Case Study: Estimating Demand for Bananas

Mathias Goller

December 27, 2007

1 Introduction

Daily Markets is a Wholesales Company specialised on selling fresh fruit. Especially, it sells fruits from overseas and exotic fruits to small supermarkets.

Dealing with fresh fruit might be very profitable if one can minimise the loss caused by fruit that became too old for sale. Therefore, determining the demand for a specific fruit properly is a success factor in that kind of business.

These supermarkets have rented some of their shelf space to Daily Markets where it can place its products. The supermarkets exchange the risky fruit segment by a steady income.

In this case study, you assume the role of a business analyst in a meeting concerning the data analysis project initialised a month ago.

Estimating the demand for specific fruits as exactly as possible is the goal of the data analysis project.

The first meeting is about establishing a process that estimates the demand of bananas.

Daily Markets buys bananas when they are still green and lets them grow ripe in greenhouses. As the growth process can only be delayed by one or two days once the ripening process has started, it is necessary to predict the demand two weeks in advance.

Prepare yourself for the meeting. The meeting will be organised by the manager of the buying department. Besides members of your team there will be the manager of the greenhouse division.

2 Prepare Data Understanding and Preprocessing

The slot of the meeting you shall fill is about data understanding and data preprocessing. To be more specific, analyse if all necessary data sources are available in sufficient quality.

If you cannot do this task with the information given in that document specify which steps are necessary to test quality.

Additionally, define how to preprocess data for data mining.

Prepare your answer in a written hand-out. Using the CRISP-DM process model might be helpful for your answer.

3 Available Data Sources

Daily Markets regularly receives sales data from the supermarkets in which Daily Markets sells its products. However, some supermarkets send sales per days while others send them per week.

Additionally, Daily Markets stores data about each delivery of fruit into a market in its data warehouse. Each time new fruits are arriving, old fruits are thrown away. Thrown-away fruits are counted and their number or weight, depending on the product, is stored in the data warehouse, too.

Each month, it is taking stock in each supermarket to detect irregular losses such as theft or dehydration.

sales(**market, product, day_from, day_to, sum_euro, quantity, kg_piece**)[1]

delivery(**market, product, day, quantity, kg_piece**)

away(**market, product, day, quantity, kg_piece**)

Case Study: Estimating Demand for Bananas - Daily Markets

Mario Hofer
0255509

Abstract

This paper gives a rough overview on data preprocessing applied to the estimation of the demand for bananas at Daily Markets. To be more specific, the paper will highlight how data mining can be used to test data quality and which steps for data preprocessing are necessary. All conclusions drawn in this document are based on the description of the case study. The paper uses the CRISP-DM process in order to structure the analysis.

1 Introduction

In order to maximise profits Daily Markets needs to estimate the demand for bananas as exactly as possible. Nevertheless the main purpose of this paper is to show how Daily Markets can test for data quality and how they can structure their data preprocessing and not the demand estimation itself. Therefore the analysis starts with a rough overiew on whether the given data sources provide enough information to give an exact estimation for the demand for bananas. After this there will be a section on how to measure data quality. In order to measure data quality the section will describe different dimensions of data quality and how erroneous data will affect the different dimension. Based upon data quality and data anomalies the proximate section will explain methods for data preprocessing. The final section concludes. As mentioned before, the whole analysis is structured along the CRISP-DM process. Since this paper deals with data quality and data preprocessing, it will focus only on the sections *"Data Understanding"* and *"Data Preparation"* of CRISP-DM.

2 Data Sources

The only sources of data highlighted in the handout are the data warehouse relations *sales*, *delivery* and *away*. A simple way to estimate the demand for bananas would be a linear regression [1]. Although these relations and their attributes show the sales for bananas for a certain market for a certain period of time, the do not give any detailed information on factors influencing the sales. Applied to a regression model this would mean that the relations only provide information for one right-hand-side variable of the model, the price. Nevertheless the demand for bananas is also influenced by the price of its substitutes, e.g. the price apples.

