Introduction to Predictive Analytics

Introduction to Analyzing Data for Business Goals

As discussed in the previous modules, there are four types of data mining analytics as follows:

* **Descriptive:** What happened? Helps uncover valuable insight into the data being analyzed
* **Diagnostic:** Why did it happen? Helps understand the relationships and patterns in the data
* **Predictive:** What is likely to happen in the future? Helps forecast the future behavior of people and markets
* **Prescriptive:** What should I do about it? How should we respond to potential future events based on the analysis? Uses optimization and simulation algorithms to provide guidance and understanding on decisions and answers

This module introduces *predictive analytics*.

# What is Predictive Analytics?

Quite simply, predictive analytics looks at the future. How will sales change over the next year? What trends in fashion and style will be popular next spring? How likely is this customer to pay back their loan? These are the kinds of questions that predictive analytics helps answer.

But it’s not just about the future. You do have to have some knowledge of the past. Historical data paves the road to understanding the future. By using historical data, you can build a mathematical model that extends the past into the future. And based on the predictions from your model, you or your company can make appropriate business decisions. Here are a few examples:

**Autonomous Vehicles:** Tesla, a well-known manufacturer of electric automobiles, is spearheading the development of autonomous vehicles and equipping them with a variety of driver assistance technologies. Predictive analytics carried out in real-time takes sensor data—video, radar, global positioning—and applies algorithms that control vehicle movement including steering and acceleration.

**Electricity Production:** Pacific Gas and Electric in California (PG&E) needs to forecast both electricity demand and fuel prices. Aided by sophisticated computer applications, PG&E can use historical data for weather and electricity consumption to forecast future demand. Combined with fuel price data and spot market costs for imported energy, the company’s engineers can determine the optimal mixture of locally produced power and imported power to maintain a high level of system reliability at an optimally low price.

**Financial:** U.S. Bank in California uses machine learning and various other quantitative tools to predict potential customers’ credit risk.

Applications of predictive analytics can be ranked in order of increasing explanatory power. The examples below are listed in order of increasing confidence in being able to make accurate and repeatable predictions:

* **Associations:** Linking going fishing with eating pizza (The two may be correlated, but it’s not possible to say with certainty that one *causes* the other to happen.)
* **Sequences:** Linking a new job to buying a house (This relationship has a greater likelihood of being causal because it’s clear that one occurred first—getting the new job. However, it’s still not possible to infer causality without having any additional information. Perhaps the family had been planning to buy a new house even before the new job opportunity came about.)
* **Classifications:** Linking observable customer behavior to churn (Classification is actually one of the most popular predictive models in use today and will be discussed below. The model takes past customer behavioral data and attempts to predict whether they will remain a customer or switch to a competitor.)
* **Forecasting:** Predicting future buying habits of customers based on past patterns
* **Estimation:** Determining an approximate but sufficiently useful value, such as for the probability that a particular marketing campaign will cause a customer to purchase a particular product (also called *conversion* in marketing parlance)

# Techniques and Methodologies

The most popular predictive models are *classification* and *regression* (which are *supervised learning* models), and *clustering* (which is an *unsupervised learning* model). In a supervised learning model, data from the past is used to *train* the model so it can make predictions about the future. In an unsupervised learning model such as clustering, the model looks at existing relationships among data and assigns the data elements into similar groups.

A *classification model* predicts discrete values that are often binary, such as yes/no or 0/1, but sometimes can take on more values (e.g., 0, 1, 2, 3,...). The example mentioned earlier is a classic application of this kind of model. For instance, telecommunication companies find it extremely useful to know whether a particular customer is likely to maintain their service or switch to a different provider so they can step in with incentive offers.

*Decision trees* are a type of classification model that examines the path a person or organization takes in reaching a decision. This path involves multiple possibilities along the way and resembles a tree in that each branch represents a set of options among which the decision-maker must choose.

A *regression model* predicts continuous values such as a person’s salary based on the number of years of education they’ve had or how much an online customer is likely to spend in the future based on their previous spending.

Although this course serves primarily as an introduction to predictive analytics, regression models play such an important role in data science and statistical analysis that they’re worth examining in more detail here. We’ll focus on linear regression.

## Statistical Methods

### Linear Regression

Linear regression is a statistical modeling technique that describes the mathematical relationship between a *dependent variable, y*, and one or more *independent variables, x1, x2, x3,*..., *xi*. The dependent variable is also called the *response variable* and the independent variables are also called *explanatory* or *predictor* variables. Continuous predictor variables are sometimes also called *covariates*, and categorical predictor variables (which have discrete values) are also called *factors*.

