Predictive Analysis Modeling and Procedure

Predictive modeling is a mathematical process used to predict future events or outcomes by analyzing patterns in a given set of input data. It is a crucial component of predictive analytics, a type of data analytics which uses current and historical data to forecast activity, behavior and trends.

Examples of predictive modeling include estimating the quality of a sales lead, the likelihood of spam or the probability someone will click a link or buy a product. These capabilities are often baked into various business applications, so it is worth understanding the mechanics of predictive modeling to troubleshoot and improve performance.

Although predictive modeling implies a focus on forecasting the future, it can also predict outcomes (e.g., the probability a transaction is fraudulent). In this case, the event has already happened (fraud committed). The goal here is to predict whether future analysis will find the transaction is fraudulent. Predictive modeling can also forecast future requirements or facilitate what-if analysis.

"Predictive modeling is a form of data mining that analyzes historical data with the goal of identifying trends or patterns and then using those insights to predict future outcomes," explained Donncha Carroll a partner in the revenue growth practice of Axiom Consulting Partners. "Essentially, it asks the question, 'have I seen this before' followed by, 'what typically comes after this pattern.'"

Types of predictive models

There are many ways of classifying predictive models and in practice multiple types of models may be combined for best results. The most salient distinction is between unsupervised versus supervised models.

- Unsupervised models use traditional statistics to classify the data directly, using techniques like logistic regression, time series analysis and decision trees.

- Supervised models use newer machine learning techniques such as neural networks to identify patterns buried in data that has already been labeled.

The biggest difference between these approaches is that with supervised models more care must be taken to properly label data sets upfront.

"The application of different types of models tends to be more domain-specific than industry-specific," said Scott Buchholz, government and public services CTO and emerging technology research director at Deloitte Consulting.

In certain cases, for example, standard statistical regression analysis may provide the best predictive power. In other cases, more sophisticated models are the right approach. For example, in a hospital, classic statistical techniques may be enough to identify key constraints for scheduling, but neural networks, a type of deep learning, may be required to optimize patient assignment to doctors.

Once data scientists gather this sample data, they must select the right model. Linear regressions are among the simplest types of predictive models. Linear models take two variables that are correlated -- one independent and the other dependent -- and plot one on the x-axis and one on the y-axis. The model applies a best fit line to the resulting data points. Data scientists can use this to predict future occurrences of the dependent variable.

Some of the most popular methods include the following:

- **Decision trees.** Decision tree algorithms take data (mined, open source, internal) and graph it out in branches to display the possible outcomes of various decisions. Decision trees classify response variables and predict response variables based on past decisions, can be used with incomplete data sets and are easily explainable and accessible for novice data scientists.
- **Time series analysis.** This is a technique for the prediction of events through a sequence of time. You can predict future events by analyzing past trends and extrapolating from there.

- **Logistic regression.** This method is a statistical analysis method that aids in data preparation. As more data is brought in, the algorithm's ability to sort and classify it improves and therefore predictions can be made.
- **Neural networks.** This technique reviews large volumes of labeled data in search of correlations between variables in the data. Neural networks form the basis of many of today's examples of artificial intelligence (AI), including image recognition, smart assistants and natural language generation.

Predictive Analytics: Steps or Procedure

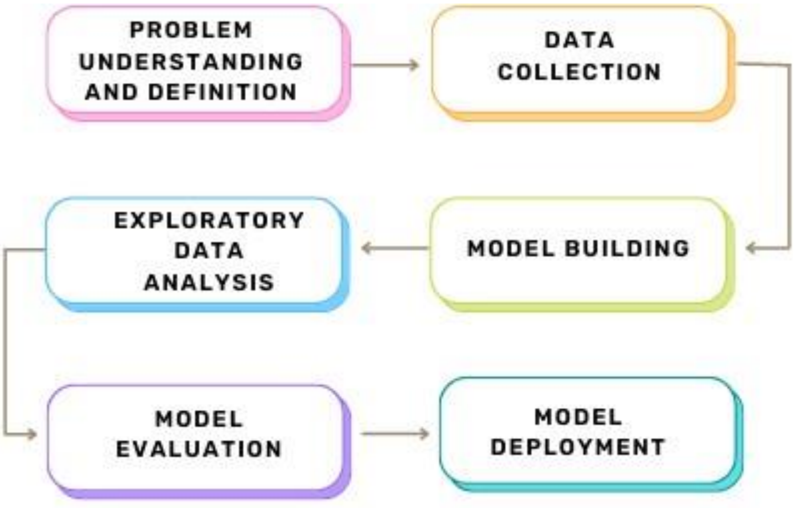
PROBLEM UNDERSTANDING AND DEFINITION
DATA COLLECTION
MODEL BUILDING
EXPLORATORY DATA ANALYSIS
MODEL EVALUATION
MODEL DEPLOYMENT

Figure 1: The Predictive Analytics Process

(1) Problem Understanding and Definition:

This is the initial stage in the process of predictive analysis. This is a vital stage because we first need to understand what exactly the problem is to frame the solution. When a stakeholder approaches you with a certain problem, the first step would be to know the stakeholders' requirements, the utilities available, the deliverables and finally, know how the solution looks from the business perspective.

