Name of the module	Reacfin Academy Data preparation and data quality

Powered by Reacfin

Contents

Scene 1: Title

Slide 1.1 Title

Slide 1.1: Introduction

In this module we explain how data is created, gathered and manipulated to be optimally used in companies. We also propose different best practices to address as efficiently as possible preparation of data with a view of creating data science projects.

Scene 2: Data source

Slide 2.1 Data source

In this module we explain how data is created, gathered and manipulated to be optimally used in companies. We also propose different best practices to address as efficiently as possible preparation of data with a view of creating data science projects.

Slide 2.2 Structured vs unstructured information

C1 In order to analyze data or information, a good practice is to first analyze where the info is coming from to ensure its value, its quality and define the best way to use it.

C2 Actually many sources exist today: web but also Internet of Things (IoT), companies' internal process, IT programs and companies' interaction with providers, clients, etc.

We can make a distinction between two main categories of information:

- C3 Internal data, on one side, represents all information that is created internally: its origin and its characteristics are usually known. It can consist of, for example, information that comes from internal reports or documents, internal website (information about visitor activity, purchases, etc.), emails, results of program calculations, etc.

- C4 External data, on the other side, represents all information that does not come from internal processes. In that situation, data characteristics are not always fully known. It can consist of information that comes from external websites, or social networks. External data can also be purchased from specialized providers or obtained within open source project such as open data government websites.

C1 Including these internal and external datasets, another important distinction must be highlighted. Some of these data are clearly identified, characterized using variables, specified through definitions, author, update frequency, etc. This is what we call C2 structured data.

Another type of info on the contrary is non-organized data, not easy to manipulate – this is what we call by opposition, C3 unstructured data

Actually, the amount of unstructured data is larger than the amount of structured data. It comes from the fact that infrastructures cannot manage the incredible amount of data that are produced nowadays and its constant generation. To give you some numbers, we estimate that today around C4 70% of critical information within insurance companies are not used, because of its format (old database, documents, papers), but also because its existence is not known.

Well, even if unstructured data requires more preparation and documentation, this kind of info has many benefits, mainly because they represent a rich complementary source of content to explore (70% as reminder).

C5 This is the new mapping of our internal and external data, considering the structured/ unstructured segmentation.

C1 Once the source of information and the type of information (structured or unstructured) have been identified, another natural question comes to mind: what kind of data are we talking about?

Actually, data has many different forms:

- C2 Number: it's basically what we usually call data, a set of numerical values stored in a table. But data is more than just numbers
- C3 It can also be a text: that is all information represented using words or characters: documents, letters, emails, books, articles, etc. Internet by nature contains a large quantity of text data. Just to give you some numbers: in 60 seconds, the web creates 156 millions of emails, more than 80 000 posts, 500 000 comments on Facebook and 1 500 new blog articles.
- C4 It can be audio data: which concerns all sound files such as calls or music. Another figure? In 60 seconds, more than 2 000 000 minutes of calls are done by skype users,
- C5 Photo and Video: it concerns all information provided through images: pictures, scans, movies, video conferences, etc. Here again, we observe an incredible quantity of content. E.g. Youtube and its 700 000 hours of video watched every 60sec.
- C6 Many other hybrid types of information exist such as computer signals, location and other specific IoT sensors, etc.

C7 To come back to our data mapping, here is what it could look like considering also nature of data. We observe that structured data within companies remain mainly "traditional" data: text data and numbers. On the opposite, unstructured data from external sources are mainly images, audio, text and other types: this comes from the fact that methodologies to manage

the understanding and use of these data are still in progress (Natural Language Processing (NLP), image recognition, Speech to text, etc.)

Slide 2.3 Structured vs unstructured information in the insurance value chain

C1 Thus, data has many available sources, interests and natures. But how can it impact concretely insurance functioning?

