Business Understanding

Identifying the business goals:

1. Background: Instagram currently has about one billion monthly active users, meaning that there are a lot of people a company could choose from to advertise their product to their audience. This project will be useful in assisting companies to determine which Instagram influencers will be the most successful at reaching mass audiences in order to sell their product(s). This project could also benefit Instagram users that are looking to get more reactions on their Instagram posts.
2. Business goals: With this project I intend to help companies better understand what the most followed people on Instagram have in common that help them to generate a large number of likes. The company will then be able to use this data to help them determine what features a person must possess when promoting their product so that their post generates the greatest number of likes. The greater the amount of likes a post receives will increase the amount of people that see the post, therefore increasing the number of people that will see the product which can increase the sales of the product. Additionally, if the project is being used to benefit a single person, the goal would then be to help that person get more likes than they are currently receiving.
3. Business success criteria: To determine if the project is successful at increasing product sales by promoting the product on Instagram using people possessing the optimal features determined by the project. I would look at the product sales a month before the Instagram post and a month after the Instagram post. If the product receives more views and more sales after the promotion was posted on Instagram, I would consider the project a success.

Assessing your situation:

1. Inventory of resources: Public Instagram accounts and the photos on these pages.
2. Requirements, assumptions, and constraints: The deadline for this project is December 16, 2019. Instagram is a free to use and free to access social media site, so all the information in my datasets will be from public pages that anyone can access. I am not using any personal information only features that describe a picture, so no legal or security issues should arise.
3. Risks and contingencies: Internet outages are always possible so if one occurs, I will simply go to the university's library. Another risk that is possible is someone finding out their data was used and wanting their data out of the training model. My solution would be to remove their data and replace it with a new person's data.
4. Terminology:
    a. Followers – How many Instagram accounts follow a certain user.
    b. Likes – How many people have interacted with a post.
    c. Features – Certain characteristics a person displays in their Instagram photo e.g., smiling with teeth, no facial hair, or eye shadow color.
    d. Reactions – the act of liking, commenting, or sharing a post. Or receiving likes, comments, or shares on a post.

5. Costs and benefits: There is no money being spent on this project.

Defining my data-mining goals:

1. Data-mining goals: The models that this project will use will be k-nearest neighbors, decision trees, and random forest.
2. Data-mining success criteria: I will use model accuracy and compare the accuracy of all three models to determine the best model to use to predict the number of likes promotional posts will receive.

Data Understanding

Gathering data:

1. Outline data requirements: The data will be in a .csv file so I can easily access the data in python using the pandas library. The data will take at least 15 hours to collect and put into the .csv file.
2. Verify data availability: All the data I will be using will be from a public Instagram pages that anyone can access. If someone is able to figure out that their data was used for this project and request that their data be removed, I would then gather more data and add it to the data set.
3. Define selection criteria: I am creating my own data source with features a person possess in a photo and the number of followers they have and the amount of likes the picture has.

Describing data:

- This data set will include information about the person in an Instagram post. It will include the top 100 most followed people on Instagram and 10 of their photo's features. Some of these features will include whether or not they are showing teeth when they are smiling, what color their hair is, color of their eyes, length of hair, facial hair, gender, etc.

Exploring data:

- First thing I will have to do to the data is making sure that no mistakes were made when creating the data set, like making sure that all things are spelled the same way and making sure all the data makes sense for each feature. I hypothesis that women with long blonde hair will generate more likes, on average, than any other women with another hair color and length because there are more blonde stereotypes. For example, the blonde bombshell.

Verifying data quality:

- A quality issue I may face with this project is having male and females in the same data set together. This could cause an issue if it finds that women get higher likes if they have a mustache So I may need to split the data into two different data sets and gather more information for both the male and female datasets.

Planning the Project

Plan for the project:

1. Creating the dataset: To create the dataset I will first find the top 100 most followed Instagram accounts. I will then start going through the list to find photos they have posted that clearly show as many features as possible. It should take about 15 or more hours.
2. Cleaning the data: After collecting the data I will need to make sure that the data is correct and not missing any features. This should take about 2 hours.
3. Creating the code: The next step will be to write the code that will read in the data and prepare the data for the machine learning algorithms. This should take about 1 hour.
4. Predicting Instagram likes: The next step is to predict the number of likes using each machine learning algorithm. It should take about 2 hours.
5. Comparing the models: This step will calculate the accuracy of the models and compare them to one another and determine which model gave the higher accuracy. This should take about 5 hours.
6. Making the program user friendly: After the base code is running, I want to make the code user friendly so that a user will be able to enter some features about a picture and the code will tell them the number of likes the person will receive. This will take about 20 hours.

Methods and tools:

- I plan to use only python for this project. To make the data collection faster I may code a Instagram scrapper to collect the data for me.