The linear regression equation with one independent variable takes the form

*y = β0 + β1x1 + β2x2 + ⋯+ βnxn + ε*

where

*y* is the response variable (the value being predicted)

*βn* is the kth coefficient, where *β0* is the constant term in the model

*xn* is the nth observation

*ε* is the error term, also called “noise”

Consider a simple example in two dimensions, in which you want to predict a person’s salary based on the number of years of working experience they have. Thus, we have just one independent variable and the regression equation simplifies to

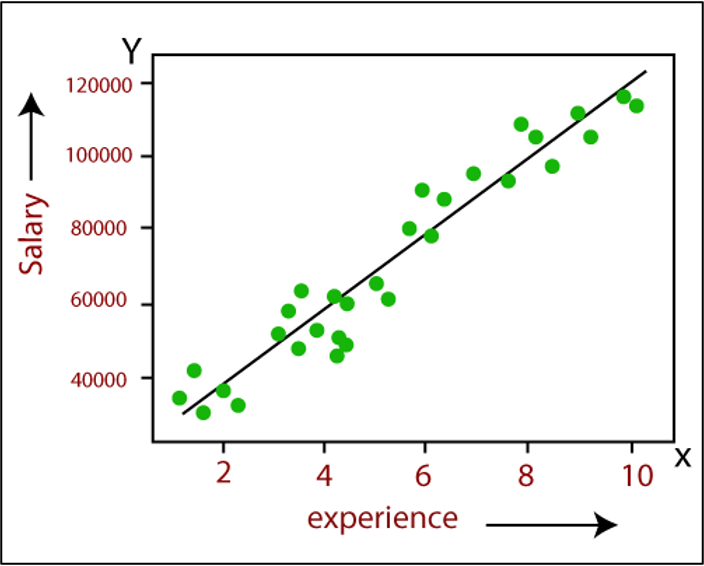
*y = β0 + β1x1 + ε*

where

*y* is the response variable (salary)

*β0* is the y-axis intercept (salary when the person has zero years of experience)

*β1* is the slope of the line (increase in salary for a one-unit increase in years of experience)



*Figure 1: Example linear regression in which salary is a function of years of experience*

Statistical analysis applications typically calculate the coefficients *βi* for a given set of input data along with “goodness-of-fit” statistics that indicate how closely the observed values fit the calculated regression line. These statistics include the:

* **Standard deviation of the residuals (**also called the **root mean square error-RMSE)**, where a residual (also called a prediction error) is the absolute value of the difference between the actual (measured) response variable and the corresponding value given by the regression equation
* **Coefficient of determination** (R2), which measures the goodness-of-fit and is the square of the correlation coefficient (R).
* **Adjusted coefficient of determination** (adjusted R2), which takes into account the number of independent variables and is used in multiple regression models.
* **Average error,** which is the numerical difference between a predicted value and its corresponding actual value

R2 and the adjusted R2 take on values between 0 and 1, where a 0 signifies that the calculated regression line does not fit the data at all and a 1 indicates a perfect fit. Note that the adjusted R2 value increases with the addition of dependent variables to the model even if the model’s explanatory power does not.

### Types of Linear Regression Models:

There are several types of linear regression models:

* **Simple:** a model with only one predictor (as in the preceding example)
* **Multiple:** a model with multiple predictors *xi*
* **Multivariate:** a model with multiple response variables

### Logistic Regression

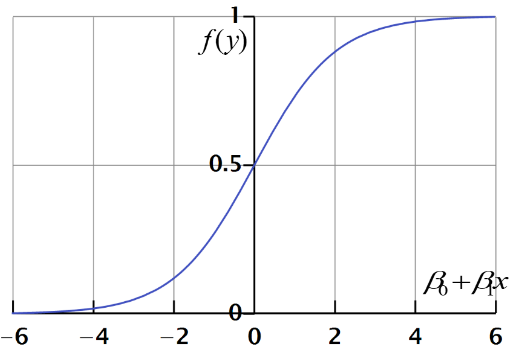
*Logistic regression* is a popular, statistically sound, probability-based classification algorithm that employs supervised learning. It was developed in the 1940s as a complement to linear regression methods and has been used extensively in numerous disciplines including the medical and social sciences fields.

Logistic regression is similar to linear regression in that it aims to regress to a mathematical function that explains the relationship between the response variable and the explanatory variables using a sample of past observations (training data). However, it differs from linear regression in one major way: its output (response variable) is a class rather than a numerical variable. That is, while linear regression is used to predict a *continuous* numerical variable, logistic regression is used to classify a *categorical* variable.

