Course Project

Course Project

The web sites https://www.sas.com/en_us/customers.html, https://www.ibm.com/case-studies/search?search, and https://www.informs.org/Impact/O.R.-Analytics-Success-Stories (among others) contain brief overviews of some major Analytics success stories. In this course project, your job is to think carefully about what analytics models and data might have been required.

(1) Browse the short overviews of the projects. Read a bunch of them – they're really interesting. But don't try to read them all unless you have a lot of spare time; there are lots!

(2) Pick a project for which you think at least three different Analytics models might have been combined to create the solution.

(3) Think carefully and critically about what models might be used to create the solution, how they would be combined, what specific data might be needed to use the models, how it might be collected, and how often it might need to be refreshed and the models re-run. DO NOT find a description online (or elsewhere) of what the company or organization actually did. I want this project to be about your ideas, not about reading what someone else did.

(4) Write a short report describing your answers to (3).

Solution - Course Project

1. Case Study - Introduction

I've picked up a Case study from the Agriculture domain featured on SAS website (https://www.sas.com/en_us/customers/boragen.html).

The Biotechnology company *Boragen* is into creating next gen crop-protection solutions.

The case is centered around the below concept taken from the webpage mentioned above:

To ensure continued bountiful production of grains, fruits and vegetables, new tools are needed to augment sustainable agricultural practices. The most effective tools often merge ancient farmer know-how and natural products with new data-driven approaches.

In the below sections I'll be analyzing the various challenges faced by the company - *Boragen* and what analytics approaches it will have to make use in order to overcome those.

2. Challenges

- The company wants to create new generation of fungicides (Chemical compounds) that can be used safely and in a more sustainable way.
- In Agriculture, introducing a new compound from the lab to the field is time-consuming and expensive process, involving several phases of testing.
- Better data analysis can help uncover innovations that escape the human eye or ancient wisdom.
- Need for tools and techniques to ensure that right mix of compounds are used in the experiments and field testing.
- Need to compare the efficiency of the existing products in the market with new compounds that *Boragen* is going to launch.
- Ensuring right data is available at every stage of research and development process.

3. Data Requirements

Assuming that we will be requiring the below mentioned datasets.

Here are the datasets needed to get closer to solutions for the challenges mentioned in section - 2 :

- Historical geographical and climatic conditions where the new generation of the fungicide need to be applied.
  - Moisture levels
  - Rainfall
  - Temperature ranges
  - Soil conditions
  - Altitude
  - Common crop diseases
- Existing solutions in the market and their efficiency in the similar conditions over the period of time.
  - Volume of the compound applied.
  - Healthy produce

  - Efficiency rate
  - Supported Crop Types

- Selection of new experimental chemicals / compounds.
  - Volume of the compound applied.
  - Healthy produce
  - Projected Efficiency rate
  - Supported Crop Types

- Crop details where the compound would have to be applied.
  - Crop type
  - Nutrient needs (NPK)
  - Suitable climatic and geographic conditions
  - Compounds applied
  - Crop yield
  - Crop shelf life
  - Geographical location

These datasets have to be obtained from internal and external data sources like:

- Crop yield data - Government or independent survey companies datasets
- Geographical data - Government or independent survey companies datasets
- Weather data - 3rd Party companies - Weather Channel / DTN
- Existing solutions and effects - Department of Agriculture and other Ag Consortiums
- New solutions and effects - Internal Research and Development

In real world this data acquisition is a large amount of effort. Company would need specialized people within the domain to ensure right data is procured and used.

When newer versions of these datasets are released by the sources then the whole process would have to restart.

4. Data Exploration and Analysis

As a starting point the Data scientists / analysts in the company would have to begin with some preliminary steps like below:

- Draw bar and line plot to analyze the data graphically.
- Capture if there're are correlations among factors.
- Draw boxplots to identify any outliers and drop them.
- Explore the Schema of the dataset(s).

Some data cleaning techniques would also have to be applied such as *Imputation* and *Scaling*.

Imputation

```
Given : That there could be values missing within the datasets \

Use : Imputation techniques :
      - Dropping the missing data points
      - Replacing missing data with Mean / Mode values
      - Regresson predicted values
      - Regression with Perturbation.

To : Correct / Replace / Compensate for any missing values within data
```

Resulting in data that does not have any missing values that can disrupt the models. This would be used as input in the next step.

