Analytics Vidhya

Hypothesis Generation for Data Science Projects - A Critical Problem Solving Step

BEGINNER DATA EXPLORATION DATA SCIENCE USE CASES

This article was published as a part of the Data Science Blogathon.

Introduction

The first step towards problem-solving in data science projects isn't about building machine learning models. Yes, you read that right!

That distinction belongs to hypothesis generation – the step where combine our problem solving skills with our business intuition. It's a truly crucial step in ensuring a successful data science project.

Let's be honest – all of us think of a hypothesis almost everyday. Let us consider the example of a famous sport in India – cricket. It is that time of the year when IPL fever is high and we are all absorbed in predicting the winner.

If you have been guessing which team would win based on various factors like the size of the stadium and batsmen present in the team with six hitting capabilities or batsmen with high T20 averages, then kudos to you all. You have all been making an educated guess and generating hypotheses based on your domain knowledge of the sport.

Similarly, the first step towards solving any business problem using machine learning is hypothesis generation. Understanding the problem statement with good domain knowledge is important and formulating a hypothesis will further expose you to newer ideas of problem-solving.

So in this article, let's dive into what hypothesis generation is and figure out why it is important for every data scientist.

Table of Contents

What is Hypothesis Generation?

Hypothesis generation is an educated "guess" of various factors that are impacting the business problem that needs to be solved using machine learning. In framing a hypothesis, the data scientist must not know the outcome of the hypothesis that has been generated based on any evidence.

> "A hypothesis may be simply defined as a guess. A scientific hypothesis is an intelligent guess." – Isaac Asimov

Hypothesis generation is a crucial step in any data science project. If you skip this or skim through this, the likelihood of the project failing increases exponentially.

Hypothesis Generation vs. Hypothesis Testing

This is a very common mistake data science beginners make.

> Hypothesis generation is a process beginning with an educated guess whereas hypothesis testing is a process to conclude that the educated guess is true/false or the relationship between the variables is statistically significant or not.

This latter part could be used for further research using statistical proof. A hypothesis is accepted or rejected based on the significance level and test score of the test used for testing the hypothesis.

To understand more about hypothesis testing in detail, you can read about it here or you can also learn it through this course.

How Does Hypothesis Generation Help?

Here are 5 key reasons why hypothesis generation is so important in data science:

- Hypothesis generation helps in comprehending the business problem as we dive deep in inferring the various factors affecting our target variable
- You will get a much better idea of what are the major factors that are responsible to solve the problem
- Data that needs to be collected from various sources that are key in converting your business problem into a data science-based problem
- Improves your domain knowledge if you are new to the domain as you spend time understanding the problem
- Helps to approach the problem in a structured manner

When Should you Perform Hypothesis Generation?

The million-dollar question – when in the world should you perform hypothesis generation?

- The hypothesis generation should be made before looking at the dataset or collection of the data
- You will notice that if you have done your hypothesis generation adequately, you would have included all the variables present in the dataset in your hypothesis generation
- You might also have included variables that are not present in the dataset

Case Study: Hypothesis Generation on "New York City Taxi Trip Duration Prediction"

Let us now look at the "**NEW YORK CITY TAXI TRIP DURATION PREDICTION**" problem statement and generate a few hypotheses that would affect our taxi trip duration to understand hypothesis generation.

Here’s the problem statement:

To predict the duration of a trip so that the company can assign the cabs that are free for the next trip. This will help in reducing the wait time for customers and will also help in earning customer trust.

Let’s begin!

Hypothesis Generation Based On Various Factors

1. Distance/Speed based Features

Let us try to come up with a formula that would have a relation with trip duration and would help us in generating various hypotheses for the problem:

TIME=DISTANCE/SPEED

Distance and speed play an important role in predicting the trip duration.

We can notice that the trip duration is directly proportional to the distance traveled and inversely proportional to the speed of the taxi. Using this we can come up with a hypothesis based on distance and speed.