Even though the original form of logistic regression was developed for binary output variables (e.g., 1/0, yes/no, pass/fail, accept/reject), the present-day modified version is capable of predicting multiclass output variables. If there is only one predictor variable and one predicted variable, the method is called simple logistic regression (similar to simple linear regression, which has just one independent variable and one dependent variable).

In essence, logistic regression takes the natural logarithm of the odds of the response variable to create a continuous criterion as a transformed version of the response variable. Thus the logit transformation is referred to as the link function in logistic regression. Even though the response variable in logistic regression is categorical or binomial, the logit is the continuous criterion on which linear regression is conducted.

Figure 2 (below) shows a logistic regression function where the odds are represented on the *x*-axis (a linear function of the independent variables), and the probabilistic outcome is shown on the *y*-axis (i.e., response variable values change between 0 and 1).



*Figure 2: Example of logistic regression*

## Machine Learning Methods

Another important technique for making predictions is called *machine learning*, which applies statistical algorithms to very large amounts of data to detect patterns. Machine learning is especially useful for these kinds of applications:

* Product recommendations based on a customer’s previous online purchases
* Speech recognition and translation
* Fraud detection (e.g., identifying unusual credit card purchasing patterns)
* Creditworthiness scoring
* Predictive maintenance (e.g., identifying mechanical breakdown patterns)

Machine learning goes beyond predictive analytics techniques such as regression modeling because it can handle huge amounts of data without being limited to numerical or categorical data. Anything from image and audio files to full-length movies is amenable to machine learning.

There are two major types of machine learning:

* *Supervised learning*, in which “training” data is labeled so the model can be fine-tuned to fit specific patterns. That is, in supervised learning, there is a target variable that supervises the learning process.
* *Unsupervised learning*, in which the model simply looks for any kind of pattern that may be present in the data. In unsupervised learning, there is no target variable, and therefore, the goal is to discover the inherent natural patterns (associations and natural groupings) within the data.

We will examine these two types of machine learning in more detail in the following section.

### Decision Trees, Neural Networks, and Support Vector Machine

There are numerous machine learning methods developed for predicting or explaining a specific variable of interest. The most popular supervised prediction-focused machine learning methods include decision trees, neural networks, and support vector machines. While decision tree methods use discretized variables (nominal values variables) to build a tree-like structure (a collection of discerning rules) for predictive purposes, neural networks and support vector machines use numerical variables to build a mathematical representation for predictions. All three of these machine learning models are capable of performing both regression and classification type prediction modeling. Theoretical and algorithmic details of these machine learning methods are not covered here but can be found in the reference material.

# Considerations for Applying Predictive Analytics

As you know, predictive analytics follows descriptive and diagnostic analytics in the analytics hierarchy. By the time you’ve accumulated the data needed to undertake predictive analytics, you have what is needed to conduct both descriptive and diagnostic analytics. Generally, you apply descriptive analytics first to answer the question, “What happened?” Next, you carry out diagnostic analytics to answer the question, “Why did it happen?” And finally, you are ready for predictive analytics.

At this point, we’re ready to explore how predictive models used in machine learning are “trained”—how they learn from the past so they can predict the future.

## Supervised and Unsupervised Learning

Machine learning models use large sets of data as part of their “training regimen.” Let’s begin by looking at the first type of learning, *supervised learning* to see how this is done.

### Supervised Learning

As the name indicates, *supervised learning* implies that a “supervisor” or “teacher” is present. We introduce input data that is *labeled* to our model, meaning that the corresponding output results are known. Returning to regression analysis for a moment, recall that your statistics software package takes existing measured values of both the input (independent) and output (dependent) variables, and determines a regression equation. You can then apply this equation to new input data and obtain a prediction of the output. In effect, your existing measured data (input *and* output) trains the linear regression model so it can be applied to new input data.

This training approach works with other models in addition to linear regression. The details differ but the concept is the same. In summary, training data with known input and output values (e.g., independent and dependent variables) are presented to the model, which then learns how to generate outputs given previously unseen input data.

In addition to regression, here are several other supervised learning models:

* Logistic regression (similar to linear regression but for discrete dependent variables)
* Classification
* Decision Trees
* Support Vector Machines

### Unsupervised Learning

Some models are designed to work without training data and are thus considered *unsupervised*. These models essentially “teach themselves” by finding patterns and similarities in the data.

There are two major categories of unsupervised learning algorithms.