Scaling

```
Given : Some historical data points for weather / climate might be in mix of
        years or months or days or hours
        Some chemical volumes would be in different units of measure
        Some crop yield data would be in different units of measure

Use : Scaling of data

To : Ensure that Models use the correct input in order to give expected and correct output
```

After the data have been scaled to the same levels and missing values have been taken care the data becomes ready for input to any Analytics Model to derive the desired results.

5. Data Clustersing

We have to choose between using Clustering vs Classification to achieve what we want, however in this case I would choose Clustering as we are looking to find innovative solutions.

Since the company is looking for more innovative chemical compositions that can treat crop diseases in a more sustainable form.

Unsupervised learning is more suited so that it can reduce any historical biases in data and knowledge, also unraveling new findings.

I would use unsupervised clustering algo. - K Means clustering.

Climatic conditions clusters

I would classify the climatic regions for creating separate chemical formulas of the next gen application solutions.

```
Given : There's data available for the climatic conditions and diseases
        that could happen in those.

Use : K-Means with elbow method for optimal number of clusters

To : Create clusters of matching climatic conditions
```

This will result in climatic clusters that would be used to treat crops with specific chemical composition.

Crops clusters

Now I would create clusters of the crops so that crops with similar growing and treatments conditions are grouped together.

```
Given : There's data available for the Crops and their growing and their treatment conditions.

Use : K-Means with elbow method for optimal number of clusters

To : Create crop clusters of matching growing and treatment conditions
```

This will result in crop clusters that have been / would be used to treat crops with specific chemical composition.

Once we have clusters of Crops and Climatic conditions in place we can design experiments that can be used for each cluster combinations.

6. Design of Experiment

Based on the clusters of data in section - 3. I would create different compositions of the "Next Gen" fungicides and design experiments that uses *Factorial Design of Experiments*.

Factors to be used :

- Number of New chemical compositions
- Number of Existing chemical compositions
- Number of Climate cluster
- Number of Crop cluster

```
Given : We have clusters of data with different properties and classes of data

Use : Factorial Design of experiments

To : Create combinations of "Compound : Crop : Climate" and eliminate the non-feasible options
```

And since testing each of these in an actual field / lab would be extremely time consuming, we can run simulations and eliminate the non-feasible options from the choices of experiments.

We can use a given filtering criteria to select a subset of experiments to simulate.

Now we have subset of experiments that we would want to use for simulations.

7. Simulations

Now I'd run simulation of the feasible design experiments and start tweaking and comparing their results.

```
Given : That there are multiple chemical compositions for the same Climate and Crop
        Clusters

Use : Simulation techniques or software

To : Check which newly created compounds works best for a combination of climate and crop.
```

Towards the end of the simulation I would have efficiency score of the "new chemical compound" with different cluster combinations.

Now also run a similar simulation with the existing chemical compounds as well.

```
Given : That there are multiple chemical compositions for the same Climate and Crop
        Clusters

Use : Simulation techniques or software

To : Check which existing chemical compositions works best for a combination of climate and crop.
```

Towards the end of the simulation I would have efficiency score of the "existing chemical compound" with different cluster combinations.

In case of existing compound the efficiency is already known, and the simulations should report close values to the actual known values.

8. Validate the simulated results with known results

With the actual and simulated results in place we can create Regression Models that will check accuracy of simulated efficiency and also can make predictions easier for future compounds as well.

We have to choose between the following regression types Linear regression, Elastic net regression, Ridge regression, Lasso regression, Polynomial Regression etc. depending upon the datasets.

We can also include the Principal component analysis to reduce the # of factors in regression:

Reduce factors to be used in decision making

```
Given : The simulation results from section 7

Use : Principal component analysis (PCA)

To : Reduce the number of factors to be used in Regression model
```

We can create regression models using the datasets to compare the simulated results with known actual results:

```
Given : Dataset DS1, DS2 and Actual results capture from Lab experiment

Use : Regression

To : Check accuracy of the simulated efficiency scores with actual lab results
```

If the Mean Squared Error (MSE) is too high for outcome of simulations and lab results then the simulations need to be re-looked and re-designed. The Model also will have to be updated.

If the MSE still continues to be high then this has to be reported to the company's R&D lab to validate.

We can also create **QQ-Norm plots** to compare the simulated vs reported efficiency of the compounds in a Graphical view. /

9. What we have so far

At this point of time I would have the following set of models:

- *Models*

  - Clustering Model for Climatic conditions
  - Clustering Model for Crop Growing and Treatment conditions
  - Factorial design of experiment
  - Simulation for calculating efficiency of - New compounds
  - Simulation for calculating efficiency of - Existing compounds
  - Regression Model(s) to validate the results of the simulations

- *Dataset*
  - **DS1** Existing compounds efficiency with crops and climate conditions
  - **DS2** New compounds efficiency with crops and climate conditions

10. Run Live experiments to collect the actual results

At this point we have enough information to run actual lab experiments.

