THE ANALYSIS OF THE PERFORMANCE OF GLS-ITALY IN BRESCIA (ITALY)

INTRODUCTION

With the **growth of the e-commerce market** and the policies of expansion of companies like Amazon or E-bay whose aim over the years has been to ease the process of acquisition of goods using the Web, the sector of transportation has seen an important improvement all over the world. From this point of view, we are assisting to an era where people can easily buy online with a simple click setting in motion a process where e-companies relives on trucking companies to get their goods delivered to their clients.

From this point of view, **Gls** represents one of the main European trucking companies and as such it has **benefited of the trend of the industry** to increase his activity.

However, logistic companies which operate on road have to face a very **complicated environment** of action and their activity depends on a series of variable that cannot always been predicted and often are very hard to handle.

From this point of view, a series of factors like traffic or weather conditions can greatly influence the performance. Also, the area of delivery with its morphological characteristics (or simply the presence or lack of adequate roads or parkings) is vital for the reason that a place can be easy or hard to reach and it is not always granted to find the right conditions to deliver into. Moreover, the decisions of a driver regarding which route to do and how intense to work and the decisions of the warehouse workers related to the number of packages to assign to each driver that specific day are vital for an optimised activity. It needs to not be forgotten that the clients also play an important role in the success of the operation. In fact, the absence of the client during the phase of delivery can cause damages to the performance of the driver and to the company that needs to organise the delivery the following days.

In other words, the industry of transportation represents a very complicated area where I believe that a lot can be done in the automatisation of some mechanisms such as the choice of the road to do, the number of packages to load every morning on a specific van, and the assignment of the right area for each driver. These elements can be optimised taking advantage of the increasing of technologies related to big data. The goal of this analysis is the one to try to analyse which situations a trucking company has to face and how its activity can be improved.

GLS factory of Brescia has kindly given me the possibility to perform an analysis with their data in order to deepen the overview of this sector. In this context, privacy rules impose me to not share the data, which remain not public.

Before to start with the description it is necessary to remind the reader that this analysis represent my first approach on the world of data analysis.

In particular, experience will be made by doing. The goal is to be able to perform a future analysis on this company in the next future where both me and the

company will have the possibility to previously set what is needed to obtain results that can truly make the difference on the activity of analysis.

DATA AVAILABLE FOR THE ANALYSIS

The dataset is a collection of over **150,000 data** granted by the GLS **field office of Brescia**. In particular, the data are the record of the delivery operations of the entire month of March 2018 in the city of Brescia and its suburban area.
In this context the activity of delivery of goods by road represents a perfect environment to obtain a various and complete dataset. On the other side, we live in a moment of transition where companies are step by step understanding the necessity to obtain clear and solid data to analyse and optimise their activity, and they are working to create the right structures to record them. However, some of these mechanisms are still not ready and the data provided to me are not complete. As a consequence, my analysis can boast the presence of precise variables and data with the consciousness that a lot more could be done the moment the client will have the right technologies to capture more data related to its activity.
As regards the characteristics of the dataset, this allows the analyst to observe for each driver every delivery he has done during the month of March 2018, with the specification of the exact day, the exact time, address (street, district) and kg of the delivery. The dataset contains other variables that are not taken in consideration now but whose description is available on the *project proposal document* and as such available for every suggest of external reader.
On the other hand, the precise data that are available for the process of delivery and above described, are not available **in relationship with the goods that are picked up.** As a consequence, this factor represents a limit for the analysis. In fact, it is important to remember that during the process of delivery of goods (see *Figure I*) each driver has to simultaneously perform services of picking up of goods that will be brought to the factory at the end of the day and delivered in other cities in order to be given to their addresser. So, basing an analysis on the performance of the driver without having fundamental details about the picking up process (that represents around 10% of the activity of the daily performance of the driver), can lead into misleading results.

