People Analytics: Measure High Performers Through Data

Rodrigo Araujo Lima Torres
Matrícula: 180150987
Email: rodlimatorres@gmail.com

***Abstract*—People Analytics is an integrated process for collecting, organizing and using data analysis techniques on sources like: social networks, e-mails, climate surveys, goal achievement, internal development, ERP and BI solutions, among others.**
As a result of data analysis, it becomes possible to anticipate trends, understand user behavior, produce insights and adopt probabilistic projection models (predictions) and present solutions to problems that affect performance, productivity and engagement of employees.
Jenny Dearborn [1] affirms that leaders severely limit their ability to solve problems and advance business strategies if they lack on leveraging data analytics.
Big Data is now everywhere, including the workplace. Real-time information on employee engagement, performance and productivity is everywhere, making it possible to tie talent to business outcomes in ways that were almost impossible before.
Data mining methods like Apriori and Linear Regression could provide valuable information for decision-makers, that are reliable, efficient and scalable method for sequential pattern mining.
Apriori algorithm is a kind of basic algorithm which looking for frequent item sets, the basic principle is to use a iteration method called search step by step.
The purpose of this work is to allow the understanding of the factors and behaviors that most influence employee productivity and engagement, using these data mining methods.

***Index Terms*—Data Mining, ETL (Extract, Transform and Load), SQL Server, People Analytics, APRIORI Method, Linear Regression.**

I. Introduction

TO identify high performers within organization and map their behavior this research focus on a methodology called People Analytics. This method promises to help organizations understand their workforce as a whole, as departments or work groups, and as individuals, by making data about employee attributes, behavior and performance more accessible, interpretative and actionable (Tom Pape, 2016 [13]).

For this it is included the use of information systems, visualization tools and predictive analytic, underpinned by employee profiling and performance data. Using data mining techniques it is possible to understand what motivates employees, what are the factors and behaviors that most influence their productivity and engagement.

A. Problem Definition and Strategic Alignment

With part of the revenues coming from the Federal Union, based on federal law nº 8.029/90[1], from 13th of February of 1990, SEBRAE must observe the principles of transparency, universality and the balance of its budget execution, giving due importance in meeting the targets, focus on results and mobilization of employees.

The definition of individual performance, employee engagement levels, and employee absenteeism are considered to be essential for adequate management of public resources, better achievement of results, and achievement of business objectives.

Companies such as Accenture, General Electric, Starbucks and Adobe have adopted a new approach to employee assessment that allows for greater production and collection of performance information, generating a more meaningful and higher quality data mass for analysis. (Jean Paul Isson and Jesse S. Harriott [8]).

However, while data is paramount to good management, alone is not enough to produce value. Human Resources analytics programs evolve to focus on measuring and modeling the strategic impact of human capital, thus creating better decision tools (Boudreau, J. and Lawler, E. [14])

During this work the executive direction was observed and raised strategic questions about the business. In a non-intrusive way, the executive board was assisted in identifying critical business issues that need to be answered. To support the discussion about engagement, productivity and employee behavior it was brought the study from Tracy and Paul [9] based on over 14 million employee survey responses across 70 countries. This study combines principles of psychology and motivation and comes with the five factors that explains what helps employees achieve higher levels of engagement, as well as how employees can have a higher performance. These concepts drove the discussion in sense to help to define the variables that were used in this study and to have a clear expectation of the team and work involved to obtain the answers for the questions below.

Technical and behavioral skills of employees have been identified so that they can address critical business aspects as well as mapping the characteristics that affect productivity through the correlation of data collected at behavior

[1]http://www.planalto.gov.br/ccivil_03/leis/L8029compilada.htm

evaluation cycle, absenteeism and deliveries, intending to find the answers for the following questions:

(1) What is the degree of influence of the types of absenteeism registered in the company and the productivity of the employees?

(2) What competences are common in higher performing employees?

B. Business Context

Sebrae needs to observe the principles of transparency, universality and the balance of its budget execution, giving due importance in meeting the targets, focus on results and mobilization of employees.

For this study it was defined with the executives the technical and behavioral skills as well as the characteristics that affect productivity. It was used data collected at behavior evaluation cycle that contains necessary information about Sebrae's employees, that are directly linked to the ability to assist the business community to develop sustainable businesses. This work produced correlations and understand which variables or characteristics are most related to high performers at Sebrae.

Managers and executives expressed the need of skills assessment and absenteeism data correlation to identify factors that impact employees performance.

The information collected at behavior evaluation cycle of year 2017 is registered on a variety of systems like Enterprise Resource Planning (ERP), Business Intelligence (BI), and internal applications and employee performance indicator that provide valuable data and variables that will be used during this study.