- **Distance**: More the distance traveled by the taxi, the more will be the trip duration.
- **Interior drop point**: Drop points to congested or interior lanes could result in an increase in trip duration
- **Speed:** Higher the speed, the lower the trip duration

2. Features based on Car

Cars are of various types, sizes, brands, and these features of the car could be vital for commute not only on the basis of the safety of the passengers but also for the trip duration. Let us now generate a few hypotheses based on the features of the car.

- **Condition of the car**: Good conditioned cars are unlikely to have breakdown issues and could have a lower trip duration
- **Car Size**: Small-sized cars (Hatchback) may have a lower trip duration and larger-sized cars (XUV) may have higher trip duration based on the size of the car and congestion in the city

3. Type of the Trip

Trip types can be different based on trip vendors – it could be an outstation trip, single or pool rides. Let us now define a hypothesis based on the type of trip used.

- **Pool Car**: Trips with pooling can lead to higher trip duration as the car reaches multiple places before reaching your assigned destination

4. Features based on Driver Details

A driver is an important person when it comes to commute time. Various factors about the driver can help in understanding the reason behind trip duration and here are a few hypotheses this.

- **Age of driver**: Older drivers could be more careful and could contribute to higher trip duration
- **Gender**: Female drivers are likely to drive slowly and could contribute to higher trip duration
- **Driver experience**: Drivers with very less driving experience can cause higher trip duration
- **Medical condition**: Drivers with a medical condition can contribute to higher trip duration

5. Passenger details

Passengers can influence the trip duration knowingly or unknowingly. We usually come across passengers requesting drivers to increase the speed as they are getting late and there could be other factors to hypothesize which we can look at.

- **Age of passengers:** Senior citizens as passengers may contribute to higher trip duration as drivers tend to go slow in trips involving senior citizens
- **Medical conditions or pregnancy:** Passengers with medical conditions contribute to a longer trip duration
- **Emergency:** Passengers with an emergency could contribute to a shorter trip duration
- **Passenger count:** Higher passenger count leads to shorter duration trips due to congestion in seating

6. Date-Time Features

The day and time of the week are important as New York is a busy city and could be highly congested during office hours or weekdays. Let us now generate a few hypotheses on the date and time-based features.

Pickup Day:

- Weekends could contribute to more outstation trips and could have a higher trip duration
- Weekdays tend to have higher trip duration due to high traffic
- If the pickup day falls on a holiday then the trip duration may be shorter
- If the pickup day falls on a festive week then the trip duration could be lower due to lesser traffic

Time:

- Early morning trips have a lesser trip duration due to lesser traffic
- Evening trips have a higher trip duration due to peak hours

7. Road-based Features

Roads are of different types and the condition of the road or obstructions in the road are factors that can't be ignored. Let's form some hypotheses based on these factors.

- **Condition of the road**: The duration of the trip is more if the condition of the road is bad
- **Road type**: Trips in concrete roads tend to have a lower trip duration
- **Strike on the road:** Strikes carried out on roads in the direction of the trip causes the trip duration to increase

8. Weather Based Features

Weather can change at any time and could possibly impact the commute if the weather turns bad. Hence, this is an important feature to consider in our hypothesis.

- **Weather at the start of the trip**: Rainy weather condition contributes to a higher trip duration

End Notes

- After writing down our hypothesis and looking at the dataset you will notice that you would have covered the writing of hypothesis on most of the features present in the data set. There could also be a possibility that you might have to work with fewer features and the features on which you have generated hypotheses are not currently being captured/stored by the business and are not available.
- Always go ahead and capture data from external sources if you think that the data is relevant for your prediction. Ex: Getting weather information
- It is also important to note that since hypothesis generation is an estimated guess, the hypothesis generated could come out to be true or false once exploratory data analysis and hypothesis testing is performed on the data.

I hope you were able to get some value from this post. If there is anything that I missed or something was inaccurate or if you have any feedback, please let me know in the comments. I would greatly appreciate it.

Article Url - https://www.analyticsvidhya.com/blog/2020/09/hypothesis-generation-data-science-projects/

kaushal113