Furthermore it seems reasonable to assume that the demand for bananas depends on other variables like for example the type of the mall or the demographics of the area surrounding the market as well. However for a short period of observation, like it is true for this case, it should be safe to assume that these factors have an influence which is constant over time. Therefore these factors will be ommited from further discussions.

3 Data Understanding

According to the CRISP DM process, in order to verify the data quality one has to collect, describe and explore the data. These activities are subsumed under the phase *"Data Understanding"* in CRISP DM.

Describe data

The case asignment is not clear whether it is still possible to collect further data, like for example the price of substitutes. For this reason it is assumed that the collection of data has been finished and therefore there will be no further discussion on data collection. The next step in the phase *"Data Understanding"* is the description of the data. The assignment already gives an overview of the structure of the data, namely the three relations and their attributes. Furthermore Daily Markets should also:

- Understand the meaning of each attribute: For example, are market, product and the two dates "meaningful" values or are they just the keys? What is the meaning of *kg_piece* in business terms? Does this attribute represent the weight in kg of a single banana or a bundle or even a

[1] cf. Davidson and McKinnon (2004) pp. 86 or Wooldridge (2003) pp. 68

whole box? Does this attribute have exactly the same meaning in all of the three relations?[2]

- Check the format of the data: What are the data types of the different fields? Is quantity given as an integer or a floating point number? What are the ranges on the different attributes?

- Check whethere the given attributes are relevant for the estimation of the demand for bananas: Do we really need *sum_euro* for the estimation? As explained later in this paper *sum_euro* may be usefull to test for data quality. The same is true for the other attributes, so therefore it is assumed that all attributes presented in the assigment are relevant for the data mining task.

- Check the quantity of the data: How many records does each of the three relations contain? What is the size of the data which has to be mined in mega- or gigabytes? How long has the data been collected for? Do all of the three relations cover the same time span?

- compute basic statistics (distribution, max, min, standard deviation, variance, etc.) for the attributes *sum_euro*, *quantity* and *kg_piece*. This helps to get a first impression of the data and it is already the first step to identify erronous data, e.g. the max/min values of either of the three attributes listed above can find outliers indicating irregularities or constraint violations (negative quantity etc.).

The findings of this step are outlined in the *Data Description Report*.

Explore Data

After the data has been described one has to get accustomed to the data. According to CRISP DM *"this task tackles the data minging questions, which can be addressed using querying, visualization and reporting"*[3]. Therefore the analysts of Daily Markets could visualize the demand for bananas accross time. This helps to identify whether the demand for bananas is subject to a linear or non-linear time trend or whether there are seasonal variations in the demand (which seems to be reasonable).

[2]This since *kg_piece* is not clearly defined in the assignment, it is assumend that it represents the weight in kg of a typical bundle of bananas and that this definition can be applied to all three relations.

[3]see [2] p. 21.

Furthermore [2] suggests to look at the distribution of the key attributes. Therefore Daily Markets should visualize the distribution of the attribute *quantity* in order to see for example whether there is a high fluctuation in demand for bananas. Furthermore Daily Markets should compute simple derived attributes like the price per kg bananas, since it is likely that this is the most important variable when estimating the demand for bananas. The price per kg can be computed by $price = \frac{sum\_euro}{quantity * kg\_piece}$.

Verify Data Quality

The next step suggested by CRISP DM is to verify the data quality. During this step among other things the following questions should be answered:

- Is the data complete? e.g. has all data on sales been recorded?

- Does the data contain errors? How common are errors? e.g. do the sales quantities exhibit negative or unrealistic high values?

- Are there missing values? How common are missing values? Do they have a specific meaning? e.g. are there tuples where the quantity is missing?

In order to measure data quality, it has to be defined first. This paper uses the dimensions of data quality as explained by [4], who defines data quality as an aggregated set of quality criteria.