To achieve the goal of identifying high performer characteristics, it is used APRIORI and Linear Regression data mining algorithms, which has the basic idea of finding all the item sets within the minimum frequency predefined support degree, and the strength of correlation between variables. (He,Lei and Qi,Jiaxin, 2013 [5])

So, this research explores statistical analyzes with a significance level of 75% in order to identify the factors and behaviors that are most strongly related to employees' performance.

The phases used in this research follow the propose made by Kitchenham [3] in the state of the art were the Strategic Alignment, Literature Review and Planning, Collection and Analysis (based on CRISP-DM model), Predictions and Correlations, and Results Publication, which have the roles of: identifying the information needs related to business challenge; recognize related studies and plan the method of collecting the information; extract usefull information related to the problem and define metrics and variables; correlate information and produce predictions; and documentation of the results, respectively. Each of these phases are organized in the rest of this paper sections where there is an extra one (Section 6) related to conclusions and future discussions.

Sebrae Nacional
September 20, 2018

II. Literature Review and Planning

The bibliography review was made using the state of the art research on subjects involving planning, human resources, people analytics, algorithms and data mining methods like Apriori, Linear Regression.

Planning: Kitchenham [3] - method of scientific research. It helps the researcher to identify, analyze and interpret primary studies available in an electronic medium containing evidence linked to a particular field of research in a way that understands what has already been done and what can still be done in this field.

Article Research:To search for related works, it was used a method to exempt the researcher's bias while selecting keywords and search bases. The source used was "CAPES Periódicos", where a training was conducted on how to carry out effective consultations in the Capes' website Journal Portal. The search criteria in this base was: Articles that contained nine terms: Achievement; APRIORI; Analytics; "Data Mining"; Engagement; "Human Resources"; "People Analytics"; Linear Regression; Performance and Prediction. This information would help to compose the parameters for the main search. The exploratory research route was as follows: (1) The research was done in the CAPES periodicals portal. (2) The five most relevant articles of each term were taken. The key words were sorted into a list of the most recurring to the least recurrent. (3) There was a cut of the five most recurring key words. (4) Articles were ranked, with those with the most recurring keywords at the top.

People Analytics: Tracy and Paul [9] - Evaluation of employee engagement. Authors define 5 factors related to technical and behavioral skills named as MAGIC: Meaning, Autonomy, Growth, Impact and Connections. This research helped to define the variables that should be used and that is related to Sebrae business problems.

Also, Yogesh Pal Sujeet N. Mishra, and Dev Raghvendra Lama [12], defines that the purpose of Human resource predictive analytics is measuring employee performance and engagement, studying workforce collaboration patterns. This can help to optimize performances and produce better return on investment for organizations through decision making based on data collection, HR metrics and predictive models.

Analytics: Mortensen, M., Doherty, N. and Robinson, S. [11] - Aims to assist in the decision-making process through the use of quantitative methods to organize, and give meaning to the data that are being generated.

Data Mining Methods and Algorithms: Mueen, A., Zafar, B., & Manzoor, U. [10], describe Modeling and predicting students' academic performance using data mining techniques such as Decision Tree, Linear Regression, Naïve Bayes, and Neural Network.

He, Lei and Qi, Jiaxin, 2013 [5] also discussed the basic idea of APRIORI algorithm model on the enterprise human resources information showing it's efficiency, practicality and effectiveness, that was very helpful for this research.

According to Wei-Chao Lin [6], Wu et al. [7] has identified

the top (or frequently used) ten data mining algorithms, which covered classification (including C4.5 decision tree [C4.5], support vector machine [SVM], AdaBoost, k-nearest neighbor [K-NN], naïve Bayes [NB] and classification and regression tree [CART]), clustering (including k-means and expectation maximization [EM]), statistical learning (i.e. EM and NB), association analysis (i.e. Apriori) and link mining (i.e. PageRank).
These works presents APRIORI as being useful at obtaining correlation patterns for a sequence information, which reflects the characteristics of a corresponding application domain.
So, this article focuses on the use of Linear Regression and APRIORI methods to identify the patterns of absenteeism, skill types and behaviors that are common among the highest performing employees, through the correlation of skills assessment, absenteeism and deliveries data, evaluating the degree of influence of these characteristics.

III. Collection and Analysis (based on CRISP-DM model)

In this paper CRISP-DM is referred as a process that consist of a particular course of action intended to achieve the results. It consists on a cycle that comprises six stages: Business understanding: Focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives. *This topic was already covered in this article*; Data understanding: Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information; Data preparation: Covers all activities to construct the final dataset from the initial raw data; Modeling: Modeling techniques are selected and applied and their parameters are calibrated to optimal values; Evaluation: Obtained models are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives; Deployment: Knowledge gained will need to be organized and presented in a way that the customer can use it. Chapman et al, 2000 [4].