To understand this mechanism it's important to come back to the understanding of the insurance value chain and which data are required at each step. Insurance value chain regarding insurance product can be seen as follows:

- C2 First step consists of a risk or client's need identification. To do this, insurers need market information, but also analysis of the risk.
- C3 Then, a pricing of this risk is necessary and related marketing studies can help refine the product design. At this stage, insurer and especially actuaries need concrete data to model as accurately as possible the price of the risk. Marketing teams also need to challenge the product design with the competition or other offers internally in order to completely fit the client's needs.
- C4 Legal aspects: in parallel of tariff calculation, the legal department needs to check feasibility of such a coverage, observing jurisprudence, local law etc. and writing general conditions based on the product intended purpose and insurance firm practices
- C5 Once technical specifications of the product are ok, commercial campaign can be launched. To be optimal, this campaign requires information about the client from the web but also from insurance channels (brokers). It also needs Key Performance Indicators to measure the impact of the offer.
- C6 Regarding the underwriting stage, insurers gather data from the clients (about age, professional situation, family, etc.) but should also for example collect information about insurance proposals that did not become contracts to analyze the underlying pattern of non contractualization.
- C7 The modelling part involves a very large quantity of data: financial data, policy behavior, insurer reserve, etc. that can be used to feed the product development, customer analysis, actuarial department or risk management, to name a few.
- C8 Finally, the insurance calculation and methodology are audited according to compliance practices and regulation information.

All these steps use but also generate new internal data. This is the beginning of a long process within the insurance company to organize and manage these data.

Scene 3: Organization and management of data

Slide 3.1 Organization and management of data

Slide 3.2 Data journey

C1 Usually, once information has been collected; data begin a long process within the insurance firm before being exploited by the end user. This process is essential to ensure harmonization, exploitability, compliance and quality of datasets.

C2 In this process, we can resume data treatment in three stages: Extracting, Transforming and Loading stage (called ETL), storage stage and front-end analytics stage.

C3 Regarding the ETL stage: it mainly corresponds to the treatment phase of information. Insurance companies use ETL tools, IT languages and advanced scheduling and monitoring tools to collect information and reshape it in order to make it more relevant and useful for the company. Details on this stage will be discussed later.

C4 Regarding the storage stage: Data warehouses are set up to make the inventory of data, on a precise period according to strict rules. It allows creating a data model that represents the general structure of data in the company. Data mart simply corresponds to a subpart of these data warehouses and allows organizing information for a specific purpose or team. There are different ways to organize and store information (data lake, Online analytical processing, OLAP for example), and different technologies (Hadoop, Spark). We will further discuss this later.

C5 Front-end analytics stage: Finally, information is called through a query that simply corresponds to a request of specific characteristics of data. Then, the results are treated and printed within a reporting table. Thus, information can be used at the discretion of the business or technical teams to make dashboards that will allow to visualize dynamically or manipulate information, or to build algorithms for predictive analytics.

C6 At the end, the user can finally access data and related analysis for its own purpose.

Slide 3.3 ETL

C1 Let's go back on the ETL part. As discussed just before, it corresponds to extraction, transformation and loading of data. More specifically, it means the following:

C2 Extracting information consists of reading data from the sources mentioned previously and retrieving all different formats (e.g. XML, JSON, Flat files) to convert them into a single appropriate one.

C3 Transforming data aims to convert data according to the requested structure. Generally, business teams, in agreement with IT teams define such structures. Concretely, tools apply many

rules on data to select columns, translate values, sort information, join or split tables, create new variables based on simple calculation, perform quality checks, etc.