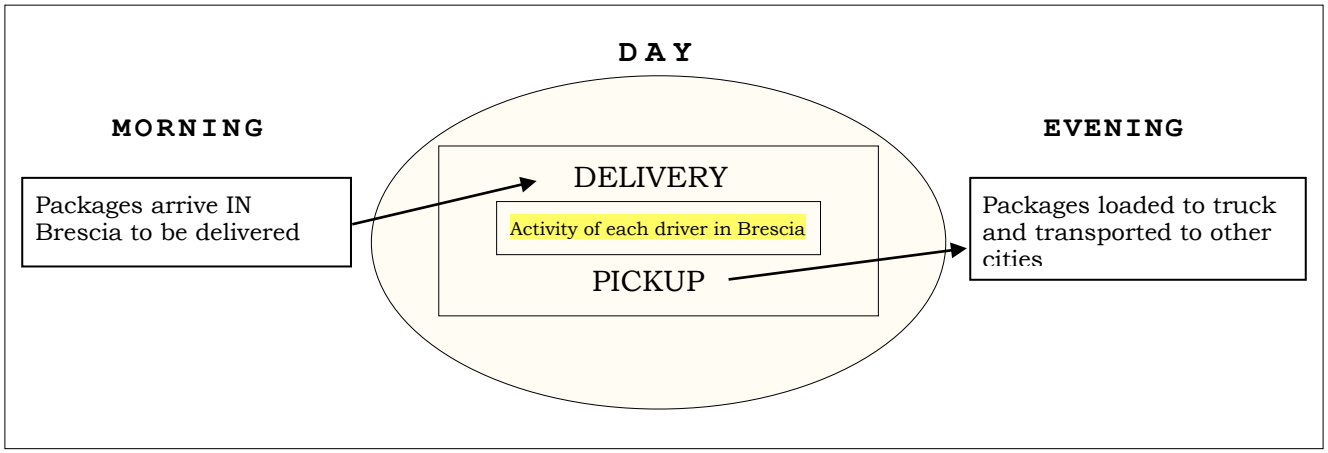

Figure I: Daily activity of the company with a focus on the

This problem have been partially curtailed thanks to a second dataset I have been able to obtain from the client. (see next paragraph for this).

As regards the analysis, on a first approach the goal was the one to have the possibility to verify if the daily itineraries of each driver were the most optimised, if the each driver could have performed a better process of delivery and if a planning model for the future could be created for this company.
However, this ambitious plan had to face a series of issues described as following:

1) Each address should have been turned into geographic coordinates.
Unfortunately, the google geocode function available on R allows people to only perform 2,500 queries per day. This situation was hard to sustain due to the presence of over 100,000 unique addresses.
2) The lack of variables such as address and pickedup_time available for the process of picking up would have turned into a non precise model.
This means that the goal is not supported by the data available. Moreover, data are probably not enough. It is necessary to work on a dataset with several month - probably at least a year - to perform a complete analysis, instead of just one month.
3) Lack of competences. Being ambitious is the best way to have results, but every long road starts from a step. From this point of view, it is better to simplify the goals. In fact, this represents my first work of analysis and that is why now I just have to face real problems and provide myself the right knowledges to growth professionally and to be able to solve complicate problem in the future.

Recently, the client has provided a new series of data that helped the analysis to become more complete. In particular, this dataset contains the daily summary of a series of variables for each driver. These variables are the total packs that arrived every day for the specific area of the driver, the number of packs loaded on the van every morning, the number of packages not delivered by the end of the day and to conclude the total number of picked up services that he has provided during the day.
This dataset have represented an important element for the goal of obtaining an analysis as more complete as possible.

GOAL OF THE ANALYSIS

Obtained all the available data, the research has moved into the direction of an analysis of the performance of each driver.
In particular, the goal is to obtain and analyse a dependent variable which has been chosen to be the ratio between deliveries and worked hours for each driver. As a consequence, the variable driver represents the statistical unit of the analysis.
In this way, the project aims to verify which factors are more relevant to define the process of distribution of goods for each driver. In particular the focus will be on the definition of the number of deliveries which can be performed per hour for each driver, based on a series of variables which will be considered in the predictive model.