A. Data Understanding

Data collection and proceeds of ETL (Extract, Transform and Load) with activities in order to get familiar with the data, to identify data quality problems were conducted. It was collected **71.979** registers involving competency data, types and quantities (hours) of absenteeism and information about employee performance. Figures 1 and 2 respectively represents the Demonstration of Personnel Availability based on types and the average long term absenteeism expressed in hours during each month of year 2017.

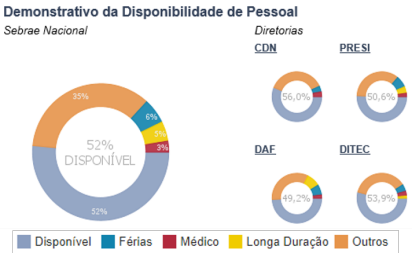

Fig. 1: Demonstration of Personnel Availability.

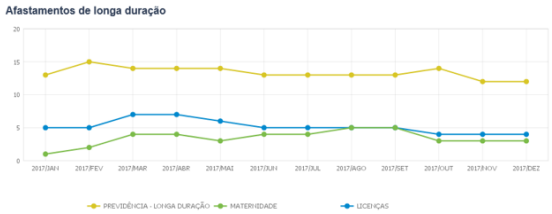

Fig. 2: Average long term absenteeism of year 2017.

Related to competences, figure 3 indicates the percentages of types of competency assessment and stratification of its subcategories expressed by the sum of occurrences.

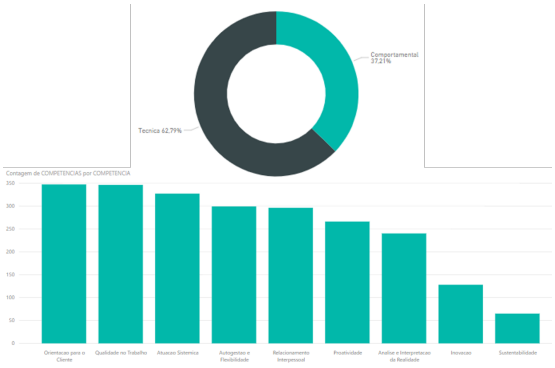

Fig. 3: Percentage of each Competence Type and amount of its subcategories.

B. Data Preparation

With help from specialists, data related to Sebrae's competency assessment were evaluated and extracted from the various systems. Data were gathered among systems like: BI, TOTVS ERP, Microsoft SQL Server databases and cycle evaluation applications systems. During this process it was conducted some transformation, data quality adjustments and some technical procedures using SQL, Excel and PowerBI tool, like:
(1) Removal of noise or spurious data (blank lines)
(2) Inclusion of 0 values for the missing notes due to lack of evaluation in that competence
(3) Formatting column names by removing characters not recognized by R (%, é, í, ó, ç ...)
(4) Reverse normalization of performance assessment data tables, behavior notes, absenteeism and deliveries, creating a single "Scoreboard"

(5) Separation of Qualitative and Quantitative Data
(6) Discrete of numerical data to delimit the appropriate interval and definition of attribute types (int, char)
Best practices define the importance of data split into training and validation portions. So, it was divided a sample of **70%** for training and the rest of remaining data **30%** were separated for Validation. The process is represented by the following R function (figure 4):

```
## Separação dos dados de Treinamento e Validação

```{r}
set.seed(123)
split <- initial_split(Desempenho, prop = .7)
Treino <- training(split)
Validacao  <- testing(split)
```
```

Fig. 4: Data split: Training - 70%; Validation - 30%.

Also, data should be prepared for APRIORI algorithm execution. It is possible to find on formula bellow the main parameter from APRIORI algorithm defining the minlen (representing the correlation of 3 factors), support of 0.4 and confidence of **75%**. The following steps were taken in place for the correct data preparation:
(1) Convert data from skill types to a categorical format
as.factor(Validacao$COMPETENCIAS)
(2) Group skills assessments by the same employee name
ddply(Treino, c("NOME"), function(Lista1) paste(Lista1 $COMPETENCIAS,collapse = ","))
(3) Convert the list to bascket format and set the Support and Trust Rule
read.transactions(file="Lista-Competencias.csv", rm.duplicates= FALSE, format="basket",sep=",",cols=1); apriori(txn,parameter = list(minlen=3, sup = 0.4, conf = 0.75, target="rules"))
With this parameters it was possible to define the minimum frequency of 3 (**minlen**), a support degree of 0.4 (**sup**), and the strength of correlation between variables of **75%** (**conf**).

