Introduction to Business Analytics

Aram Balagyozyan

Department of Operations and Information Management
Kania School of Management
The
University of Scranton

January 27, 2022

Outline

1. Introduction to Business Analytics
2. Some Core Ideas in Data Mining

Outline

What is Business Analytics?

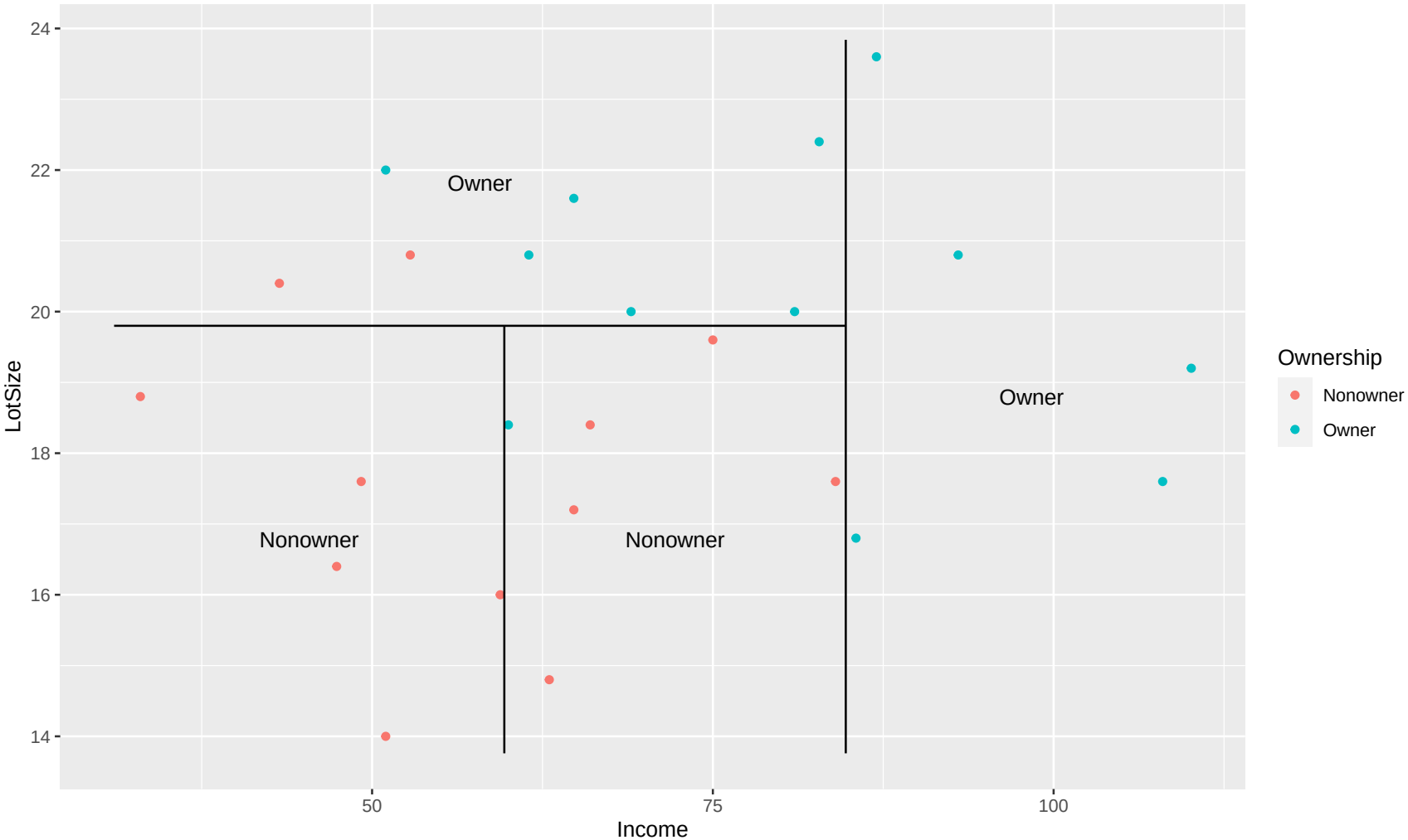
- *Business Analytics* (BA) is the practice and the art of bringing quantitative data to bear on decision making.
- Term *Business Analytics* means different things to different organizations and scenarios, ranging from simple A/B testing to customer segmentation and forecasting.
- BA includes a range of data analysis methods. Thus, this course also will cover several data/business analytics methods.
- In contrast to BA, term *Business Intelligence* (BI) refers to data visualization and reporting for understanding "what happened and what is happening." BI utilizes charts, tables, and dashboards to display, examine, and explore data. BA now includes BI as well as sophisticated data analysis methods used for exploring data, quantifying and explaining relationships between measurements, and predicting new records.