C4 Load: the goal of this stage is to move and load data in a new location such as a data warehouse or a dedicated data mart in order to "publish" them. That is usually performed on a regular basis, so that tools can keep a history of data and allow data audit trail (in order to follow data consistency over time)

C5 Many tools exist to address this data management stage. The most famous ETL tools are: Cognos, Informatica, IBM, DataStage, Microsoft SQL Server, Oracle Data Integrator, Pentaho, SAP Data Services, SAS,…

Slide 3.4 Storage details

C1 Data storage is a broad but critical topic for companies: it's basically the heart of the data management mechanism and more generally of a data driven company (its DNA). It's a place where all information are gathered, and structured according to ETL rules and outputs needs. One of the classical approaches regarding data storage is the enterprise data warehouse which is a system we can describe through the following criteria:

- C2 Unique Data model reference which will represent how data are organized within the structure
- C3 Transversal content subjects that represent the main company activities
- C4 Frequency and history are clearly detailed in order to keep everything "in mind" and updated on a regular basis
- C5 A very large memory space is required to set up such a project

C6 Data marts are also an interesting structure to mention. They are smaller than data warehouses and mainly focus on a unique dedicated area such as finance system, HR databases, etc. Memory is usually limited to a specific size, and contents are a subpart of data warehouse systems. The idea behind this is to address immediate data needs of business teams.

C7 Another storage method recently appeared with the emergence of big data. Data lakes are somewhat different from data warehouses as the main principle of data lakes is to entirely retain data from sources, conserving also their original shape including unstructured data.

More precisely, differences between data warehouses and data lakes are the following:

- C8 Structured data on one side against many types of data for data lakes
- C9 Fixed process and storage against agile methodology for data lakes
- C10 IT specialists to manage the system against data scientists teams for data lakes

Slide 3.5 Front-end analytics

C1 Last stage of the data process for the company is to let the organized information be used by the end user. Depending on the degree of expertise, its use of data and its goals, final front-end analytics can take many aspects.

C2 Query and reporting will be performed to address a simple question or to answer a simple request. Basically it's "what information can I obtain?". A basic response to this is to set up an extract of the data warehouse or the data mart. For example, we could make the request to extract all insurance policies linked to the life insurance business line where the policyholder is more than 60 years old to assess the proportion of old clients in the portfolio. A more advanced answer is to develop dedicated reporting that could present the information including additional statistics or results from computation.

C3 End user could also ask himself, "what can I do with this data, how can I analyze it?" Analytics and data mining can help you find out. In this case, the final user will be able to make statistics based on extracted information, but also develop advanced analytics to create predictive models. For example, user could develop a machine learning algorithm to predict lapse rates of the saving policies extracted in the first stage.

C4 A last aspect is to answer a larger question: "why do I observe this effect?" or "why are my results like that?". The answer could be suggested through data visualization which is basically composed of charts that represent data and that could provide many points of view to understand them. A complementary solution is to develop an interactive dashboard that allows users to visualize info dynamically, and thus to "play" with data and test them in order to solve their business problem. For example, we can observe on a dynamic map the importance of the broker activity and deeper study what can influence profit or losses in a dedicated region.

Scene 4: Data preparation and quality checks

Slide 4.1 Data preparation and quality checks

Slide 4.2 Successive steps

C1 Within the process presented in part 2, one critical element requires specific attention: data preparation and data quality. Preparing data is an important stage in order to ensure results' accuracy. Data preparation consists of applying a set of actions on data in order to improve it. Many categories of actions exist:

- C2 Missing values: it mainly consists of identifying data not completed and finding a way to fulfill information in a smart way
- C3 Format and types: this action aims to reach homogenous format and types for each variable in order to avoid specific manipulation for specific observations of the variable
- C4 Abnormal treatment: it consists of observing if there is too large or too small values in the database compared to what's expected (e.g. an age of 140)

- C5 Redundant variables treatment consists of hiding or deleting information that is not relevant for the analysis (for example observations that appear twice)
- C6 Many other categories of treatment exist: e.g. feature engineering which is a very useful way to increase data base focus for data science studies.

Slide 4.3 Data quality process

C1 As we have seen earlier in this module, working with quality data is very important when applying data-driven modelling techniques. We will now have a more precise look at practical elements of the data quality process.