C. Modeling (APRIORI and Linear Regression)

The APRIORI method is used to obtaining correlation patterns for a sequence information, which efficiently reflects the characteristics of a corresponding application domain. During this process it is also included multi-source data collection, metrics definition, modeling and parametering of the tools, synthesizing all available information that will help answer the performance and engagement questions. The implementation of new processes and methods of data collection is conducted involving analysts from the different areas of SEBRAE, service providers and specialists.
During modeling it was checked the correlation level between **7 variables** and one predictor **final punctuation Score**. The code used to measure strength of variable correlation to its predictor variable **Placar-Pontuacao-Ciclo**, using R was: *cor (Validacao$Placar-Pontuacao-Ciclo, [each one of the seven variables])*.
However, CompTec category variable has 7 subcategories and Comp.Comport category variable has more 4 subcategories coefficients that were analyzed individually since it was discovered both of them have a high correlation with performance. The function above produced a result where was possible to observe that variables related to Technical (Comp.Tec), Behavior (Comp.Comport) and number of courses have the highest correlation with the performance of Employees, as shown on table I below:

TABLE I: Correlations between variables and predictor

Variables	**Predictor Placar_Pontuacao_Ciclo**
Afastamentos_Medicos	0.3374574
Soma_Horas_BH	0.3215776
Comp.Gestao	0.1154404
Comp.Tec	0.8678192
Comp.Comport	0.871462
Escolaridade_Qtd.Cursos	0.6753996
Placar_Qtd.Steps_Ciclo	0.3514448

D. Evaluation

The information gathered in Data Mining phase will allow the correlation of information, define support degree, and identify the confidence related to the strength of correlation during analysis. Analytical information (algorithms, indicators and statistical methods), graphical correlations, tests and validations of the hypotheses, comparative analysis, and clear and concise interpretations of the data will be produced in the context of the problem. R and RStudio software was used to produce plots.

1) Linear Regression: Linear regression was used to create a predictive line through data. In this study the intent was to identify the amount of variability of the Productivity Scores explained by the Technical, Behavioral and Number of Courses competences. For this various of steps were executed as fallows:
(1) Identification of standard error of each coefficient: the value indicates how much the coefficient varies from your estimate. It is better as much lower this value is;
(2) Significance of the coefficient (t-value): The greater the absolute value, the greater the probability of the coefficient being significant;
(3) Probability of the correlation of the variable approaching zero: Pr (> | t-value |);
(4) Calculate the p-value level: Test analysis and hypothesis rejection;
(5) Calculation of R2: Indicates how much of the total variation is common to the analyzed coefficients and pairs.
Based on correlation factors presented on table II, variables Comp.Tec (Technical Competences), Comp.Comport (Behavior Competences), Escolaridade-Qtd.Cursos (Amount of courses and degree) and Afastamentos-Medicos (Medical Absenteeism) were considered for the analysis due to the obtained grade. R^2 of all Coefficients is 0.8199 indicating that 82% of the variability of the performance scores were explained by the technical and behavior competences, number of courses and absenteeism sick notes.

Table II presents the results of calculations described above:

TABLE II: Linear Regression Evaluation

Coefficients	**Std. Error (SE)**	**t-value**	**Pr (>|t-value|)**	**p-value**	R^2
Comp.Tec	0.85875	4.375	1.91e-05	2.2e-16	0.7531
Comp.Comport	1.66293	3.098	0.00222	2.2e-16	0.7594
Escolaridade _Qtd.Cursos	0.04982	7.054	2.45e-11	2.2e-16	0.4562
Afastamentos _Medicos	0.03215	5.232	4e-07	4.005e-07	0.1139
				Total R^2	0.8199

To evaluate if a linear model is reliable, we need to verify:
(1) linearity;
(2) constant variance; and
(3) normal residues distribution.
The linearity and variance constancy of the residues are indicated by its dispersion width, when the value of x increases. If this width varies as the value of x increases, residues' variance is not constant. Figure 5 shows dispersion width of the residues from variables Comp.Tec and Comp.Comport. There is a linearity and the width of residues has just a little variance.

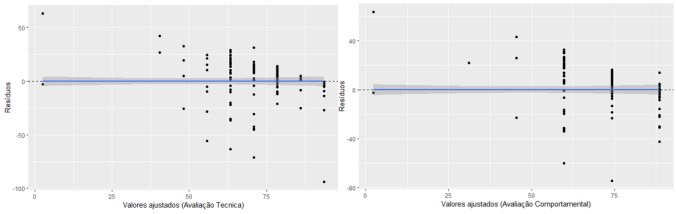

Fig. 5: Comp.Tec Comp.Comport Residues Dispersion

Therefore the variables related to courses and absenteeism have significant amount of residues dispersion, representing a not constant variance, as we can see on figure 6 below.

