upGrad

BUSINESS PROBLEM SOLVING

Once you have learnt the necessary theory and application of different ML models, it is essential to learn how you can utilise them in solving business problems. Real-world business problems are seldom well defined. It is upto the data scientist to understand and convert the open-ended business problem to a data science problem. Furthermore, it is also important to understand the pros and cons of each model for specific business scenarios. Finally, you need to make sure that the solution you're proposing is feasible and can be implemented at scale if needed.

BUSINESS UNDERSTANDING

There are two main aspects of the business understanding part

- Understanding the business problem
    - What is the business problem?
    - What are its implications?

In the beginning, you should understand what the business problem is and what its implications are. These implications might be cost-based, revenue-based, time-based and so on.

- Domain understanding
    - Build your domain knowledge
    - Significant in development of the final solution
    - Understand the product, business model, KPI, etc.
    - If possible, try to understand customer behaviour

Domain knowledge can be gained through a detailed discussion with the clients. Here you will understand

- What product is the client selling?
- What is their business model?
- What are the key Resources and processes involved?
- What are the KPIs involved?
- Who are the competitors? How are they solving the same business problem?

..and so on.

DEVELOP HYPOTHESES

Once you have gone through the business problem and have a clear understanding of the processes involved, the next step is to develop hypotheses regarding what might be the root cause of the problem and the possible solutions as well. This process will be further augmented once you have collected the data necessary for solving the problem.

Developing hypotheses will be a key part of your job role as a data scientist when you're working on real-world problems. You need to bring all your domain knowledge to the forefront and try to identify the potential root causes of the given problem. Also, it is here where you can discuss hypothesised solutions as well.

An understanding of the magnitude of the problem/ potential impact of the solution, if possible, can greatly help prioritize. You're often in situations where there are multiple problems that need to be solved but have limited resources. You could be leading an analytics team and need to allocate resources on projects. A sense of the potential impact/benefit can be very helpful.

DATA COLLECTION

Once you have understood the business processes, you need to take a look at the available data. Using the domain knowledge, identify the critical variables and check if the data is in a structured format (tabular) or an unstructured format(text, images, audio, video, etc.). Here, you need to identify the data that you can finally leverage for your analytics model as well.

The data may also be present in multiple sources and in different formats. You will need to identify which data sources are important and will be useful in formulating the final solution. Also, evaluate the efficacy of extracting the data from the given sources.

Next, you need to identify the data which may be needed to prove or disprove your hypothesis. Some of the hypotheses made earlier may be rejected outright due to the lack of available data. For the rest, modify your hypotheses if necessary in order to accommodate the data.

upGrad

PROBLEM MAPPING

Once you have the domain knowledge and an understanding of the available data, it is now time to convert the business problem to a data science problem and develop a solution. The approach should be in tune with the hypotheses formulated and should give an idea of your approach of the overall solution i.e. the data that you'll be using, the EDA that you'll perform, the ML algorithm that you'll be using, the Evaluation metrics you'll be tracking, mapping those evaluation metrics to KPIs and making business decisions.

This solution also needs to pay attention to the following factors:

- Cost
- Resources
- Availability of data
- Importance of decisions
- Frequency of decisions

In many scenarios, you don't necessarily need a machine learning solution for the given business problem. Here, an EDA analysis of the data is more than sufficient for solving business needs. In some cases, the data may not be enough for a machine learning model to be used. Discuss with the businesses regarding the same as to what is feasible in such cases.

If an ML solution is required, identify the ML algorithms that can help solve the problem. For example, a classification problem can be solved using either logistic regression, SVM or Random Forest models. It's up to your knowledge of the given business problem and the amount of clarity or interpretability demanded by the business to solve the problem using a particular ML algorithm.

Also, don't let the inclination towards novelty, if any, drive these decisions. There could be situations where a simple analysis/ model can solve the problem at hand. In such a situation, that simple analysis/model should be the approach that needs to be employed

Here are some additional practical constraints

- Frequency of solution
- Model lifecycle
- Offline vs online model
- Batch processing vs real-time processing
- Infrastructure constraints

upGrad

SOLUTION APPROACH

Once the business problem has been mapped to the data science problem, there are a few steps that need to be considered while developing the solution approach.
The first important step is to build a simple POC or proof of concept model. This will get your client excited about the prospects of your solution and also help define the success metrics that you'll be utilising further down the line. It also helps in identifying additional data needs if you have any.

Then you can evaluate its performance using offline validation methods and A/B testing as well.

EXPLORATORY DATA ANALYSIS

Once the solution approach is finalised, it's time for the EDA steps to begin. Here's a brief summary of EDA steps that can be used

- **Data staging and clean up**: This is the basic data cleanup and preparation stage. You collect the data from various sources, clean it and prepare the master dataset.
- **Sanity checks**: The next step is doing a quick sanity check of the entire dataset to observe any unusual data points that should not exist.
- **Univariate Analysis**: Finally we begin with the univariate analysis part. This is where visualisation tools like histograms and boxplots come in handy as they help in analysing numerical features.
- **Bivariate Analysis**: Then, you go ahead and evaluate the relationship between the target variable and the rest of the features. Here plots like scatter plots, pair plots, correlation matrices come in very handy to do the analysis. Some segmented analysis can also be done in this step as well.
- **Hypotheses validation**: In this step, you'll be getting some directional insights on whether the hypotheses that you built earlier are showing any promise or not.
- **Feature Engineering**: Finally, if you want, you can do feature engineering to extract useful features from the given dataset.

upGrad

MODEL BUILDING

Here are the steps that you need to perform during the model building phase.

- First, build a simple enough model with good interpretability so that you can demonstrate the results to your client. This will act as a proof of concept that the suggested solution approach is feasible.
- After that, increase the complexity of the model and try to optimise the parameters involved to get the best results.
- Avoid overfitting the results or else your model can't be generalised for unseen data
- Use validation set or a cross-validation score
- If possible, check the statistical significance as well.

MODEL EVALUATION

Now that you've prepared the model, it's time to evaluate the results and present it to your client. Here's a summary of the model evaluation steps that need to be performed here:

- Evaluate performance on unseen data
    - Can use validation sample
    - Can use out of time sample
- Identify the right evaluation metrics - Depending on the problem statement at hand, you should be tracking the correct evaluation metrics
- Evaluation metrics should align with business outcomes
- Model performance summary for all stakeholders - You need to make sure that you discuss a high-level summary of the entire model's performance with all the stakeholders.

Showcase the potential results - how much tangible benefits can be achieved by utilising the recommendations from the machine learning solution. Also, discuss the ways in which the model can be recalibrated in the future for even better decisions.