Based on the datasets - DS1 and DS2 run lab and controlled experiments and capture real results in terms of chemical compositions and their efficiency on crops.

In the lab tests there would be some constraints that could be reported like - cost, availability of chemicals etc.

Based on the Lab results if we run into constraints like a certain compound can only be given in a certain climate or is more effective on a particular crop type only, we can run optimization as well:

```
Given : Actual results capture from Lab experiment highlight contraints

Use : Optimization

To : Minimize the violation of the contraint in the future experiments and datasets.
```

11. Observations

- In reality this will not a linear process, there are multiple iterations of the same to ensure the accuracy of data
- Cases like this require good domain understanding to achieve best results
- There could also be lots of back and forth between the stages

- A large amount of time goes into data acquisition and clean related activities
- Multiple Models would have to be tried and tested before we land on some final choices
- At each stage more than a few models would prevail
- Using analytics platform with proper tools and techniques would be a preferred option for the company to expedite their work.

12. Flow

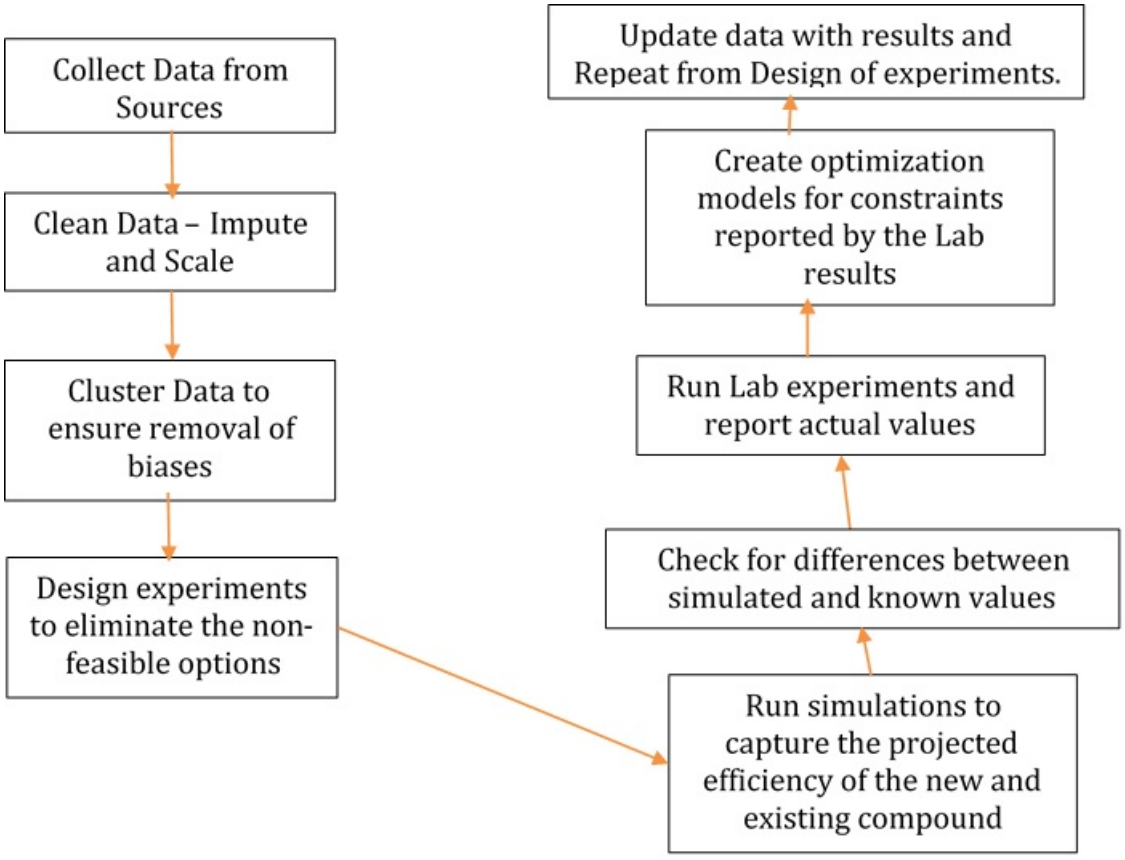

Figure 1: Flow image