Project Overview

OBJECTIVE

1. To undertake a multi-faceted project that demonstrates your understanding and mastery of the key conceptual and technological aspects of Machine Learning
2. To develop a visceral understanding of how analytics problems are solved using a combination of tools and techniques.
3. To understand the trade-offs that need to be made when solving a problem in real life.
4. To reinforce the analytical problem solving methodology.

THE 2 MILESTONES:

This project has 2 mile stones, each consisting of 3 deliverables. You get 3 weeks for completing each milestone. The 1st mile stone should be submitted in the interim report format

Mile Stone -1:

1. Define the Problem and Get the Data, Explore (Data Report)

   Split this into two. First one is define the problem. Second one define what kind of data is required, potential source, size of data, challenges faced in this step (if any) and how it was overcome. Give example of a good problem statement and the data acquisition challenge and steps (could be dummy example). Problem statement should answer What, Why and How.

   For e.g.

   "**Current Situation** - The current methodology of matching a job requirement to applications involves manual scanning of the applications to filter out applications that do not match the mandatory skills requirements for the job. This requires the concerned person to go through every application downloaded from the site and the ones on the email. This takes at least 1 minute per application which totals up too many hours given the number of applications.

   Opportunity for improvement -
   We see an opportunity here to automate the high level filtering of unwanted applications using machine learning based systems which can do the same task in milliseconds per application. This not only brings down the time to filter by a factor of 100 but also eliminates the need to involve skilled people in the mundane job.

   Data Requirement - we need job descriptions that were advertised in the last one year from all the departments and functions. Similarly, we need data in from of past applications and the job profiles. The sample applications should include all possible formats, file types for all the roles. The applications should include those that were found unsuitable for a job and those that were found suitable but did not considered and those that were suitable and considered.

Source of data and challenges - we are looking at the HR database for the applications that went through the first level filter. However, the applications that did not go beyond first level filter will be difficult to get as they may not be recorded in the system. Further, some applications may have come by post and hence not available in electronic form.

Size of the data that we need is at least 1000 data points per job description. We expect a ratio of 50:50 for reject: accept in the first level filter. Since the data can come various formats such as MS Word, PPT, Paper etc.... we need sufficient storage space and computation requirements to pre-process the data and bring it to a consistent state before modelling

2. EDA and Pre-processing (Feature engineering and selection)

   Include any insightful visualization you have teased out of the data. If you've identified particularly meaningful features, interactions or summary data, share them and explain what you noticed. Visual displays are powerful when used well, so think carefully about what information the display conveys. Some basic steps to follow:

   - Remove unwanted variables
   - Check for missing values
   - Plotting (Boxplot: outliers and scaling)
   - Splitting: Train and test validation

3. Modelling (Focus on accuracy and generalization)

   Describe what you have learned so far, what models you have used and the progress you have made towards your intended solution.

   For e.g.

   Since this is a text manipulation based project, we would like to build a NLP based Bayesian model. The input would be a labelled data of those who made through the first filter and those who did not. Both the classes with have same attributes. The accuracy will be a function of both True Positive Rate and False Positive Rates. We will evaluate the Naive Bayes, Logistic Regression and Decision Tree Classifier using ROC to select top 3 models and build an ensemble based on these to ensure the accuracy is high and the model generalizes.

Mile Stone-2:

4. Evaluation of your model (Comparison of different models, performance tuning)

   Describe how you will proceed with analysis, compare different models and choose which models to use, how will you build on your initial analysis to increase the accuracy of your solution.

   For e.g.

Since this is classification domain, we intend to use the Recall, Precision and F1 metrics along with the ROC / AUC for model comparisons.

5. Model deployment (need to cover)

   In order to start using a model for practical decision-making, it needs to be effectively deployed into production. It is one of the most difficult processes of gaining value from machine learning. It requires coordination between data scientists, IT teams, software developers, and business professionals to make sure the model works reliably in the organization's production environment. With Anaconda Enterprise, you can easily deploy:

   - Machine learning models as REST APIs
   - Dashboards with Bokeh, Plotly, and other viz libraries
   - Web applications with Flask and Tornado

   For e.g.
   We plan to deploy the model both as a service and in the batch mode. For this we will have a model object managed by a webserver such as Django based servers. The frontend will be a HTML based GUI.

6. Presentation and report

   You should start preparing the final report at least 2 weeks prior to the project completion date. No later than 1 week prior to completion date, teams should send a draft to the mentor. The format and expectations for the final report will be included in your Capstone course page.