Data Mining, Machine learning, and Statistics

- *Statistical models* refer to methods that reveal global tendencies in the data. Statistics focuses on inference from a sample to the population regarding the "average effect."
- As an example, statistics may focus on the the effect of \$1 price increase on the **average** demand for a product.
- In this course, term *machine learning* (also known as *artificial intelligence*) refers to algorithms that learn directly from data, especially local patterns, often in layered or iterative fashion.
- Machine learning mostly focuses on predicting individual records. It may for example attempt to predict the effect of \$1 price increase on the demand of a specific person.
- *Data Mining* stands at the confluence of the fields of statistics and machine learning. While most methods used to explore data and build models have been around for a long time, the core tenets of classical statistics – computing is difficult and data is scarce – don't apply to data mining applications, where both data and computing power are plentiful.

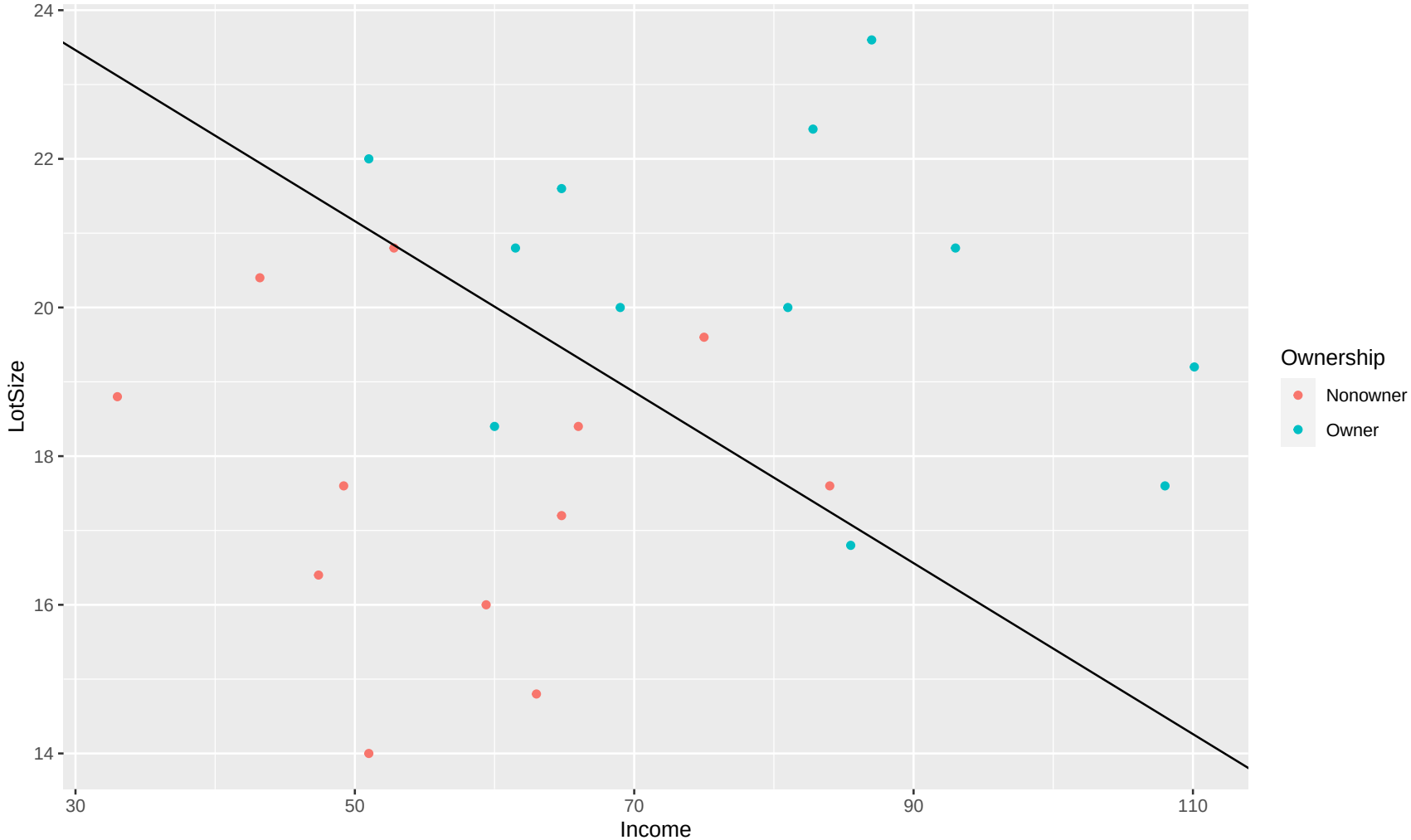
Why so Many Different Methods?

