Q1) Predictive Modelling Given training data on tweets and their retweets, how would you predict the number of retweets of a given tweet after 7 days after only observing 2 days worth of data?

Ans: Retweets can depend on the main key features of a tweet that need to be considered while analysis:

|  |  |
| --- | --- |
| Tweet metrics | Description |
| URLs count | # of URLs in the tweet |
| Mentions count | # of mentions/citation of Twitter users in the tweet |
| Hashtags count | # of hashtags included in the tweet |
| Favorites count | # of favorite obtained by the tweet |
| Publication time | Local hour H24 in which the tweet has been published in the day according to the author’ local time. |
| Author of tweet metrics | Description |
| Days count | # of days since the tweet’s author created its Twitter account |
| Statuses count | # of tweets made by the tweet’s author since the creation of its own account |
| Author network metrics | Description |
| Followers count | # of followers the author of the tweet |
| Followees count | # of friends the tweet’s author is following |
| Listed count | # of people added the tweet’s author to a list |
|  |  |

More points should be considered like is the person a politician or celebrity and what is the motive of tweet, is it a burning issue ( Eg. Metoo campaign ), what is the age of account ?

On this basis the problem can be approached towards the solution. We can check is all the features are influencing the retweets or if not then what are those features for that dimension reduction can be used.

The problem can be treated as regression where number of retweets can be count on the basis of what data it is the type of the regression can be decided.

Or the problem can also be treated as classification problem where the classes can be divided 1-10, 10-20… according to the dataset.so based on proper exploratory data analysis the problem can be solved by standard ML multiclass classification techniques.

Q2) Predictive Modelling How could you collect and analyse data to use social media to predict the weather?

Ans: Data can be taken from the Twitter, Facebook, Flickr et al public APIs filtered by location.

* Define and label the weather in the past x days for each feed.
* For each feed, find the words or phrase using with text matrix. There may be some featured terms, like ‘umbrella.… This could be a very sparse table, but feature reduction can be performed based on the weather information (only choose the highly frequent ones for a particular weather).
* Other possible attributes could be date (month, quarter), location, etc.
* Extract the real time new feeds.
* Extract the text and other attributes from the feeds, and then data manipulation.
* Perform suitable classification technique with the historical data for a general weather prediction.

Q3) Predictive Modelling How would you construct a feed to show relevant content for a site that involves user interactions with items?

Ans: This is a classical problem of a recommendation system where the sites like amazon, flipcart or Netflix ( involves user interactions ) recommends the items in a customised way for every particular user.

Different sites uses different approaches for the recommendation.

Example:

**Netflix**: This is one of the simplest approach where, they save all the information related to each user in a vector form. This vector contains the past behavior of the user, i.e. the movies liked/disliked by the user and the ratings given by them. This vector is known as the profile vector. All the information related to movies is stored in another vector called the item vector. Item vector contains the details of each movie, like genre, cast, director, etc.

The content-based filtering algorithm finds the cosine of the angle between the profile vector and item vector, i.e. cosine similarity. Suppose A is the profile vector and B is the item vector, then the similarity between them can be calculated, which ranges between -1 to 1, the movies are arranged in descending order.

Q4 ) Predictive Modelling How would you design the people you may know feature on LinkedIn or Facebook?

Ans :Facebook or linkedin maintain the information in a graph structure where each person is a vertex and  edge between two people will simply tell whether or not two persons are friends. Weight in that edge may correspond to something similar between two people, e.g. same location, hometown, likes, mutual friends etc.

It suggest the friends based on the number of mutual friends of 2 people, or a person saw your profile that would be a suggested friend or facebook takes the from your whatsapp contacts and from that data facebook suggest that this person you may know.

Or It can be designed as For a vertex ***v*** find vertices in the graph or sub-graph, which are one or two hop away and have maximum or higher weights on their edge from others.

Q5) Predictive Modelling How would you predict who someone may want to send Snapchat or Gmail to?