A quality "check and fix" process should always be applied to data, from wherever it comes. This diagram presents such a process. C2 After the collection and organization of the data, a series of tests should be performed, C3 each one concerning a different aspect of data quality. If the database passes the test, C4 the modeler can go to the next test. If it is not the case, he has to take some action to "fix" the diagnosed issue, C5 and then go to the next test. Actions can be to modify the table or to simplify it.

C6 This process seems quite simple from a general perspective but it's actually the most time-consuming activity within a data science project as usually tests are not ok from the first try.

In the next slides we present in detail different types of tests that can be performed as described in the introduction.

Slide 4.4 Databases

C1 But before, let's give a short definition and formalization regarding database structure. It is convenient to think about the database as a large table C1. Each line represents one data record (or observation) while each column represents one variable (or characteristic).

C2 For example, let us consider a motor-insurance database, which we will follow throughout this module. In this case, each line represents one policy (one contract) whose characteristics are given in the columns.

C3 Each column represents one characteristic, which is normally given for each policy. For example we could define in column the contract start date, driver's name and age, bonus-malus, etc.

Slide 4.5 Filters

Once data structure is defined it's also important to understand how quality checks are applied.

C1 The data quality actions transform the database and can be understood as data tables that will go through the filters one by one. Some of these filters will shrink data into smaller sub-databases, by deleting unnecessary lines from data.

C2 Other filters do the same with columns, decreasing the number of variables that can be considered.

C3 Finally, filters can also act on each cell of the table (not only columns or rows) by just modifying their value. It means that, in this situation, it will keep initial dimensions of dataset. We are now going to look at these filters in details.

Slide 4.6 Missing values

C1 As we have seen previously, the process of producing and extracting data is very rarely straightforward, easy and fully reliable. For this reason, databases often present "holes", generally referred to as "missing" or "not available" (NA) data, here represented as red cells.

Many tools let the user count and visualize the number of missing values in databases, in a similar way it is shown here. Sometimes these tools also consider incomplete values (missing digits) as missing values.

C2 In our sample database, we can for instance remark that some information about a particular group of policies is missing. Here, we see that the "leasing" information is not given for two specific policies. Many reasons could explain these breaches that should be further analyzed if recurrent.

C3 When the missing data are sparse, they can be somehow ignored. The techniques to be applied generally handle rare missing data rather well.

C4 However, when a large number of cells in the same row or in the same column are missing, the situation is more problematic. In this case, the corresponding rows or columns may be deleted. A threshold of missing values is generally set: if the number of missing values in a row or a column exceeds this threshold, it is removed from the database.

C5 Other more complex alternatives exist when missing values are numerous in order not to delete information. It aims to use data analytics to replace missing values per approximation. One basic approach is to replace information by the average value of the column if the column

is numerical or the most frequent label if it is a categorical variable. Nevertheless, these corrections should be avoided as much as possible as it can distort the conclusions of the analysis.

Slide 4.7 Format and type

C1 Another important element of data quality concerns the format in which data are provided, which corresponds to the way the value is printed. Many problems exist regarding format aspects. An important one is the date format. In our sample database, a column gives the starting date of the policies. However, all the dates are not given in the exact same format. In order to use them, it is necessary to standardize them. This is a very commonly seen problem, especially for companies operating in different countries and regions.

C2 In addition, a special attention should be paid to type mixture: it is critical to verify that variable types are consistent, or in other words, that the values stored in one column are all of the same type. The most frequent types are the following:

- Some variables are logical (also called Boolean), which means that it is either TRUE or FALSE (also denoted by 1 or 0). For example, we could indicate in a variable named "leasing" if the insured car is leased or not.
- Some variables are given as integer numbers, as for example the power of the car.
- Some variables are given as decimal numbers, as for example the premium paid by the policyholder.
- Some variables are given as intervals, as for example the driver's age.
- Some variables are given as unstructured strings, as for example the driver's name.
- Some variables are given as structured strings, as for example the starting date of the contract.