Below is the list of quality criterie as introduced by [4] applied to the case of Daily Markets:

- Accuracy: Aggregated value of Integrity, Consistency and Density
  - Integrity: Aggregated value of Completness and Validity
    - Completness: Quotient of entities from the mini-world (M) represented by a relation (r) and total number of entities in M. Applied to the case completness would be e.g. the ratio of recorded sales and total sales.
    - Validity: The percentage of tuples in r representing valid entities from M. In the case of Daily Markets validity would suffer from for example typing errors in the quantity.
  - Consistency: Aggregated value of Schema Conformance and Uniformity

    - Schema Conformance: percentage of tuples in r conforming to the schema structure (e.g. domain). Because of the limited information on data sources, this point will be left out from discussion.
    - Uniformity: percentage of tuples in r which do not contain irregulariets[4]. Relating to the case, uniformity would suffer from misspelling in the name of a market or if e.g. Daily Markets is an international company and did not account for different currencies when merging the sales data.
  - Density: Quotient of missing values in tuples of r and the number of total values (that are expected to be known). If there are 100,000 records in the sales relation then there should also be 100,000 values for e.g. quantity.

- Uniqueness: Quotient of tuples the same entity in M and total number of tuples in r.

In order to test for data quality Daily Markets has to measure above described quality criteria. The criteria cannot be measured directly but rather indirectly by analysing different data anomalies which influence the performance of the different criteria. In order not to go beyond the scope of this case study, there will only be a short presentation of data anomalies and even a shorter presentation on methods which can be used to detect them.

This part descibes in accordance to [4] and [5] the different anomalies which influence data quality. For each anomaly there will be a short explanation and a list of methods on how to detect and/or eliminate the anomaly. An explanation and introduction into the methods would be beyond the scope of this case study, therefore only references will be provided.

Lexical Errors

When data is transferred to the data warehouse it may be that sometimes values are not available. This could result into attributes which are note completely defined. e.g. if the sales data from the different markets is retrieved via text files, it may be that a *sum_euro* is missing leading to a vialotion of the expected number of values for a sales record. Methods to detect and/or eliminate this error a the same as for domain format errors.

Domain Format Errors

Such an error occurs if a value violates the domain format specified for an attribute. In the Daily Markets sales table it could the case that the required format for *day_from* is *DD.MM.YY* but suddenly a market delivers its data in the format *YYYY-MM-DD*. In order to detect and also clean those errors [6] suggests:

- Format & domain transformation
- Standardisation
- Normalisation
- Dictionary look-up

Irregularities

This anomaly often arises from the process of merging data from multiple sources, e.g. if a multinational enterprise uses operational data from local stores to fill a data warehouse. It may be the case that there are shops in Austria where the local prices are denoted in Euros, whereas the Japanese stores denote the prices in Yen. Since the assignement does not specify whether Daily Markets is an international company it could very well be that this anomaly is relevant. Another example of an irregularity would be if there are some markets where *sum_euro* includes VAT and some markets do not include VAT. This would lead to an error in the derived price for bananas, which then in turn would bias the estimated demand.

To detect this kind of error one could:

- carry out an outlier analysis[5] or
- carry out an regression before merging the data from different markets. One could e.g. regress the price of one market on the price of an other market. The coefficient of the price gives an indicator whether one market includes the VAT and the other does not.

Integrity Constraint Violations

[3] define integrity constraints as *"a predicate or query such that if the predicate holds on a state of the data, or equivalently if the query produces an empty answer, then the database is considered valid"*. Applied to the

[4] [4] defines irregularities as *"the non-uniform use of values, units and abbreviations"*.

[5] as described in e.g. [1]

case study it could be that there is a negative quantity or that *day_from* greater than *day_to*. As stated in the assignment the data comes from a data warehouse, therefore I assume that integrity is enforced by constraints specified in the data warehouse.

Duplicates

Duplicates deal with the problem that two different tuples represent the same real world entity. For Daily Markets this would occur if a market changes its name or sends its sales data using a different abbreviation for its name. e.g. Groceries&Co Inc. may suddenly transmit its sales under G&C Inc.

The data warehouse of Daily Markets only consists of a small number of tables with only a few attributes. Therefore detection methods which are based on the comparison of individual attributes of tuples should suffice. Such method would be for example Q-grams as explained by [7].

Invalid Tuples

An invalid tuple is a tuple where there is no corresponding entity in the real world, e.g. a recorded sale, which just never happened. So far I have not found any specific methods to detect or clean this kind of error.