Consider an attempt to predict whether a customer will buy a lawn mower. Using data on incomes and lot sizes of previous customers we may separate buyers from non-buyers using vertical lines produced by the *regression tree* model as below:



Why so Many Different Methods?

The *logistic regression* model will result in a single diagonal line as below:



Why so Many Different Methods?

- The two models above could be used to predict whether a given new customer, with say $75,000 income and 18,000 sqr. feet lot size, will buy a lawn mower. The regression-tree model predicts a non-purchase while the logistic regression model predict a purchase.
- Different methods can lead to different results and their performance can vary. Making a determination which model is more appropriate is rather an art than science.
- It is therefore customary in data mining to apply several different methods and select the one that appears most useful for the goal at hand.
- Note also that in the regression-tree diagram, there were four terminal classes. But it is up to the analyst to choose the correct number of separation classes. If there were as many classes as data points, then the model would produce a perfect fit for the collected data, but would be miserable at predicting new data points. This problem is well known in data science as *overfitting*.

Outline

Supervised vs. Unsupervised Learning

- Consider a question: Do our customers naturally fall into different groups? Here no specific purpose or *target* has been specified for the group.
- *Unsupervised Learning* is an analysis in which one attempts to learn patterns in the data other than predicting an output value of interest. Unsupervised learning algorithms are those where there is no outcome variable to predict or classify. Hence, there is no "learning" from cases where such an outcome variable is known.
- Association rules, dimension reduction methods, ans clustering techniques are all unsupervised learning methods.

Supervised vs. Unsupervised Learning (continued)

- Consider another question: Can we find groups of customers who have high likelihood of canceling service soon after their contract expires? Here there is target, will the customer leave when her contract expires.
- *Supervised Learning* is the process of providing an algorithm with observations (records) in which the variable of interest (the variable we are trying to predict, otherwise known as target or output variable) is known and the algorithm "learns" how to predict this value with new observations where the output is unknown.
- Simple linear regression,logistic regression, regression trees, etc. are examples of a supervised learning algorithms. The Y variable is the known outcome variable and X variable is a predictor variable. After we estimate the best-fit regression line, we can use it to predict Y values for new values of X, for which we don't know the Y value.
- Supervised and unsupervised methods are often used in conjunction.