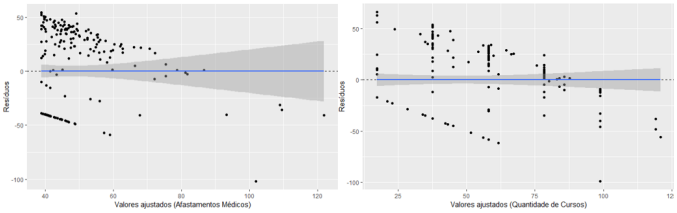

Fig. 6: Afastamentos_Medicos and Escolaridade_Qtd.Cursos Residues Dispersion

Among the coefficients, Comp.Tec and Comp.Comport variables had the lowest residues dispersion and a more constant variance.
Finally, the normal residues distribution occurs if the quantiles are positioned on a line with intercept at the origin and positive orientation, following the same distribution (normal). As we can see on figure 7 the square root of residuals from Comp.Tec and Comp.Comport variables, follows exactly that condition.

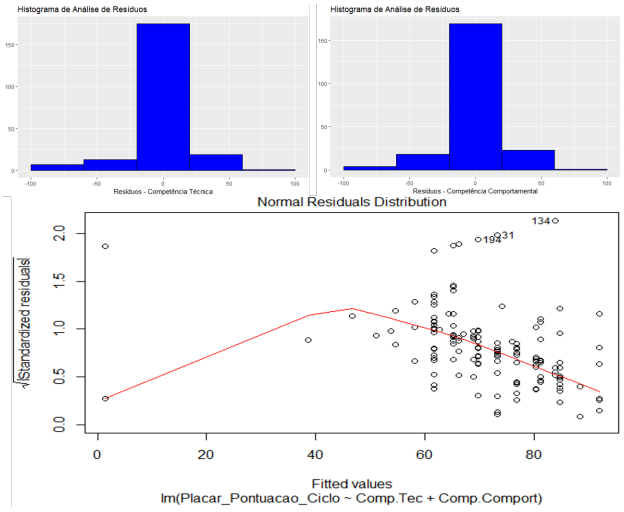

Fig. 7: normal residues distribution

Find below the R formula used to produce the graphs above.

```r
# Histogram and Residuals Dispersion
```{r}
qplot(x = .resid, data = model1, geom = "histogram", binwidth = 40, main = "Histograma de Análise de Resíduos", fill=I("blue"), col=I(1)) +
xlab("Resíduos - Competência Técnica")

plot(model1, which = c(1:3, 5),
     caption = list("Resíduos vs Valores Adequados", "Normal Q-Q",
       "Localização-Escala", "Distância",
       "Resíduos vs Influência",
       expression("Distância vs Influência  " * h[ii] / (1 - h[ii]))),
     panel = panel.smooth,
     sub.caption = NULL, main = "",
     ask = prod(par("mfcol")) < length(which) && dev.interactive(),
     id.n = 3, labels.id = names(residuals(model1)), cex.id = 0.75,
     qqline = TRUE, cook.levels = c(0.5, 1.0),
     label.pos = c(4,2), cex.caption = 1, cex.oma.main = 1.25)
```
```

Fig. 8: normal residues distribution

2) APRIORI: Once it was identified that the variables representing Technical (Comp.Tec) and Behavior (Comp.Comport) competences are the most relevant to represent a employee productivity, it was used APRIORI to obtaining correlation patterns which reflects the characteristics of the corresponding domain, analysing all of the 11 subcategories as follows:
(1)Technical Competences: Analysis and Interpretation of Reality; Management of contracts and agreements; Management of programs, projects and activities; Innovation; Customer orientation; Quality at Work; Sustainability.
(2)Behavior Competences: Systemic Actuation; Self-management and Flexibility; Proactivity; Interpersonal relationship.
During evaluation of those subcategories, it was found that on technical skills, Customer Orientation and Quality at Work are the most frequent ones in the employees of greater performance. Figure 9 presents the competences that are common on high performers employees ranked by the higher ones and figure 10 shows the parallel coordinate plot presenting strength and direction of each competency converging to technical skills of Customer Orientation and Quality at Work.

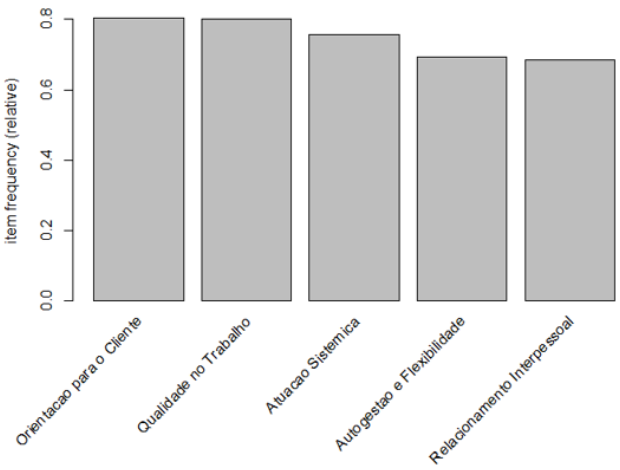

Fig. 9: Most frequent competences on high performers.