Missing Value

Missing values occur if there is no value for one or more attributes of a tuple, like a sales record. [1] suggest the following methods to deal with this problem:

- Ignore the tuple.
- Fill in the missing value manually, which is very time consuming.
- Use a global constant to fill in the missing value, which could lead to useless mining results.
- Use the attribute mean to fill in the missing value, e.g. the average quantity sold.
- Use the attribute mean for all samples belonging to the same class as the given tuple, e.g. the average quantity sold per market.
- Use the most probable value to fill in the missing value, e.g. by using a linear regression.

Missing Tuples

Especiall in a data warehouse where the data is merged from multiple sources it can occur that tuples get lost because of incorrect join-operations. Therefore Daily Markets should check whether the join-operations are correct.

Further steps for Daily Markets

In addition to the methods suggested above Daily Markets should also check for inter relation constraints which might not be captured in the data warehouse. e.g. the equation $delivery = sales + away + irregularlosses$ should hold roughly for a chosen period of time. Furthermore should the price, as already defined befor, be reasonable. When looking at the basic statistics of the attributes, Daily Markets should also query the sum of sales per market for a specific time in order to carry out an ABC-analysis. It could be that from the profits point of view it is beneficial to look at larger markets in greater detail, because they may account for a large part of total revenues.

4 Data preparation

After the *"Data Understanding"* phase has been completed the next phase, *"Data Preparation"*, starts. But there is no strict separation between the phases. Since one can observe from the last section, the methods which are used to verify the data quality are also applied for data cleaning, which, according to the CRISP DM, is part of *"Data Preparation"*. This section highlights the different steps of the *"Data Preparation"* phase. When it comes to preprocessing [1] lists the following activities:

- Data cleaning: remove noise and correct inconsistencies
- Data integration: merge data from multiple sources
- Data transformation: e.g. normalization
- Data reduction: aggregation and/or deletion of tuples

The data preprocessing process described by [1] cannot be matched 1:1 into the CRISP DM reference model. Therefore I will try to allocate the different activies from above to the different CRISP DM steps.

Select data

After the quality has been verified, Daily Market has to select the data which it wants to use for data mining. If the verfication process shows, that the data

on sales and delivery is complete and exhibits a satisfying data quality, one can drop the relation *away* because it represents redundant information (under the assumption that irregular losses are neglectable). The analyists of Daily Market also have to decide whether they want to use the whole dataset or only a subset, e.g. it is highly likely that the sales of bananas 10 years ago do not really have an impact on current demand. Especially for such a long time period, dropping old tuples can improve the performance a lot. Furthermore one has to check whether the three tables cover the same period of time. e.g. if there is no sales data for the last week, then data on losses for the last week will not be relevant for estimating the demand.

Above paragraph suggests that *Data reduction* fits to this step. [1] describe a large list of data reduction techniques. In order not to go beyond the scope of this paper the only method presented in here is sampling, which is used to represent a large set of data by a smaller random sample. This approach seems to be reasonable for Daily Markets because it is highly likely that alle of the three relations contain a large amount of data. [1] explains sampling by defining D as a large dataset containing N tuples and then they suggest the following sampling possibilities:

- Simple random sample without replacement (SRSWOR) of size n: this is achieved by randomly drawing n_iN tuples from D, where each tuple has a chance of *1/N* to be drawn.
- Simple random sample with replacement (SRSWOR) of size n: Similar to above sample, but the drawn tuple is not removed from D, i.e. a tuple can be recorded more than once.
- Cluster sample: D is split into M mutually disjoint clusters. After that a random sample of m_iM clusters is drawn.
- Stratified sample: D is split into mutually disjoint parts which are called *strata*. Then a sample is created by drawing a random sample from each strata. This way of sampling can reduce skewness. Daily Markets could use the attribute market to build the different strata (though this may not be reasonable, because the skewness in this case may be advantegous for Daily Markets).

Clean data

This step tries tu assure that the data quality goals required by the selected analysis techniques (in this case presumably regression analysis) are met. The activities in this step are basically the same as the ones described by [1]. Main aim of this step is to find and remove noisy and inconsistent data. In addition to the methods which have already been presented in section *Verify Data Quality* [1] suggest *binning* to smooth data and remove noise, which is an adequate method for Daily Markets for the price variable, which has been defined as $price = \frac{sum\_euro}{quantity * kg\_piece}$[6].