**Clustering:** A clustering model seeks to identify inherent groupings in a dataset. For example, the model might group online shopping customers according to their geographical locations or ages for better targeting of advertising. Types of clustering methods include:

* Hierarchical clustering
* K-means clustering

**Association:** An association model looks for patterns, correlations, and associations among data elements. For example, product suggestion models use association methods to identify specific products that are often sold together or purchased by the same individuals over a period of time. Statements like, “People who buy product X often buy product Y” is the end result of association mining. Types of association methods include:

* Apriori
* Anomaly detection
* Neural networks
* Principal component analysis
* Independent component analysis
* Singular value decomposition

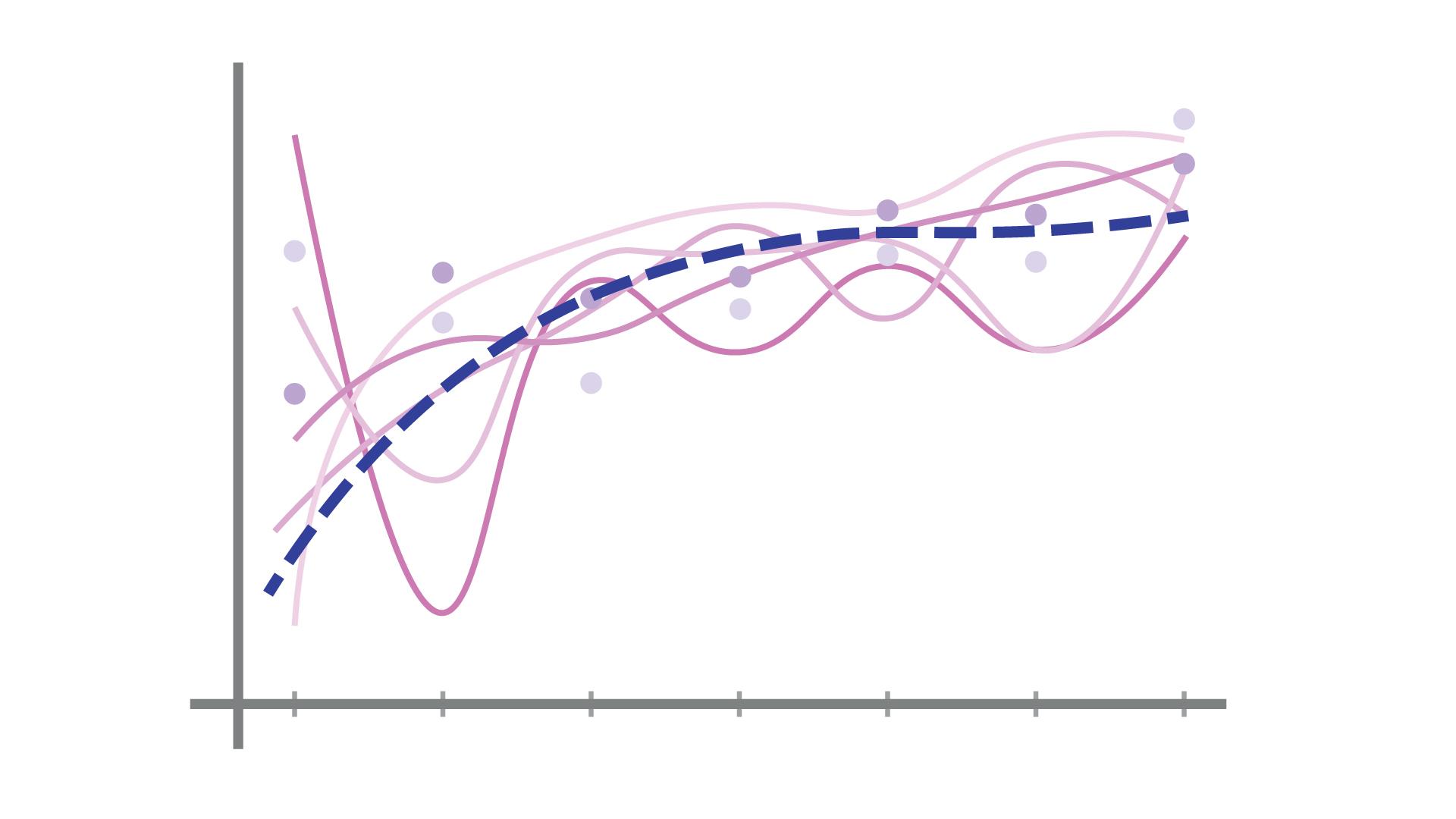
## Training Data Versus Test Data

A common way of training a machine learning model (making it fit a given set of data) is to divide the dataset into two parts: one part is used to do the training (the training data) and the second part is used to test how well the model “learned” (the validation data). Remember, both training and validation datasets include known input and output values, which are used to determine the model’s **parameters**, which are internal configuration variables that are estimated using the training data. The coefficients of a linear regression equation are parameters.