CLEANING PHASE RESULTS

Dealing with real and raw data is not easy and it has resulted in a long and delicate phase of wrangling which was necessary to obtain clean data for the future analysis.
From this point of view, several variables contained a series of **missing values** which have been analysed and handled based on the kind of variable. In particular, the observations related to missing values for the driver and the picked up time have been deleted, while the ones related to the address of delivery have been substituted with the name "UNKNOWN", as they could result useful for the aim of the research.
The big amount of data and variables of class character led to the necessity to deal with the **difference of spelling** for a big amount of strings. The districts of delivery have been fixed on their spelling before to be handled for probably the hardest part of this phase: in this context, each district have been associated to its cap using the google geocode function. This allowed, from one point of view, to simplify the analysis (461 districts against 77 postal codes) and for the other to verify the effective existence of each district in the city. From this point of view, the *district_delivery* variable contained over 461 unique districts and this lead to the impossibility to manually verify the existence of everyone.
For the second dataset the first necessity regarded the need to join each dataset together. In fact, the client provided a different dataset for each day (with the same variables). As a consequence, it was necessary to obtain a unique dataset.
After have deleted not useful observations such as the presence of driver codes from another field office of the company - not object of my analysis - and added a variable regarding the day, very useful to allow the comparison between the two dataset, the new dataset has been created.

```
> str(data_exploratory)
'data.frame':   146896 obs. of  16 variables:
 $ arr_to_bs        : chr  "2018-02-27" "2018-02-28" "2018-03-01" "2018-02-27" ...
 $ day_deliv        : chr  "01" "01" "01" "01" ...
 $ month_deliv      : chr  "03" "03" "03" "03" ...
 $ address_delivery : chr  "via frua," "via arnoldo bellini, " "calchera, " "via g.verdi ," ...
 $ weight_pack      : num  6.1 2 22.7 6 0.8 0.4 2 4.2 6.8 3.8 ...
 $ num_pack         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ driver_code      : chr  "113" "113" "113" "113" ...
 $ district_delivery: chr  "roe' volciano" "roe' volciano" "villanuova sul clisi" "roe' volciano" ...
 $ postal_code      : chr  "25077" "25077" "25089" "25077" ...
 $ delivery_date    : chr  "2018-03-01" "2018-03-01" "2018-03-01" "2018-03-01" ...
 $ weekday_deliv    : Factor w/ 6 levels "lun","mar","mer",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ delivery         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ pickup_time      :Class 'times'  atomic [1:146896] 0.629 0.678 0.467 0.635 0.594 ...
  .. ..- attr(*, "format")= chr "h:m:s"
 $ range_hours      : Factor w/ 12 levels "8_9","9_10","10_11",..: 8 9 4 8 7 9 4 9 8 8 ...
 $ lon              : num  10.5 10.5 10.5 10.5 10.5 ...
 $ lat              : num  45.6 45.6 45.6 45.6 45.6 ...
```

Figure II: dataset number 1, ready for the analysis

```
> str(aggregate_data4)
Classes 'tbl_df', 'tbl' and 'data.frame':       1918 obs. of  13 variables:
 $ sedi           : chr  "BS" "BS" "BS" "BS" ...
 $ driver_code    : chr  "425" "116" "664" "612" ...
 $ driver_name    : chr  "USMAN MUHAMMAD QULB" "KATIR GSA" "KUMAR GORAV K2" "
 $ pack_arrived   : int  103 97 111 80 56 69 88 77 64 86 ...
 $ pack_loaded    : int  98 99 62 62 69 64 63 85 94 112 ...
 $ not_delivered  : int  26 24 14 1 12 1 4 9 12 11 ...
 $ kg_delivered   : int  466 960 627 356 297 286 369 469 802 808 ...
 $ pickup_services: int  4 21 1 11 1 11 8 11 13 12 ...
 $ pickup_packs   : int  6 71 1 41 1 64 27 28 53 19 ...
 $ kg_rit         : int  16 841 1 1384 2 775 635 229 535 103 ...
 $ tot_serv       : int  102 120 63 73 70 75 71 96 107 124 ...
 $ tot_kg         : int  482 1801 628 1740 299 1061 1004 698 1337 911 ...
 $ day            : chr  "1" "1" "1" "1" ...
```