Construct data

This step is similar to the process of Data transformation as described by [1]. The main aim of this step is to construct the necessary data, like e.g. derive the price from quantity, sum_euro and kg_piece. Furthermore since some markets report their sales on a daily and some on a weekly basis the data has to be brought on equal footing, meaning that the daily data might be aggregated to the basis of 1 or 2 weeks (because like mentioned in the assignment, Daily Markets wants to estimate the demand 2 weeks in advance). For the demand estimation it might be useful to have price elasticities rather than absolute values, i.e. the analysts might want to know the percentage change in demand if the price changes by x%. In that case one has to calculate the logarithmic values of the prices and the quanitities sold.

Another method which might be useful if the estimation is carried out in a single regression is *Dichotomising*. Sinc the attribute market is a nominal attribute, it cannot be included in a regression. Therefore the process of dichotomising maps the n-markets into *n-1*[7] dummy variables, each representing a single market.

Integrate data

Data integration is necessary because the data comes from multiple tables. Nevertheless if regression is used for the estimation of the demand, the most important table will be the sales relation, because this is the only relation containing data on demand. Neglecting the irregular losses, the relation away and delivery table represent redundant information because $sale = delivery - away$. If the data still is needen for the estimation, Daily Markets has to be carful with the matching process to avoid duplicate records.

[6] Since the suggested method for demand estimation is a regression it is not recommended to used *quantity* and *sum_euro* in the same model, becaus there will be a high correlation between those two variables, leading to very high variances of the estimated coefficients. For a more detailed discussion on multicolinearity see [8]

[7] *n/1* variables, because one market is considered to be the base market, see. [8] Chapter 7

Format data

According to the CRISP DM process this step refers to primarily syntactic modifications which might be required by the modeling tool. Sincer the assignment gives no further information on the modeling tool, any further discussion on this step is left out.

5 Conclusion

This paper has shown how Daily Markets can proceed according to the CRISP DM process in order to, firstly test the quality of the data sources and secondly describe how the data can be preprocessed. Large parts of the discussion in this paper are based on the assumption that the demand for bananas is estimated by building a regression model. Keeping this in mind, the paper has outlined how data quality can be defined and by which factors data quality is influenced. Furthermore the paper has given information on how to detect and remove data anomalies in the case of Daily Markets. Finally there has been a section on data preprocessing showing the preprocessing steps according to the CRISP DM. These steps have been brought in alignment to the preprocessing chapter of [1]. Lastly it has to be mentioned, that this paper only covers the preliminary steps for the estimation process based on the limited information given in the assignment of the case.

References

[1] J. Han and M. Kamber, *Data Mining - Concept and Techniques,* Morgan Kaufmann, 2001.

[2] P. Chapman et al., *CRISP-DM 1.0*, 2000.

[3] A. Gupta, Y. Sagiv, J. D. Ullman, and J. Widom, "Efficient and complete tests for database integrity constraint checking," *Principles and Practice of Constraint Programming*, pages 173-180, 1994.

[4] H. Mller, J.C. Freytag, "Problems, Methods, and Challenges in Comprehensive Data Cleansing," *Technical Report HUB-IB-164*, 2003.

[5] E. Rahm, H.H. Do, "Data Cleaning: Problems and Current Aproaches," *IEEE Technical Bulletin on Data Engineering*, 2000.

[6] K. Rother, H. Mller, S. Trissl, I. Koch, T. Steinke, R. Preissner, C. Frmmel, U. Leser, "COLUMBA: Multidimensional Data Integration of Protein Annotations," *Data Integration in the Life Sciences*, pp. 156-171, 2004.

[7] E. Ukkonen, "Approximate string matching with q-grams and maximal matches," *Theoretical Computer Science*, Vol. 92(1), pages 191-211, 1992.

[8] J. Wooldridge, *Introductory Econom*trics - A Modern Approach,* Thomson South-Western, 2003.