Sometimes the requirements of the stakeholders may not be clearly defined. It becomes our responsibility to understand precisely what is to be predicted and whether the outcome solves the defined problem. The dynamics of the solution and the outcome completely change based on the problem definition.

Converting a business problem into an analytical one is the most important part of predictive analysis. Hence explicitly define what is to be predicted and how does the outcome look like.

(2) Data Collection:

This is the most time-consuming stage. Sometimes, the required data may be provided by the stakeholder, from an external database or in some cases, you may have to extract the data. It is possible that the data so collected may not be sufficient for framing the solution. You may have to collect data from many sources. Think about how much access you have to the dataset that is required.

Since the outcome of the predictive model relies entirely on the data used, it is important to gather the most relevant data that aligns with the problem requirements. Here are a few things to be kept in mind while searching for a dataset:

- Format of the data
- Period across which the data is collected
- The attributes of the dataset
- Does the dataset meet your requirements?

(3) Exploratory Data Analysis:

Once you have the dataset ready, you now may be willing to build your predictive model. But before we start, it is crucial to know the properties of your data. Understanding the kind of data you have, the features it possesses, the target or outcome variable, and the correlation among these features all play a role in designing a suitable model. The main aim of EDA is to understand the data. This may be achieved by answering the below few questions:

- What are the data types present in the dataset?
- What is the dimensionality of the dataset?
- What does the data distribution look like?
- Is there any missing data?
- Is there any prominent pattern in the data distribution?
- Do you observe outliers?
- How are data features correlated to each other?
- Does their correlation affect the outcome?

Sometimes the data collected contains a lot of redundant data. If such data is fed as input to the model, there is a high possibility that the model makes wrong predictions. Hence it is important to perform EDA on the data to ensure that all the outliers, null values and other unnecessary elements are identified and treated. Identifying the patterns in the data makes it easier to decide the model's parameters. EDA helps us improve the model's accuracy even before it is built.

EDA generally has two components- numerical calculations and data visualizations. Calculating Standard Deviation, Z-score, Inter-Quartile Range, Mean, Median, Mode, and identifying the skewness in the data are some ways of understanding the dispersion of data across the dataset. Graphical representations such as heat maps, scatter plots, bar graphs, and box plots help get a wider view of the dataset.

(4) Model Building:

After applying EDA, it is finally time to build predictive models using machine learning. In the dataset, we use the predictor variables to make predictions on the target variable.

Target: The dependent variable whose values are to be predicted.

Predictors: The independent attributes in the dataset that are used to predict the value of the target variable. Once the target is identified, all other columns become the predictor variables.

Here we consider the model a calculator that takes in inputs and gives out the predicted output. We may have to build a Regression or a Classification model based on the problem.

Regression algorithms such as Simple Linear Regression, Multi Linear Regression, Decision Tree Regression etc., may be used to get desired results. Such models are used when the target is a numeric feature.

Example: Predicting the house prices

While classification models are used when the target is a categorical feature, the classification problems may be a binary classification or multiclass classification.

Binary classification: The target has only two possible categories.

Multiclass classification: The target has more than two possible outcome categories.

Apart from these, unsupervised learning algorithms such as Clustering and Association algorithms can also be used based on the requirement.

(5) Model Evaluation:

Once the model is built, the next stage would be to analyze the performance of the model. Evaluating the model based on different scenarios and parameters thereby contributes to deciding ‘the most effective’ model for solving the given problem. Usually, one or more metrics are used to know how good the model performs.

Different measures are used for rating the performance of machine learning models.

For regression models: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R Squared (R2 Score)

For classification models: F2 Score, Confusion Matrix, Precision, Recall, AUC-ROC

(6) Model Deployment:

Now that the model has been built, tested and evaluated, it is time to deliver it to the stakeholder. Model deployment involves placing the model into a real-world application that can be used for its intended purpose. This may be done by using the model in a software application, integrating it into a hardware device, building a framework around the model or using the model itself as a ‘data product’.