It is therefore very important to pay attention to the variable types. Imagine for example the issue that could be created when treating equally all values of the age column while they are stored in several different formats: integer numbers, decimal numbers, intervals (age classes), strings.

Most of tools give a warning when a format error is detected. They generally replace ill-formatted values with a specific "bad format flag", for example replacing them by NA.

Slide 4.8 Abnormal values

C1 In the same vein, it is important to perform a check for abnormal values in the considered database. Some variables of the considered database may indeed present values that do not make any sense. In our sample database, three such values are given:

- According to the starting date column, one policy dates back to January the first 2016, which is impossible for our 1-year contracts considered in 2018.
- One of the drivers is recorded as being 14 years old, which is obviously incorrect.
- The power of one of the cars is given as a negative integer number.

This data quality issue does not seem very different from the previous one. It however is way more difficult to detect and fix, as the "abnormality" is very often an arbitrary human-based constraint that cannot be discovered by a computer if precise instructions are not given to it. For example, how could the computer know that an age below, say, 16 is a problem, if we have not explicitly instructed it to do so?

C2 One way to detect this kind of issues is to carefully inspect each column using data visualization tools. For example, we could produce histograms for each columns, which allows gaining an insight of the variable distribution, and thus to the values it takes. Standard statistical tools also help detecting abnormal values. For example, studying the variability of the numerical variables allows detecting extreme values or outliers, i.e. values which are way larger or smaller than the other ones.

Slide 4.9 Redundant values

C1 Many real-life databases come with a large number of variables/columns. The more columns a database has, the slower are the processes using this database. Manipulating and modelling huge amount of data can be a real challenge from the point of view of time and machine resources. For this reason, it is sometimes easier to reduce the number of variables of the database, a process which is usually called "dimension reduction" in data science. As we are interested in prediction, we want to perform this dimension reduction with the minimal loss of explanatory power for our model.

C2 It seems therefore natural to search for the most "redundant" columns of our database. This redundancy can be expressed through different statistical concepts, among which we find correlation. This metrics spans from -1 to 1 and measures the dependence between two variables:

- When the correlation is close to 1, the variables are very "similar" in the sense that they channel very similar information. They are therefore someway redundant and one of them can be removed from the database.
- When the correlation is close to -1, the variables are very "dissimilar" in the sense that they channel opposite information. They are therefore someway redundant and one of them can be removed from the database.
- When the correlation is far from 1 and -1 (for example when it is close to 0), the two variables are not sufficiently dependent to remove one of them from the database.

C3 Of course, a high dependence between variables can also be detected visually, for example in a scatter plot. In such a graph, each dot represents a record of the database: its position on the horizontal axis gives the value of variable A and its position on the vertical axis gives the value of variable B. If we observe a linear trend in the scatter plot, i.e. if the data points form approximatively a straight line, a high (positive or negative) correlation exists between the variables.

In this example, the record highlighted in blue corresponds to a car with power equal to 90 and fuel consumption equal to 1.6.

The dots somehow form a straight line: it seems therefore pretty clear that power and fuel consumption are positively correlated variables. When we know one of them, we know approximately the other one with a rather small error.

C4 On the contrary, here is a plot of the power of car versus its age. As we only see here a cloud of points, there is no linear trend in the scatter plot, so that we can conclude that the correlation between car power and its age is not material, a result which of course confirms our intuition.

Scene 5: Conclusion

Slide 5.1 Conclusion

Thus Data preparation is a long journey that begins with the identification of numerous and heterogeneous data sources. We have also noticed that data can take many forms and is essential at each step of the insurance value chain. Once data interest is identified, another process starts within the company: the data management. It begins with the extraction and the transformation of these data; the information is then stored in different places according to the business purpose. And finally information is used by individuals according to its purpose and expertise. One fundamental aspect has been highlighted in the data management process: data preparation. It can be represented as a filter that improves data quality. Filters can have many influences: to identify or delete missing values, to verify coherence and accuracy of data, study format and type or to point out redundant and non-relevant variables.