Diabetes Challenge Saturdays.AI Guide

Saturdays.AI

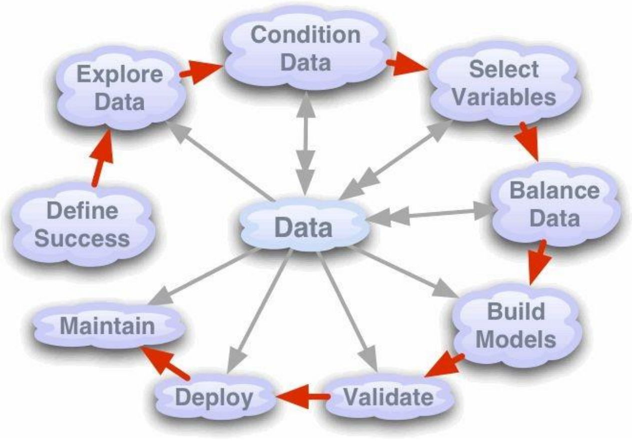
Index

Phases of creating a model

What is a predictive model?

- A predictive model is simply a toolfor predicting future events/characteristics using historical data
- Examples of day-to-day:
    - Decide the route based on traffic
    - Next word to type on mobile phone
- Building a model follows a very defined process for having a predictive model.

Predictive model lifecycle



1. **Defining the problem**

Defining a problem is the first part of building models. It is necessary to respond to the "why". For example: we assume that you could opt for the "Science of

Data" as a race option. But why? If you know the slob,great. But, if you're still thinking about an answer, it may be time to consider the "why" you're taking this workshop.

In model building, defining the problem helps us know what we are trying to solve. Ideally, this stage requires an exhaustive Q&A brainstorming session with business customers to know what exactly he/she expects from the analytics team.

When it comes toco-structuring a definition of the problem, it is expected that a framework will be put in place to the problem and then possible solutions will be found.

For example, let's say the task at hand is to increase the benefits of credit cards.

So, while calculating the profitability of credit cards, we can say that it can be increased either by increasing revenue or by lowering costs.

Income can be increased by changing several things:

- Increase/decrease the interest rate
- Changing customers' credit limits
- Increasing the rate on credit cards

Similarly, costs can be reduced by making anyof the following changes:

- Renegotiate the terms of the contract for the interest rate that the bank pays its creditors
- Reduce the cost of operations by reducing or automating customer support
- Close the accounts of people who never pay interest!

de Although there is no correct or incorrect response base in the information we have, it would probably have beenlegit to change the interest rate for customers. This is because this could have the most direct impact on the profitability of customers who are taking the credit month by month.

In the presentation of the challenge you must explain how you have defined the problem.

2. Hypothesis

The most important thing is to make a list /brainstorming of the variables that you consider to be most important.

Remember that the quality of your model will depend directly on the quality of your hypothesis..

For example, what variables can affect the most when we try to predict whether a credit card falls within the group of delinquents?

- Salary: A high salary is a predictor of financial stability
- Type of work: A person with a more stable job has greater financial stability
- History
- Socio-educational level
- And many more...

Should we think about our hypothesis before or after exploring the data? This can allow you to think without a previous bias and explore additional data sources.

In the presentation you will have to explain what your **hypotheses are**

3. Data extraction and analysis

That's not the case with this challenge. de However, it is very useful to collect data from different sources and combine it before data exploration:

- Identification of variables. Which is our target and which are not. Data type and category.
- Univariable analysis
- Analyze multivariable
- Lost data.
- Remove outliers
- Variable transformation s

As important as understanding the problem is understanding the data we have available. It is common to do an exploratory analysis of data to familiarize ourselves with them.

Exploratory analysis often does graphs, correlations, and descriptive statistics to better understand what story the data is telling us. It also helps to estimate whether the data we have is sufficient, and relevant, to build amodel.

4. **Developing the predictive** model

- How to explore the dataset
- Prepare the dataset
- Choosing a model
- Model evaluation
- Interpret the model and insights..

