Regression Projects

Hi, learner. Great to see that we finally rounded up the majority of the concepts in Regression. I hope you were able to see how most of these algorithms play a role in solving ML-related problems and how they are generally applied at an implementation and conceptual level.

Now it's your turn to build something! You are going to embark on 2 projects throughout the month of August. Here are the details of the project:

1. **Project 1** has been provided for you in 2 sets; either you choose a capstone case study or come up with your own regression problem project that you want to solve. We have listed out two capstone case studies that you might want to consider. Take a look at them and see if you are willing to pick up one of them.

   If you want to create your own personal project instead, you can write a proposal to us detailing what this project is, the data set you want to use, and the necessary details that follows the project. Send this to us via the community's email address.

   If you do not choose a capstone project or propose your project before the deadline, we will choose one for you!

2. **Project 2** is a compulsory capstone project where you will get to write a blog post on how to solve a Machine Learning regression problem from end-to-end detailing the steps/phases involved in the standard ML workflow. Please read that description for more.

Project 1: Propose a Project or Pick One of the Capstone Case Study Projects Below:

Case Study 1: Removing Inconsistencies in Concrete Compressive Strength

Description

Andrew had an hour-long meeting with Smith regarding the frequent complaints he had been receiving from their existing clientele. Smith was the sales representative at a construction aggregates company and was facing difficulties in meeting his sales targets. Smith believed that the reason for his difficulties was that the existing clients had complaints regarding quality inconsistencies—this, when coupled with negative word of mouth in the market, was making it tough for him to seal deals with potential customers.

Andrew, the plant supervisor, was Smith's friend. At first Smith tried to find a solution for this problem, but when things got out of his control, he decided to have a meeting with Andrew.

There is usually friction between the production and sales departments in most companies. Both of them usually fight over their yearly budgets and their plans for meeting their annual targets. Smith knew this but he felt sure that things would be different with Andrew as they had been friends for some time. However, as the meeting kicked off and Smith presented the matter to Andrew, to his dismay Andrew wasn't ready to acknowledge any inconsistencies in production. As Smith recalled Andrew saying, "These complaints don't make any sense to me. You ask why! Well first of all, we had an upgrade at our production plant in the last quarter, and you know for a fact that the raw material we use goes through a thorough testing prior to being used for production."

The meeting didn't help Smith in any way, and he let the matter go for some time. One day, when he was in a networking session during one of his company's corporate events, Smith heard some executives from the production department talking about the recent upgrade of the plant. He recalled hearing one of the executives say, "Though we don't have any testing improvement in our recent upgrade for measuring concrete compressive strength, then too our product quality is superior thanks to our robust testing of raw material used."

When Smith heard this, he did some research and was surprised to discover that the complaints had gained momentum ever since the upgrade had taken place. However, he also discovered that the procurement of the testing equipment was put on hold due to a

shortage of funds allocated to the procurement department. Smith knew for a fact that he couldn't afford to wait until the next year for the issue to resolve on its own, and he had to find a way to make things right.

He had heard that the recent upgrade meant that machines had recording data. Smith decided to set up a meeting with Claire (the manager of analytics) to see if she could help in figuring out the reason behind the inconsistencies in concrete compressive strength. The meeting went well for Smith, as Claire assured him of the in-house analytics capabilities. As Claire recalled, "At that time we had started up the analytics vertical, but I knew for sure that we had the muscles to formulate a model which can answer this Anomaly."

Smith was relieved and now had to formulate a strategy to take his findings forward—that is, either approach management with his findings or share his findings to the production department for them to integrate those into their inner processes. However, Smith was curious if the data at hand was strong enough to deduce some quality findings. **Moreover, he was interested to see which factors influence concrete compressive strength the most.**

Smith was interested to know which features of the dataset Claire would be working on. Hence, Claire came up with the data dictionary in the table below.

Attribute	Description
Cement (kg in a m3 mixture)	Amount of cement used in a m3 mixture (unit: kg)
Blast furnace slag (kg in a m3 mixture)	Amount of blast furnace slag used in a m3 mixture (unit: kg)
Fly ash (kg in a m3 mixture)	Amount of blast fly ash used in a m3 mixture (unit: kg)
Water (kg in a m3 mixture)	Amount of water used in a m3 mixture (unit:kg)
Superplasticizer (kg in a m3 mixture)	Amount of superplasticizer used in a m3 mixture (unit: kg)
Coarse aggregate (kg in a m3 mixture)	Amount of coarse aggregate used in a m3 mixture (unit: kg)
Fine aggregate (kg in a m3 mixture)	Amount of fine aggregate used in a m3 mixture (unit: kg)
Age (days)	Age of concrete (unit: days)

Concrete compressive strength	Concrete compressive strength which is measured in MegaPascal (MPa). This is the unit for pressure or stress and is the common unit to determine compressive strength of concrete. This is what you are going to be predicting.

It is clear that Claire would have to create a model that predicts the concrete compressive strength given newer records from the machine.

Your Challenge:

Your challenge is to build an end-to-end ML project that solves this problem. We will not consider deploying the solution at this stage but we can stop at making sure the model works well on the test data so we can be at least sure that it will successfully generalize in production.

Data dump: http://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength

Number of rows: 1030 rows/observations.

If you choose to accept this capstone, you will be given further details on how to go ahead with the project and what an acceptable performance is.