Machine learning models also have **hyperparameters**, which can be adjusted externally by the user to optimize a model’s performance. You can think of a hyperparameter as a value that controls the learning process.

## Model Selection

The simple rule for selecting a model is to choose one that is the most accurate and the least complex. Consider a regression problem in which you want to predict some quantity, say the selling price of a piece of real estate, based on certain attributes of that quantity, say the size, location, and condition of the property. Figure 3 shows a scatterplot of selling prices as a function of one of the property’s attributes. Each curve represents a different model. By examining these curves, you can pick the model that best fits the data in terms of minimizing the average error.



*Figure 3: An example of how several different models can be fit to an existing dataset with varying degrees of accuracy*

# Business Challenges

Google CEO Larry Page once described the perfect search engine as “something that “understands exactly what you mean and gives you back exactly what you want” (Cole, 2017). Today’s shift toward contextual or predictive search is driven by data—big data—and incorporates predictive models that analyze current and historical data to produce scores and other metrics that reflect expected future activity. One of today’s major business challenges is to collect historical data and prepare it for the model. It’s also challenging to select the best predictive models and choose the best model parameters or hyperparameters. Finally, applying predictive models to real-time decision making poses its own set of unique challenges.

## Error-Prone Data

Historical data presents several challenges to the analyst. It can be inconsistent as well as have missing data points, which may lead to incorrect model parameters. In some training datasets, perfectly good data may inadvertently end up being treated as unobservable, leading to the development of a model-based solely on observed, error-prone data, which would lead to erroneous parameters. Such a model would obviously yield incorrect predictions. To improve upon the model development process, a dataset is usually divided into *training data*, *testing data*, and *validation data*, which when used together, allows the analyst to immediately identify models that need to be corrected.