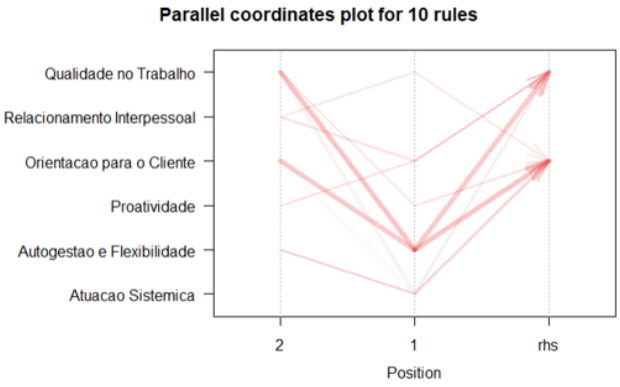

Fig. 10: Strength and direction of each competency.

Find below the R formula used to produce the graphs above.

```r
# split lhs and rhs into two columns
library(reshape2)
df_basket <- transform(df_basket, rules = colsplit(rules, pattern = "=>", names = c("lhs","rhs")))

set.seed(8000)

itemFrequencyPlot(txn, topN = 5) # Variable Frequency

#Parallel Coordinate Plot
plot(basket_rules[1:10,], method="paracoord",  control=list(alpha=.5, reorder=TRUE))
```

Fig. 11: Variable Frequency and Parallel Coordinate Plot

IV. Predictions and Correlations

In this section it is presented predictions as results of APRIORI and Linear Regression techniques. The projections produced allows to identify the factors that are most correlated to competences identified on high performers of SEBRAE employees.

1) Linear Regression: To predict trending of some statistic, Linear Regression uses a representation of a scatter plot with the minimum squares line, or in other words a linear regression is to select the line that minimizes the sum of the squares of the residuals. Figure 12 present the minimum square line of residuals from variables Comp.Tec and Comp.Comport. As we can see there are a low dispersion of errors.

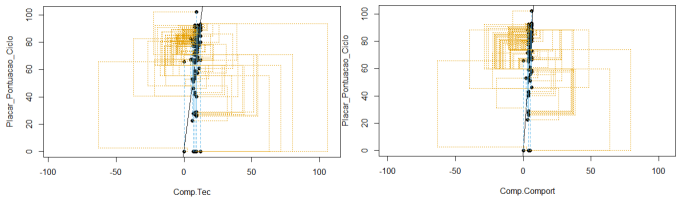

Fig. 12: Comp.Tec Comp.Comport minimum square line

Find below the R formula used to produce the graphs above.

```r
# Avaliação Técnica
plot_ss(x = Comp.Tec, y = Placar_Pontuacao_Ciclo, data = Validacao, showSquares = TRUE)

# Avaliação de Comportamento
plot_ss(x = Comp.Comport, y = Placar_Pontuacao_Ciclo, data = Validacao, showSquares = TRUE)
```

Fig. 13: Square Line Formula

Matching both variables in formula and showing the standard error "*SE*" associated with the line we get the Linear Regression Prediction Graph represented by figure 14 below:

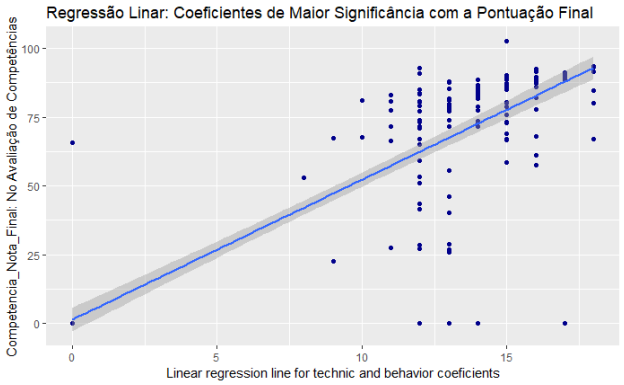

Fig. 14: Square Line Formula

The code used to produce the linear regression graph of figure 14 is described below:

```r
#Regressão Linear
```{r}
#ggplot(Desempenho, aes(Horas_Afastamento_Atestados,Competencia_Nota_Final)) +
Regressao_Linear <- ggplot(Validacao, aes(Comp.Tec + Comp.Comport, Placar_Pontuacao_Ciclo)) +

xlab ("Linear regression line for technic and behavior coeficients") +
ylab ("Competencia_Nota_Final: No Avaliação de Competências") +
ggtitle ("Regressão Linar: Coeficientes de Maior Significância com a Pontuação Final") +
geom_point(color ="dark blue") +
stat_smooth(method = "lm", se = TRUE)

#Gráficos
grid.arrange(Regressao_Linear,ncol=1)
```
```

Fig. 15: Linear Regression Formula

With Linear Regression Graph, produced by observed values of the population, it is possible to predict that: (1) 81.74% of the variability of the Productivity Scores were explained by the Technical, Behavioral and Number of Courses competences; (2) 77.43% of the variability of the Productivity Scores were explained by just considering most relevant variables (Technical and Behavior); (3) Absenteeism affects only 0.3% of employee performance, considering variables technical, behavior and courses; and (4) An employee who is evaluated by the same criteria and receives a score higher than 15 in the criteria of technical

and behavioral competence, has a high probability to be classified among the employees with highest performance at Sebrae. The table III shows the R^2 for each coefficient and for the combination of them.

