
  - Get details about the activities of the downloads of the User, Items shared by users in particular time periods based on some conditions.
  - Get details of the User's visit to a particular playlist of the tutorials and it's counts based on some conditions.

- Data Analysis using Python(Pandas and NumPy):
  - Get session details and average time spent by a user based on a particular session.
  - Get Count of Users region-wise, group user activities and get a count of activities sorted by latest activities.

- Machine Learning Model for Detecting Technical Issues :
  - During Live sessions on Vedantu Platform, some students may face Technical Issues which becomes a barrier in Live Classes and student may miss important lectures, so some of these technical issues are reported by the students in the chat area, so I was supposed to make a machine learning model to detect these Technical issues from this Chat data.

Tools, Technologies, APIs, and Libraries used.

- PyCharm
  - PyCharm is an integrated development environment used in computer programming, specifically for the Python language.

- MongoDB
  - MongoDB is a cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemata.

- Amazon Redshift
  - Amazon Redshift is an Internet hosting service and data warehouse product which forms part of the larger cloud-computing platform Amazon Web Services. It is built on top of technology from the massive parallel processing data warehouse company ParAccel, to handle large scale data sets and database migrations.

- NumPy
  - NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays

- Pandas
  - *Pandas* is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

- Scikit-learn
  - **Scikit-learn** is a machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, *k*-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

- Natural Language Toolkit (NLTK)
  - The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language.

Testing Strategies and Reports

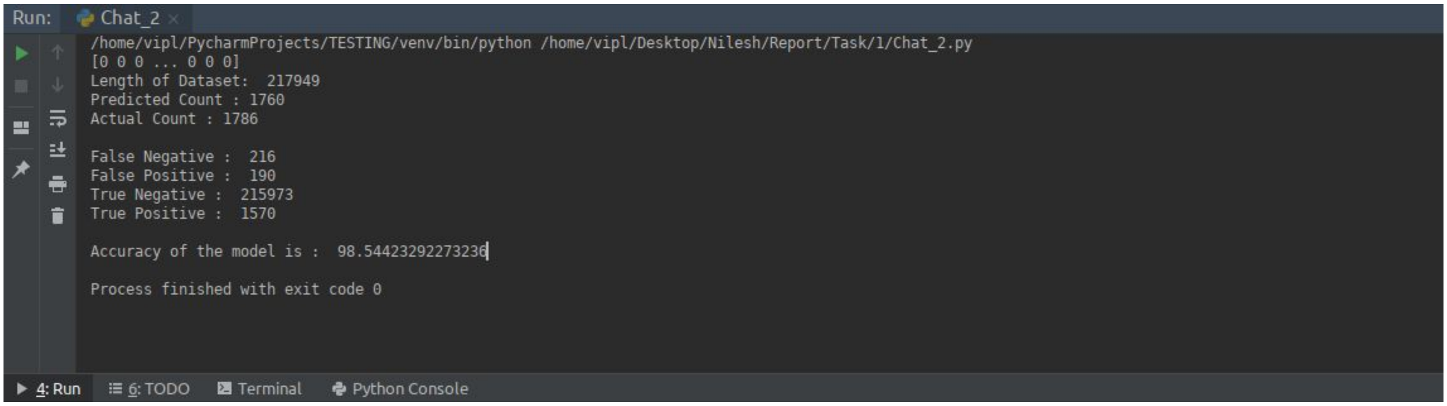
- Confusion Matrix:
    - A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm.
    - A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix.
    - In confusion matrix, our model is basically classified in four different parts which are as follows:

False negative: When the test says you don’t have it but you actually have it.

False positive: When the test says you have it but you actually don’t have it.

True negative: When the test says you don’t have it and you actually don’t have it.

True positive: When the test says you have it and you actually have it.



```
/home/vipl/PycharmProjects/TESTING/venv/bin/python /home/vipl/Desktop/Nilesh/Report/Task/1/Chat_2.py
[0 0 0 ... 0 0 0]
Length of Dataset:  217949
Predicted Count : 1760
Actual Count : 1786

False Negative :  216
False Positive :  190
True Negative :  215973
True Positive :  1570

Accuracy of the model is :  98.54423292273236

Process finished with exit code 0
```

Lessons Learnt

- I learnt New Database Amazon Redshift as well as MongoDB.
- How to work with Big Data as well as Optimizing Large queries to reduce the computational time and get results as fast as possible.
- I learnt new Python Libraries like Pandas, NumPy, Scikit-Learn, Natural Language Toolkit(NLTK), etc.
- How to work with huge Datasets and getting the desired outcomes as per the requirement from the huge Datasets.
- How to develop a Machine Learning Model.
- How to improve the accuracy of the Machine Learning Model to get accurate results.
- How to apply the knowledge practically which you learned while studying.
- How to work effectively in a professional environment.