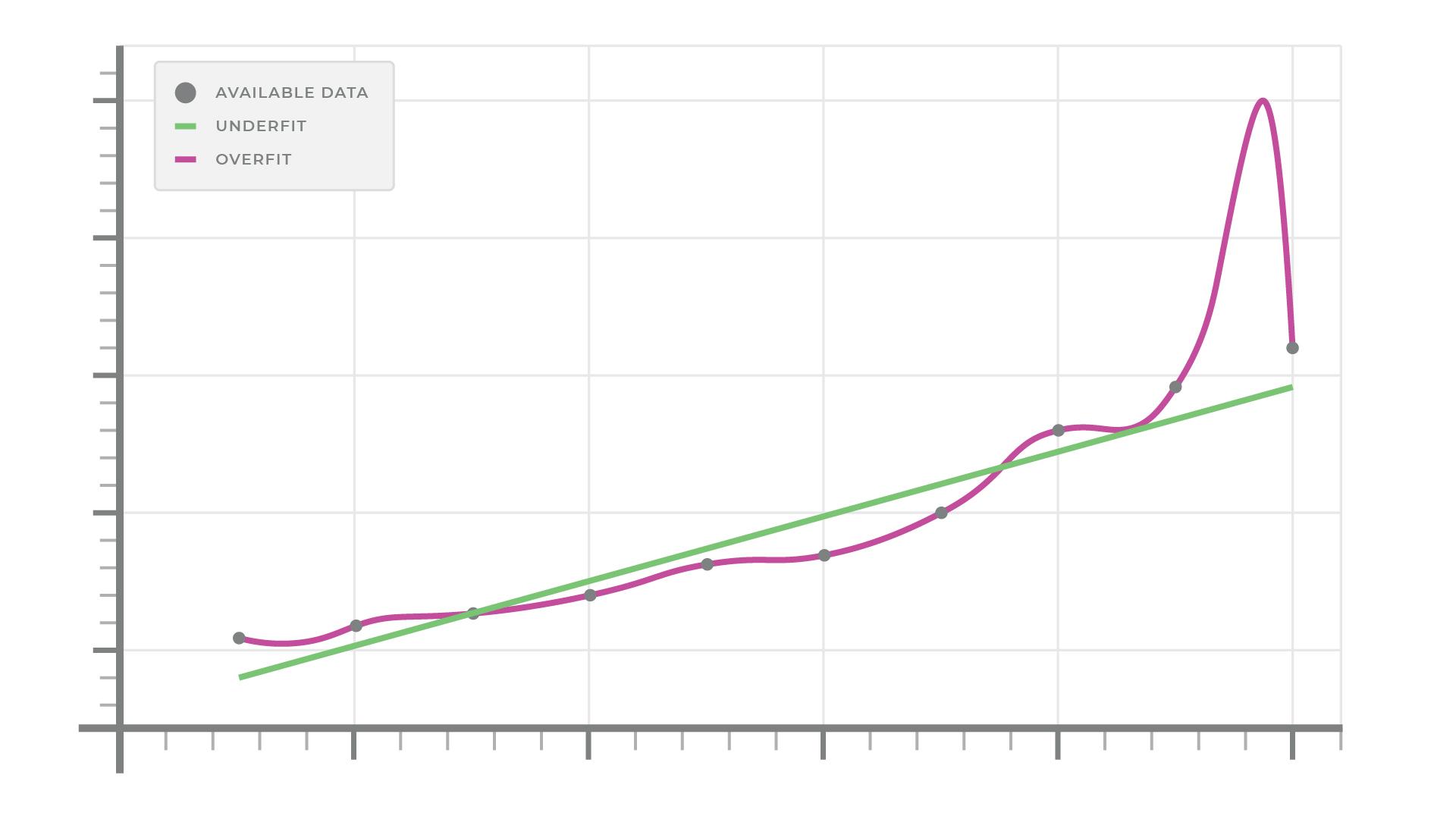
## Bias-Variance Tradeoff

In data modeling, *bias* refers to the “fit” between a model and its associated training data. If the model predicts training data output closely, it is said to be *unbiased* or to have *low bias*. If it doesn’t predict the training data output, it is said to be *biased*. (This is not to be confused with a person being biased.) Bias comes from overly simple models that don’t capture the effects of variations present in the data. When working with observational data, measurement errors can lead to systematic bias in the inferences derived from the model. However, if the modeling assumptions are correct, the model can give unbiased inferences without relying on validation data or specific knowledge about errors that may be present.

The degree to which a model fits a given dataset determines how well the model can be relied upon to provide accurate predictions. There are two kinds of fit: underfit and overfit.

**Underfit:** An *underfit* model is unable to accurately predict variations in the output. You would realize this when applying the model to a test dataset. For example, suppose you develop a model designed to predict an automobile’s gas mileage given its weight. Light cars have higher gas mileage while heavy cars have lower gas mileage. If your model shows little variation in its predictions regardless of the cars’ weight, it is probably underfitting.

**Overfit:** In contrast, an overfit model shows variations in the output that are not actually present. Again, you would realize this when applying the model to a test dataset. For instance, suppose you’re developing a model designed to predict a person’s income tax given their income (see Figure 4). An overfit model manages to correctly predict each data point. However, this doesn’t necessarily mean that it would be accurate for “new” data (data that the model hasn’t previously “seen”). In the figure, the model diverges significantly from what you might expect in between the last two data points. A model that is less complex might actually be more useful and accurate for new data. The figure also shows the output of an underfit model, for comparison.



*Figure 4: Tax as a function of income showing actual data (points), the results of an underfit model, and the results of an overfit model*