TABLE III: Linear Regression Evaluation

Coefficients	R^2
Comp.Tec	0.7531
Comp.Comport	0.7594
Escolaridade_Qtd.Cursos	0.4562
Comp.Tec + Comp.Comport	0.7743
Comp.Tec + Comp.Comport + Qtd.Cursos	0.8174
Comp.Tec + Comp.Comport + Qtd.Cursos + Absenteeism	0.8199

2) APRIORI: The idea of APRIORI is to use a interaction method called search step by step algorithm that looks for frequent item sets. It uses k-itemset to explore(1)k+− item sets. This method needs multiple scanning on transaction database even the big, and it needs a great deal of I/O load. Apriori algorithm principles are: (1) Finding out all the item sets firstly, the appearance incessant of these item sets would be better at least the same with predefined minimum support degree;
(2) Producing strong association rules by the item sets, these rules must meet minimum support (fraction of transactions that contain X and Y) and minimum confidence (measures how often Y appears in transactions that contain X); (3) If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets; and (4) If the lift is < 1, that lets us know the items are substitute to each other. This means that presence of one item has negative effect on presence of other item and vice versa. This research found 10 rules and 432 transaction(s), for a minimum frequency of 3, a support of 0.4 and a confidence of 75%. To predict trending the focus of this research is on rules with lift > 1. Table IV present the *four* rules that matches this criteria:

TABLE IV: APRIORI Results

LHS	RHS	Sup.	Conf.	Lift	Count
Proatividade, Qualidade no Trabalho	Orientacao para o Cliente	0.405	0.814	1.013	175
Autogestao e Flexibilidade, Qualidade no Trabalho	Orientacao para o Cliente	0.458	0.811	1.010	198
Orientacao para o Cliente, Proatividade	Qualidade no Trabalho	0.405	0.810	1.012	175
Autogestao e Flexibilidade, Orientacao para o Cliente	Qualidade no Trabalho	0.458	0.805	1.005	198

Based on results presented above it is possible to observe that behaviors characteristics of Pro-activity, Self-Management and Flexibility have a direction to technical competences Customer Orientation and Quality at Work. Also, there is a strong dependence between this two technical competences were each one have a direction to the other in a total of eight rules from 10 produced APRIORI algorithm. The figure 16 illustrates those relations and its strength based on lift and support:

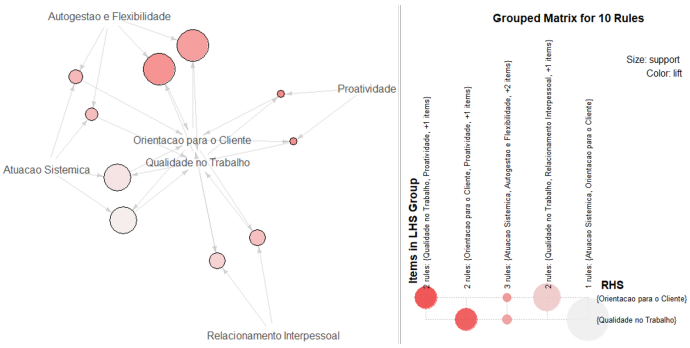

Fig. 16: APRIORI Relation Direction and Strength

With this results is possible to predict that a employee with behaviors of pro-activity, self-management and flexibility tends to be customer orientated and have quality at work produced.

V. Results Publication

During the research it was identified that variables representing technical (Comp.Tec) and behaviors (Comp.Comport) competences are most related to high performers inside Sebrae organization. This two sets of variables have almost the same strength to explain competence evaluation grades where Comp.Comport set has a higher R^2 and a lower sum of the residuals squares, representing a better linear regression line. Applying APRIORI algorithm it was observed that behaviors characteristics of Pro-activity, Self-Management and Flexibility leads to technical competences of Customer Orientation and Quality at Work. Also it was found a strong dependence between this two technical competences. With this results it is possible to answer the questions raised during the problem definition and strategic alignment, as follows:
What is the degree of influence of the types of absenteeism registered in the company and the productivity of the employees? It was observed that medical absenteeism affects only 0.3% of the employee performance, while analyzing technical, behavior and courses variables, being considered a weak influence in relation to the other coefficients evaluated by the study.
Which competency assessment factors are most strongly linked to employee performance? It was observed that the technical skills of Customer Orientation and Quality at Work are direct linked with each other. These factors are most frequent in people with higher grades at competency assessment. Also it is possible to

predict that a employee with behaviors of pro-activity, self-management and flexibility tends to be customer orientated and have high quality at works produced.