Ans: Things can be considered while predicting the solution:

* for each user, assign a score of how likely someone would send an email to
* the rest is feature engineering:
  + number of past emails, how many responses, the last time they exchanged an email, whether the last email ends with a question mark, features about the other users, etc.
* People who someone sent emails the most in the past, conditioning on time decay.

Q6 ) Predictive Modelling How would you suggest to a franchise where to open a new store?

Ans :

**first step is data acquisition**:

1. build a dataset with local demographic information available for each location.
2. local income levels, proximity to traffic, weather, population density, proximity to other businesses
3. a reference dataset on local, regional, and national macroeconomic conditions (e.g. unemployment, inflation, prime interest rate, etc.)
4. any data on the local franchise owner-operators, to the degree the manager
5. Is any other outlet of any other franchise which provide the similar product is available in that area.

These are the key features that needs to be focus while acquiring the dataset.

Then based on this features the data of the other outlets of same franchise can be merged so that it can be predicted that if in **x** conditions the revenue of a outlet is **y**

So in following conditions what can be the revenue of new outlet if it is opened in some area.

Q7) Predictive Modelling In a search engine, given partial data on what the user has typed, how would you predict the user’s eventual search query?

Ans: first we need huge amount of text ( search query ) data like Google have. Then based on that data of search queries and  on the past frequencies of words shown up given a sequence of words, we can construct conditional probabilities of the set of next sequences of words that can show up (n-gram). The sequences with highest conditional probabilities can show up as top candidates. we can put more weight on past sequences which showed up more recently and near your location to account for trends.

Q8) Predictive Modelling Given a database of all previous alumni donations to your university, how would you predict which recent alumni are most likely to donate?

Ans: This is a binary classification problem where the class would be donate or not donate(0 or 1) First it needs to be checked what is the information or features available in data, suppose the passing year of alumni, the income of alumni, number of times donation, amount of donation, age of alumni, current working company. After the data understanding we will see what features are more important by finding the variance of the features and if not all the features are important we can remove those features then by plotting the data we can see is data linearly separable or if not then can apply some feature engineering then can apply suitable classification technique according to data.

Q9 ) Predictive Modelling You’re Uber and you want to design a heatmap to recommend to drivers where to wait for a passenger. How would you approach this?

Ans: Uber has a massive database of drivers and passengers, it leverages that database and all real time data for more business revenue. Uber uses its the data to predict where the demand can be high so that drivers can wait there to get the more rides, prediction could be based on key features like peak hours, particular time of the day, any important occasion.

**Before getting the ride ( predicting high demand area )**

We can develop a predictive model based on the previous data such that if these are the conditions ( time of the day , day of the week, any special occasion , traffic of the route etc.) given we can find the probability of having high demand on particular regions.

**After getting the ride (where he should wait )**

we can collect all previous journeys from the similar area (can be done by clustering).

Then you can use the metric which analyses how long it took for a driver to find the client once they arrived to the pick-up location. The locations with smaller ‘search for a customer’ times will be the most appropriate.

Q10) Predictive Modelling How would you build a model to predict a March Mad- ness bracket?

Ans : March Madness pools are a form of sports betting based on the annual NCAA Division I Men's Basketball Tournament each spring in the United States, where 64 college teams take participation in the tournament.

**Data selection** Two easily accessible sets of predictors for NCAA basketball tournament outcomes are information from prior tournaments and results from regular season competition. Regular season data would generally include information like each game’s home team, away team, location, and the final score. For tournament games, additional information would include each team’s seed (No. 1 to No. 16), region, and the distance from each school’s campus to the game location.

To predict the outcomes of the matches a probabilistic model can be design based on the previous data available like team A is more likely to win against team B but have lost more times against team C in past matches. Which are the players playing in the teams there scoreboard and the rankings of the players. Which teams are more likely to win at which place on this data analysis can be done to predict the results of matches.