Figure III: dataset number 2, ready for the analysis

EXPLORATORY ANALYSIS

1) The phase of exploratory analysis took place with a generic overview of the situation of the company. From this point of view it has been showed some interesting trends that regard the majority of deliveries are performed in the morning, with a peak in the range between 11 and 12. The analysis of the days of the month and weekdays suffer the lack of enough day observed in order to have possibility to find a trend. In this context, the issue regarding a strike which took place the 20th of March have probably modified the process in the following days, making hard to be sure about the trends observed. In conclusion, the following map allowed to identify the area with the majority of deliveries performed. More informations are available on the Data Exploratory document.

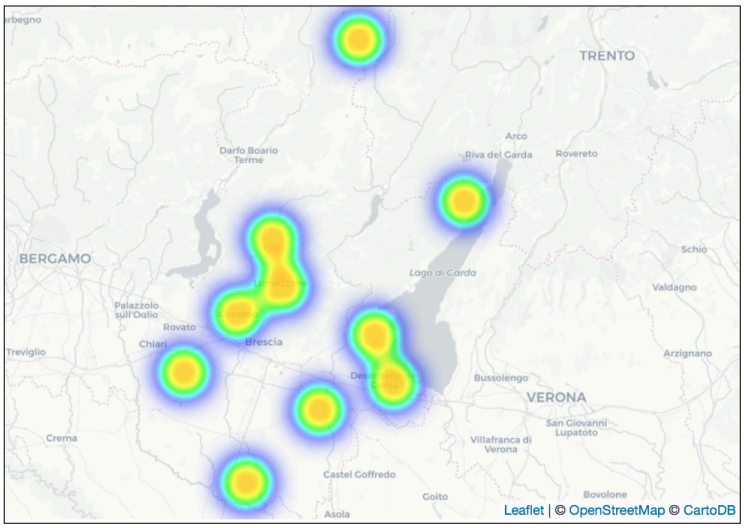

Figure IV: heat map containing the top five district of deliveries, after the city centre of Brescia

2) The second phase, whose aim is to obtain the dependent variable of the predictive model, aimed the necessity to create aggregate data. In particular, after having decided to work on each driver as statistical unit the goal has been defined to be the one to summarise the situation in relationship to each unit. The dependent variable regards the ratio between total deliveries and worked hours for each driver.

During this phase the main focus has been on the aggregation of data using the group_by and summarise function of the R Dplyr package. In particular, starting from the time of each packages delivered, it has been possible to obtain the total amount of hours each driver has worked and the total amount of deliveries that have been performed by them. As a consequence the dependent variable has been obtained as following:

$$y = \frac{tot.deliveries}{worked.hours} \quad \text{for each driver}$$

In addition, the independent variables obtained aggregating data and useful for the analysis are:

_ total packs loaded on the van
_ total packs arrived for each driver/zone
_ total packs not delivered at the end of the days
_ total packs picked up
_ total weight of the packs picked up

To conclude, the necessity to use the information relative to the area of delivery leaded to the opportunity to split the city into three different area based on the distance from them and the centre of the city: In fact, it is supposed to have similar characteristics of performance due to different distances between points of delivery and different traffic. In order to do so I had to use the R mapdist function to turn addresses into coordinates

3) The last phase allowed to put in relationship the dependent variable with a series of independent variables in order to analyse and observe or verify some trend.
From this point of view strong relationship do not seem to be obtained. However, the weight of the packages delivered seems to have a negative influence on the number of packages delivered per hour. On the other hand, it is possible to observe that the performance of the driver in term of packages delivered each hour can also be influenced by the total amount of services done. In this context, the increasing of the number of services result in a higher value of the dependent value, with driver with a better performance.
To conclude, the area between 15 and 30 km from the centre (the area defined as Zone2) is the area where the performance of the driver is less constant. From this point of view, there is a positive relationship between deliveries performed in that area and the ratio between deliveries and hour worked, which range from 6 to 14.