Case Study 2: Estimating the price of a house based on the data available.

In this course, we will constantly be simulating fictitious business problems that very well resemble business challenges you will encounter in your local society as a Data Scientist. This week, you are immersed in a working experience as a Machine Learning consultant hired for Ahmed Zakka LTD. by the operations manager Mr Dokubo Kabiri. They have a problem they think requires Machine Learning. Find more below.

Company Overview

Ahmed Zakka LTD. is a real estate company with their headquarters in the Minneapolis area of US, and with a branch in a very urban area in Port Harcourt. Ahmed Zakka LTD. (or AZ) has been around for 10 years and has explored areas from marketing consultancy to print services, and only settled for offering real-estate services 6 years ago when they acquired a start-up that already had lots of historical data on housing across different cities in the United States.

They currently offer a wide range of real-estate services from rent agency to property acquisition. The core of their business lies in properly advising clients on the best prices to get housing properties before they sky-rocket. Majority of their clients have had testimonials on the company helping them save a lot of money on housing properties acquisitions. They have been largely successful at this because they employ a lot of highly qualified property and pricing analysts to help that process.

Problem Overview

The problem here is that the stakeholders of AZ LTD. have not been very happy with AZ's revenue performance. They have been running on loss for the past 2 years because of the cost of paying highly qualified property and pricing analysts to help their customers. The stakeholders passed down their aggression on the current CEO of AZ LTD. Tonye Nnamdi. Under pressure, Mr Tonye prompted Mr Dokubor, the operations manager to look for a solution to the current problem. After some "little research", Mr Dokubo found out that Artificial Intelligence can "magically' solve the problems they are currently having.

Mr Dokubo asked around and eventually, you were recommended by a couple of sources as the best person to meet to solve this problem. Mr Dokubo gets on the phone with you;

Mr Dokubo: Hello, my name is Mr Dokubo from Ahmed Zakka LTD... Am I speaking to Mr You?

You: Hello, Mr Dokubo, yes you are speaking to Mr You. How may I help you?"

Mr Dokubo goes on to discuss the problem...

Mr Dokubo: We would want to hire you to come to provide a solution to our current problem. We agree to all your payment terms, that is how desperate we are. So are you in?

Your Challenge:

Your challenge is to build an end-to-end ML project that solves this problem. We will not consider deploying the solution at this stage but we can stop at making sure the model works well on the test data so we can be at least sure that it will successfully generalize in production.

What you are trying to do is to build a system that helps solve their problem in the US as a proof of concept, then you can now come up with the necessary strategy of replicating it here in Port Harcourt city.

Data dump:
https://github.com/Port-Harcourt-School-Of-AI/Practical-Machine-Learning-Course-2020/blob/master/AZ_housing_data_US.csv

Number of rows: 42,704 rows/observations.

If you choose to accept this capstone, you will be given further details on how to go ahead with the project and what an acceptable performance is.

Data Attributes	Description
year	The year the house was built. (Correctly formatted).
stories	The number of story buildings the house has. If it is a single-story house or a two, three-story house.
num_bedrooms	The number of bedrooms available in the house.
full_bathrooms, half_bathrooms	Number of full bathrooms and/or half bathrooms are in the house in question. Half-bathrooms are the ones without showers while full-bathrooms

	are the ones with showers.
livable_sqft, total_sqft	The livable square feet and the total square feet tells us the size of the house. The difference is that the total_sqft includes any outside patio space that is not part of the actual living area of the house.
garage_type	There are 3 possible outcomes here; an attached house garage, a detached house garage, and no house garage.
garage_sqft	This tells us the size of the garage if it exists or 0 sqft if it does not.
carport_sqft	These are like tents for cars and are not as enclosed as garages. This tells us the size of the carport in square feet.
has_fireplace, has_pool, has_central_heating, has_central_cooling	This tells us each feature the house has, does it have a fireplace in it? True or False? How about a swimming pool? Does it have a central house heater and/or Cooler? These are of course likely to affect the price of the house.
house_number, street_name, unit_number, city, zip_code	These are just the addresses of the house if such information is available.
sale_price	The actual price that the house was sold.This is the value our ML system is going to be predicting, giving newer data sets.

Project 2: Compulsory Capstone (Blogging):

Your compulsory capstone will be to write a detailed blog post on how you think an end-to-end Machine Learning problem can be solved. Taking into account all the steps and workflows necessary for a successful Machine Learning project.

You could decide to come up with your own hypothetical problem, or pick a problem and propose, or use any of the problems from the capstone case studies above. Although this is not strict, you can decide to follow the format below;

1. Introduction (An overview of what the blog is all about and what the readers will hope to learn at the end of the blog post).

2. Get readers to understand the most important concepts of Machine Learning and Data Science relevant to the blog post like; definition of the term, examples of applications of the technology, typical steps involved in the successful execution of a Machine Learning project, challenges involved in building Machine Learning projects.

3. Introduce the problem you are trying to conceptually solve, including the company you are solving the problem for.

4. Detail the steps in a typical Machine Learning workflow and try to conceptually go from needs analysis to deploying the solution as you explain what the various steps for solving the problem will entail.

5. Conclude your blog post, helping readers understand what they might have gained reading your blog post.

6. If you can, make “further readings” recommendations to your readers.