👉 LINK

Challenge assessment criteria

The jury will evaluate the following characteristics* of the solution submitted by the group:

- Data display.
- Data **allocation,** variable transformation, and data exploration.
- Development and evaluation of the model.
- Hypothesis and solutions beyond dataset variables. .
- Model score.

In both the presentation and the code, each part of the solution must be duly justified and commented. The jury will ***not*** *ask any questions to the team.*

What is the challenge?

The goal of the data challenge is to diagnose whether or not a patient has diabetes, based on certain diagnostic measures included in the dataset.

The dataset consists of several predictive medical variables and a target variable. Predictor variables include the number of pregnancies patients (all women) have had, their BMI, insulin level, age, etc.

What is diabetes?

According to the NIH, "Diabetes is a disease that occurs when blood glucose, also called blood sugar, is too high. Insulin, a hormone produced by the pancreas, helps the glucosa of food enter the cells to be used as energy. Sometimes the body doesn't make enough or no insulin or doesn't use insulin well. Glucose remains in the blood and does not reach the cells.

Over time, having too much glucose in your blood can cause health problems.

What are the different types of diabetes? The most common types of diabetes are type 1, type 2, and gestational diabetes.

Type 1 diabetes: If you havetype 1 diabets, your body does not produce insulin. The immune system attacks and destroys cells in the pancreas that produce insulin. Type 1 diabetes is usually diagnosed in children and young adults, although it can occur at any age. Peoplewith type 1 diabetes need to take insulin daily.

Type 2 diabetes: Eu body does not produce or use insulin well. You can develop type 2 diabetes at any age, even during childhood. However, this type se of diabetes occurs most often in middle-aged and older people. Type 2 is the most common type of diabetes.

Gestational diabetes: Gestational diabetes develops in some women when they are pregnant. La Most of the time, this type of diabetes goes away after the baby is born. tiene However, if you have had gestational diabetes, you are more likely to develop type 2 diabetes later. Sometimes diabetes diagnosed during pregnancy is actually type 2 diabetes.

Challenge phases

Part 1 - Description of data:

The set of data that we obtain and want to introduce into the model will be our **dataset.**

Because of the dataset, each computer is expected to read the variable name and understand the information it encodes. From this information you should extract the ranges in which each variable should move and therefore look for nulls.

In addition to performing this task, they are expected to obtain a description of the data distribution of each variable, basic information of media, fashion, correlations, etc.

Part 2 - Data Set Alterations

Once the dataset is understood, the next step is to edit and clean it up.

We'll do it in a ***notebook/book*** (that. Jupyter's ypinb. A notebook is a cloud programming document.

And becauseonly one data source is available, these variables (if deemed relevant) will be obtained as a combination of the variables already present in the dataset. In order to obtain these variables, it is recommended to represent them in different diagrams (bar plot or scatterplot) to study their distribution visually and see the possible existence of patterns.

Part 3 - Training

Once the dataset has been edited at the liking and consideration of the group, we will proceedto find and train a model that solves the exercise. Since each model depends on one or more hyperparameters that have to be configured by each group, it is expected that a study will be made of the different values that you can take and give a reason why you are choosing a certain value. Again, the graphic representations are of great help in this section.

Finally, the results of the model should be analyzed, in classification usually a confusion matrix isused, to analyze which cases are most conflicting for the model and analyze whether it is a suitable scenario or whether another metric should be looked for the selection of the model.

How to process the data?

Real-world data is rarely clean and homogeneous. This is due to the following reasons:

- They tend to be incomplete.
- We found noise.
- Corrupt data.
- Information loading failures.
- Incomplete extraction of data.

Therefore, it is an important task for a data scientist to treat the data by filling the missing values to have a robust model. It's important to manage the lack of data as they could lead to a misprediction or classificationfor any model used. Real-world data often has lost values.

- If you want to read more (step by step) how to face a data challenge. [LINK]

- Tutorial on how to launch the notebook in Sagemaker and train a model [LINK]

- Otro step by step de XGBoost [LINK]

How do I get my solution?

Stick to the evaluation criteria. Explain clearly how you have defined the problem and hypothesis, and then focus on detailing the processing of the data and the model. You'll have a few minutes. Part of the jurywill bereviewing your code solution.

Now... get to work! 💪