## Interpretation & Subjectivity

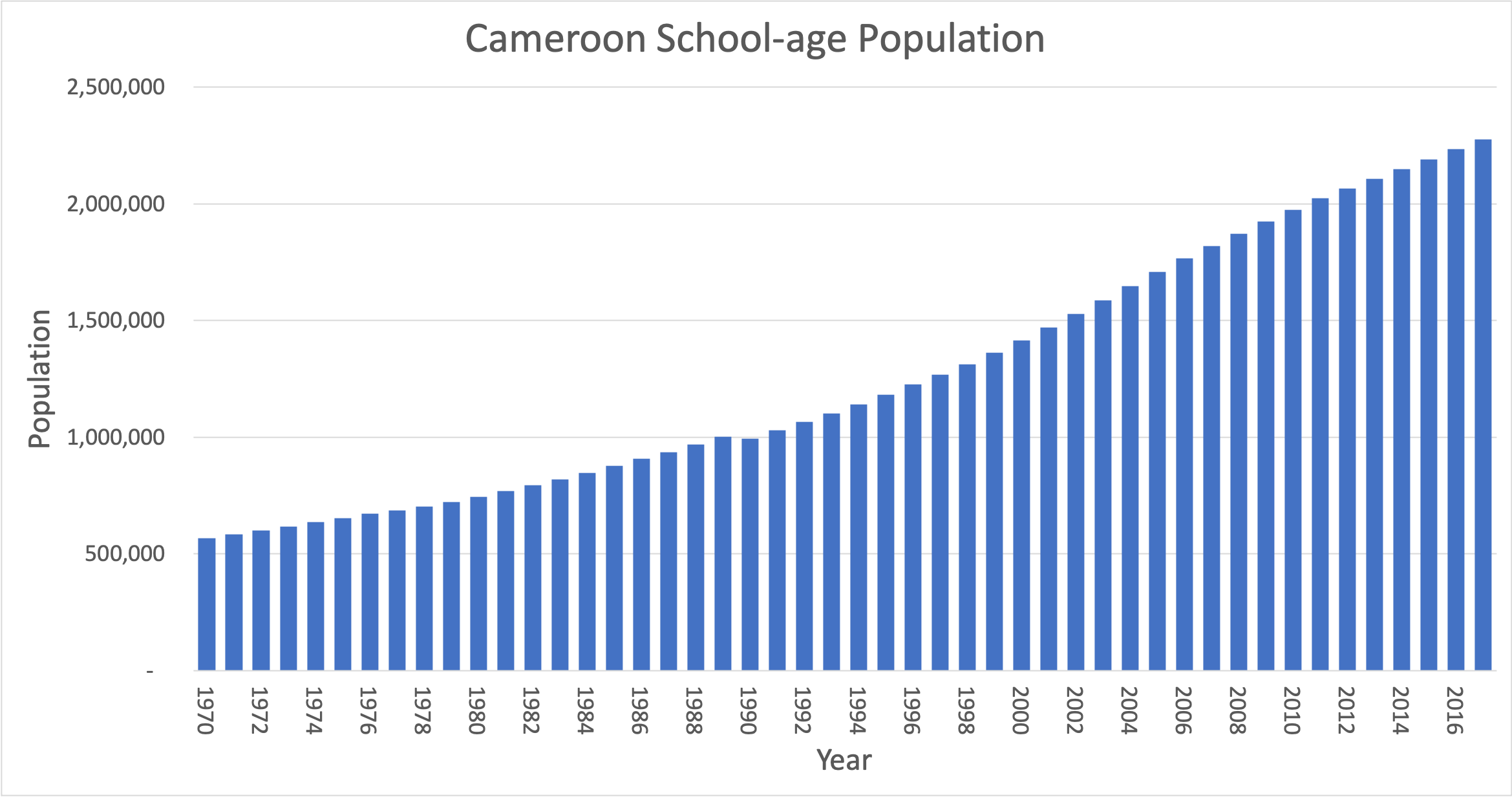
After applying your model to a dataset, you (in your role as a data scientist) need to interpret the predictions it makes and communicate those interpretations to your audience. You would do this with words, tables, or charts that feature scores or performance metrics. You and/or your stakeholders can then take the results and develop a course of action based on the specific business problem you are trying to solve. Of course, your interpretation of the results and the specific actions you take depend on your familiarity and expertise with your business.

# Achieving Business Goals with Predictive Analytics

As you’ve seen, predictive analytics can provide a significant amount of value to your business. Being able to predict customer behavior in response to marketing campaigns, popularity trends for the kinds of products your company sells, and the trustworthiness of potential credit card clients all have obvious benefits for your organization’s profitability. These are just a few of the many applications of predictive analytics.

As an example of how a simple predictive model can provide useful information, consider education in the country of Cameroon. If you are a government official responsible for ensuring that schools have sufficient resources, it would obviously be very useful to be able to predict the number of school-age children that could potentially be attending school in the future so you can ensure that resources are available.

Starting with [historical data](https://drive.google.com/file/d/1EzBGu2S3aUvwz3OXNRlypmRltAqIH4E1/view?usp=sharing) for the number of school-age children in Cameroon, you can develop a simple linear regression model to make predictions for the future. Figure 5 shows this data graphically. While the population does not exactly increase linearly, it appears to be close enough that a linear predictive model can probably provide some value.



*Figure 5: School-age population for Cameroon*

Using the raw data, the resulting linear regression equation is

where

*y* is the population,

*x* is the year,

*a0* is the y-axis intercept (when x = 0), and

*a1* is the slope of the regression line.