VI. Conclusions and Future Discussions

According to the state quo of human resources information mining, this paper applies Apriori, Linear Regression algorithms and data mining technology into the human resources information mining at Sebrae, and brings a method on how to analyze different aspects and characteristics of employee competences and behaviors to identify factors that impact employees performance. The algorithms results and methods used allowed to answer business questions with high precision and has a strong practical application, which worth's for spreading use. As result of this work, new discussion and questions might rise related to human resources information and data mining technologies such as:

(1) What actions can be taken to help people develop and have technical and behavioral characteristics that are close to those identified as most strongly linked to employee performance?

(2) Could mining techniques such as Naïve Bayes and Neural Networks bring different and / or complementary results to the studies?

(3) What is the level of engagement and satisfaction of the highest performing employees?

Acknowledgment

I'd like to thank Prof. Dr. Marcelo Ladeira - CIC/UnB - who works at Department of Computer Science of the University of Brasília and is Member of the CNPq Advisory Committee.

References

[1] Jenny Dearborn. Data analytics: The four steps every leader should take now. Leader to Leader, pages 38–43, 2018.

[2] Huan-Jyh Shyur, Chichang Jou, and Keng Chang. A data mining approach to discovering reliable sequential patterns. Journal of Systems and Software, 86(8):2196 – 2203, 2013.

[3] B. Kitchenham, "Procedures for performing systematic reviews," Keele, UK, Keele University, vol. 33, no. 2004, pp. 1–26, 2004.

[4] Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C. & Wirth, R. (2000), 'CRISP-DM 1.0 Step-by-step data mining guide' , Technical report, The CRISP-DM consortium.

[5] Lei He and Jiaxin Qi. Enterprise human resources information mining based on improved apriori algorithm. Journal of Networks, 8(5):1138–1145, 05 2013. Copyright - Copyright Academy Publisher May 2013.

[6] Wei-Chao Lin, Shih-Wen Ke, Chih-Fong Tsai, (2017) "Top 10 data mining techniques in business applications: a brief survey", Kybernetes, Vol. 46 Issue: 7, pp.1158-1170, https://doi.org/10.1108/K-10-2016-0302

[7] Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. (2008), "Top 10 algorithms in data mining", Knowledge and Information Systems, Vol. 14 No. 1, pp. 1-37.

[8] Jean Paul Isson and Jesse Harriott. People analytics in the era of big data: Changing the way you attract, acquire, develop, and retain talent. hoboken, nj: Wiley, 416 pages, $49.95, hardcover. Personnel Psychology, 70(4):929–930, 2016.

[9] Tracy Maylett and Paul Warner. Magic: Five keys to unlock the power of employee engagement, $20.95, hardcover. Greenleaf Book Group Press, 1(1):245–253, 2014. Copyright - 2014 - Decision Wise, Inc.

[10] Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. International Journal of Modern Education and Computer Science, 8(11), 36-42, 2016.

[11] Mortensen, M., Doherty, N. and Robinson, S. (2015). 'Operational research from taylorism to terabytes: a research agenda for the analytics age'. European Journal of Operational Research, 241: 3, 583–595.

[12] Yogesh Pal Sujeet N. Mishra, Dev Raghvendra Lama. Human resource predictive analytics (hrpa) for hr management in organizations. INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, 5(5):3, 2016.

[13] Tom Pape. Prioritising data items for business analytics: Framework and application to human resources. European Journal of Operational Research, 252(2):687 – 698, 2016.

[14] J. Boudreau and E. Lawler. 'making talent analytics and reporting into a decision science. Centre for Effective Organisations, University of Southern California, 2015.

[15] David Angrave, Andy Charlwood, Ian Kirkpatrick, Mark Lawrence, and Mark Stuart. Hr and analytics: why hr is set to fail the big data challenge. Human Resource Management Journal, 26(1):1–11, 2016.

[16] Sjoerd van den Heuvel and Tanya Bondarouk. The rise (and fall?) of hr analytics: A study into the future application, value, structure, and system support. Journal of Organizational Effectiveness: People and Performance, 4(2):157–178, 2017.

[17] Rodrigo Rezende Ferreira. Avaliação de necessidades de aprendizagem no trabalho: Proposição e exploração de um modelo. Universidade de Brasília - Instituto de Psicologia, page 198, 2014. Orientadora: Prof a Dr a Gardênia da Silva Abbad.

[18] Elziane Bouzada Dias Campos. Competências empreendedoras : uma avaliação no contexto de empresas juniores brasileiras. Universidade de Brasília, Programa de Pós-Graduação em Psicologia Social, do Trabalho e das Organizações, page 161, 2015.