PREDICTIVE MODEL

The possibility to deal with a various dataset has led to the possibility to perform a Machine Learning analysis in which the goal has been to obtain a model that could predict the ratio between deliveries and hour for each driver, based on the conditions described by the above mentioned predictors.

From this point of view, the first necessity regarded the need to fix the issue related to one of those variable. In fact, the parameter regarding the number of pack arrived each morning from other field offices to the specific driver/area contained a series of missing values that would have led to the impossibility to adopt some predictive models. The problem has been fixed applying a model of prediction to the dataset, where the variable with missing values could represent the response variable of analysis. In this way, the 26 missing values have been treated as the dependent variable of analysis of the test dataset, and they have been predicted with the regression tree model.
Secondly, the differences of scale between the different parameters have been fixed through a standardisation of each value. This situation have made possible to perform the standard Linear Regression to the dataset, as the response variable could have reached also negative values. Moreover, two penalised linear models have been applied:
_ Lasso
_ Ridge

In conclusion, the analysis has led to the application of the Regression Tree which, despite the generic efficiency of this model in different context, has demonstrated to not be the best model to fit this specific case. In fact, the mean of the squared difference between the sum of the y predicted by the model on the test data and the sum of the effective y of this data has been used to compare the different models. This difference has been the biggest when applied to the Regression tree predictive model obtained. On the other side, the Ridge model has resulted to be the best in comparison with the other obtained.

CONCLUSION

The project of analysis of the data Kindly furnished by GLS-Italy has represented my first personal approach to the world of the analysis of data.
From this point of view, a lot of efforts have been done by myself in order to perform a precise and adequate analysis, with the consciousness of this project to be a gym for me to improve myself. As a consequence, the same analysis in the future will surely lead to better results and an improved analysis. The awareness of my gaps is conscious on myself. At the same time, some suggests should be adopted by the company as well in order to create the conditions to improve its activity and organisation in the future.

1) There is a necessity to create databases and structures of recording of data that can be more objective and defined. From this point of view, activity should be more automatised and every time a data is entered on a database, this should be objective and previously defined, avoiding the risk that different people enter the same information in a different way (name of cities, addresses, comments about the delivery etc).
2) The precise data available for the process of delivery should be created for the process of picking up of goods as well. This situation could allow future analysis to be more precise and accurate with the possibility to obtain models that can automatically predict the road of each driver based on the parameters of interest. Also, the possibility to have such an accurate database for a period longer than a month (an year, minimum) could allow to obtain and verify strong trends that can lead to the application of modifications on the activity and new strategies for the future.
3) The company, such as all the other logistic companies in the world, is facing the increase of online shopping by almost every kind of client. From this point of view, the type of person who is waiting for the delivery can greatly influence the activity of these companies. In particular, some people can be not enough aware of how a activity of delivery works and it can happen that often when the driver tries to delivery the good the client is not home; this situation can happen for different reason, but it can become a problem when the client already knew he would have not been home at that time. In fact, in this way the client caused a damage to the company because of unconsciousness. This situation has to be avoided and companies should work on it. In this context, there is a necessity to:
    - Educate the clients about the process of delivery , and let them know and understand that they can have different possibilities when they order something like delivering the good at their work place, to a family member or to specific place on the city, and that the process of delivery is necessarily a collaborative operation between the deliver and the client.
    - Provide to the client more choices about their possibility, for example asking them what is the best hour or day they would want to receive the package, being more communicative with them through a specific app where they can have the possibility to actively participate to the process of delivery feeling more involved on the activity.

04/07/2018,
Simone Zanetti