For the Cameroon data, we have

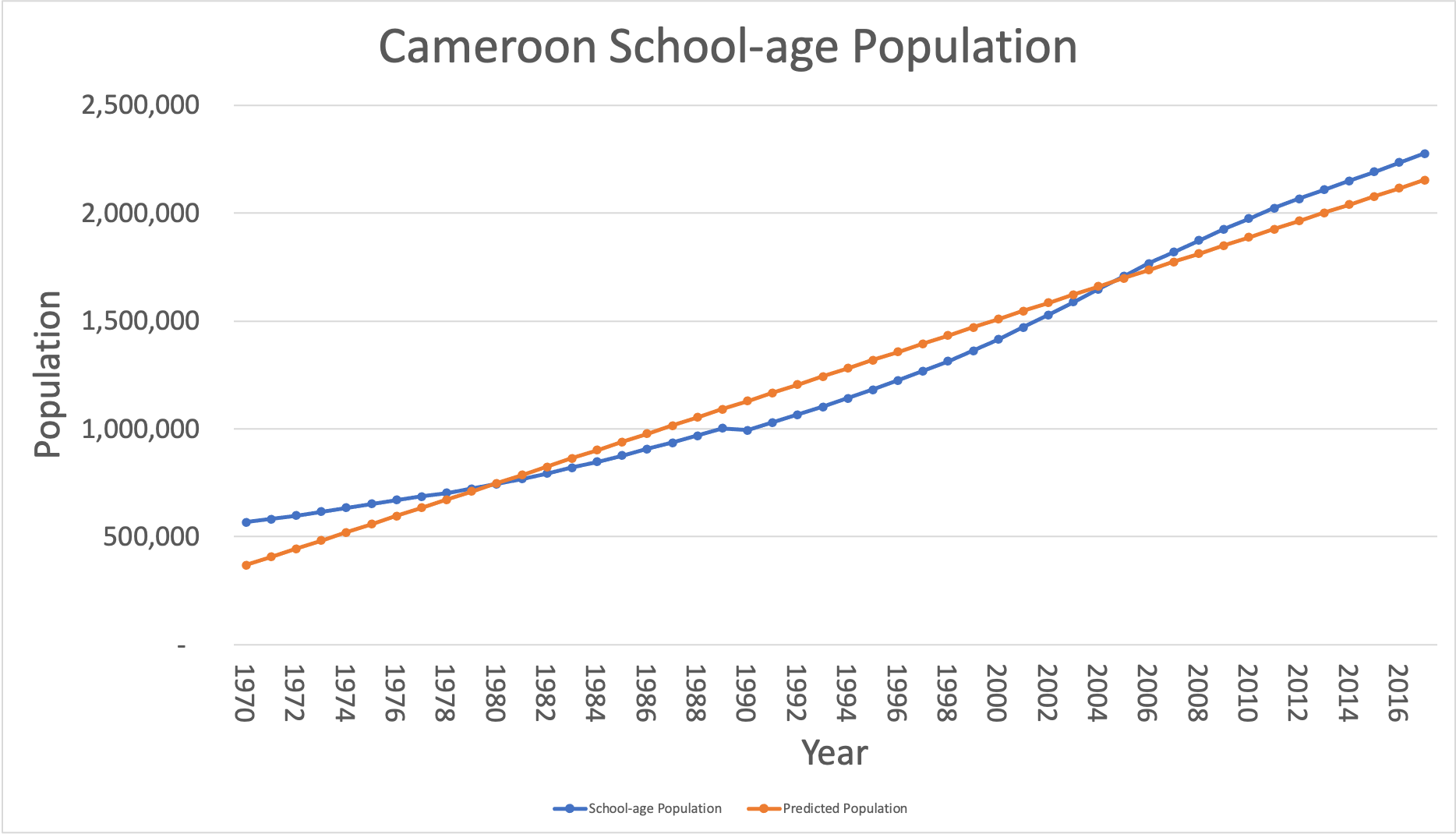
*a0* = -74,377,136.97 and

*a1* = 37,942.52

The correlation coefficient R is 0.9826, which suggests that the data are reasonably well-represented by the linear regression equation.

(R = 0 means there is no linear relationship among the data and R = 1 means that the data points all lie on a line. But beware! There is much more to determining if this equation is appropriate, as you’ve seen earlier. Nevertheless, for the purpose of illustrating the concept of using a predictive model, this equation is adequate.)

Figure 6 shows this regression line superimposed on the actual data.



*Figure 6: School-age population for Cameroon showing linear regression line*

Using the regression equation, it’s possible to predict the school-age population for any year. For example, in 2020, the predicted population is 2,266,765. For 2030, it’s 2,646,190. Keep in mind that populations generally don’t grow linearly forever. At some point, downward pressures become important, such as limitations on resources to sustain that high of a population or simply parents deciding to have fewer children. This is, of course, an example of having to apply your expertise in population dynamics to the case at hand as you interpret the model’s results.