UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Vector representation of texts applied to prediction models

Deborah Bassi Stern

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Deborah Bassi Stern

Vector representation of texts applied to prediction models

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Statistics

Advisor: Prof. Dr. Rafael Izbicki

USP – São Carlos
February 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

S839v

Stern, Deborah Bassi
Vector representation of texts applied to prediction models / Deborah Bassi Stern; orientador Rafael Izbicki. -- São Carlos, 2020.
41 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. WordVectors. 2. Prediction models. 3. Natural language processing. 4. Neural networks. I. Izbicki, Rafael, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Deborah Bassi Stern

Representações vetoriais de textos aplicados a modelos preditivos

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *EXEMPLAR DE DEFESA*

Área de Concentração: Estatística

Orientador: Prof. Dr. Rafael Izbicki

USP – São Carlos
Fevereiro de 2020

ACKNOWLEDGEMENTS

I thank my advisor Rafael Izbicki for the inspiration and help he has given me throughout this dissertation.

I would like to thank all the support my parents and siblings have provided me.

I would like to thank my professors from Undergraduate and Master courses, specially, Sônia Regina Leite Garcia. I would like to also acknowledge the professors of the evaluation board.

I thank my former coworkers, specially, Marcelo Aparecido De Paula Rosa, Jairo Cavalcante de Souza and Wilson Freitas for their encouragement and guidance.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

RESUMO

STERN, D. B. **Representações vetoriais de textos aplicados a modelos preditivos**. 2020. 41 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Processamento de linguagem natural sofreu uma grande mudança com o tempo. Abordagens estatísticas passaram a ganhar atenção apenas recentemente. O modelo word2vec é uma destas. Ele é uma rede neural rasa desenhada para ajustar representações vetoriais de palavras segundo seus valores semânticos e sintáticos. As representações de palavras obtidas por este método são o estado da arte. Este método tem muitas aplicações, como permitir o ajuste de modelos preditivos baseadas em textos. Na literatura é comum um texto ser representado pela média das representações vetorias das palavras que o compõem. O vetor resultante é então incluído como variável explicativa no modelo. Nesta dissertação propomos a obtenção de mais informação sobre o texto através de outras estatísticas descritivas além da média, como outros momentos e quantis. A melhora dos modelos preditivos é estudada com dados reais.

Palavras-chave: Processamento de linguagem natural, Redes neurais, Representação vetorial de palavras, Modelos de predição.

ABSTRACT

STERN, D. B. **Vector representation of texts applied to prediction models**. 2020. 41 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

Natural Language Processing has gone through substantial changes over time. It was only recently that statistical approaches started receiving attention. The Word2Vec model is one of these. It is a shallow neural network designed to fit vectorial representations of words according to their syntactic and semantic values. The word embeddings acquired by this method are state-of-art. This method has many uses, one of which is the fitting of prediction models based on texts. It is common in the literature for a text to be represented as the mean of its word embeddings. The resulting vector is then used in the predictive model as an explanatory variables. In this dissertation, we propose getting more information of text by adding other summary statistics besides the mean, such as other moments and quantiles. The improvement of the prediction models is studied in real datasets.

Keywords: Natural language processing, Neural networks, WordVectors, Prediction models.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

D2V	Doc2Vec
LASSO	Least Absolute Shrinkage and Selection Operator
NLP	Natural Language Processing
PCA	Principle Components Analysis
SLRM	Standard Linear Regression Model
W2V	Word2Vec

CONTENTS

CHAPTER

1

INTRODUCTION

Natural Language Processing (NLP) has gone through substantial changes over time. Specialists would implement fixed rules to assess texts until 1980. It was only in the late 80s and mid 90s that statistical approaches started receiving attention. Models based of word counting, vector representation of words and most recently the use of machine learning algorithms became standard in NLP field.

The Word2Vec (W2V) model is a shallow neural network designed to fit vector representations of words according to their syntactic and semantic values. The word embeddings acquired by this method are state-of-art. This method has many uses, one of which is the fitting of prediction models based on texts. It is common in the literature for a text to be represented as the mean of its word embeddings. The resulting vector is then used in the predictive model as an explanatory variable.

In this dissertation, we propose getting more information of text by adding other summary statistics besides the mean, such as other moments and quantiles. The improvement of the prediction models is then studied in real datasets.

This dissertation is structured as follows. In chapter 2 the W2V model and its application to texts is explained. Chapter 3 gives an overview of the prediction models used in this dissertation. In chapter 4 we use real datasets to compare the use of different summary statistics when fitting prediction models. Chapter 5 concludes the dissertation.

CHAPTER

2

VECTOR REPRESENTATION

The W2V method is used to fit coherent vector representations of words. It is based on the assumption that the higher is the association between two words, the higher is the probability of them appearing near each other in a text (GOLDBERG; LEVY, 2014). The notations and functions used to develop this hypotheses are presented in this chapter.

The collection of texts that compose our observed sample is called corpus. Given a specific word (w) in a text, the set of $2k$ words positioned closest to w is denoted context window of w of size k. Each word of the context window is called context (c) of w. Table 1 exemplifies the context window of size 1 of each word in a phrase:

This is a sentence.

Word (w)	Context (c)
this	is, a
is	this, a
a	is, sentence
sentence	is, a

Table 1 – Example of words and their respective contexts from a window of size 1.

The model estimates the conditional probabilities of the contexts given a word through the objective function

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta) \tag{2.1}$$

where D is the set of all word and context pairs that are present in the corpus and the conditional probability is given by the soft-max function

$$p(c|w;\theta) = \frac{e^{v(c)\cdot u(w)}}{\sum_{c'\in C} e^{v(c')\cdot u(w)}}$$

where $v(c)$ and $u(w) \in R^d$ are the vector representations of c and w, C is the set of all possible contexts, and, $\theta = \{u(w), v(c) : w \in Text, c \in C\}$.

Hence the log of Equation 2.1 has the form

$$\arg\max_{\theta} \sum_{(w,c)\in D} \left(\log e^{v(c)\cdot u(w)} - \log \sum_{c'\in C} e^{v(c')\cdot u(w)} \right) \tag{2.2}$$

The way the objective function is constructed makes similar words have similar representations (GOLDBERG; LEVY, 2014) because words that are alike should be strongly associated with similar contexts.

Notice that the representation of words are fitted separately of contexts, therefore each term has two different vector representations. The reasoning behind this is that a word a being its own context is rare, so $p(a|a)$ should have a low value. If contexts and words shared the same representation b then $b \cdot b$ would have to be small which is impossible (GOLDBERG; LEVY, 2014).

The maximization of the objective function 2.2 is done through a neural network with one hidden layer of size d. The words w will serve as input and the contexts c as output, as shown in Figure 1.

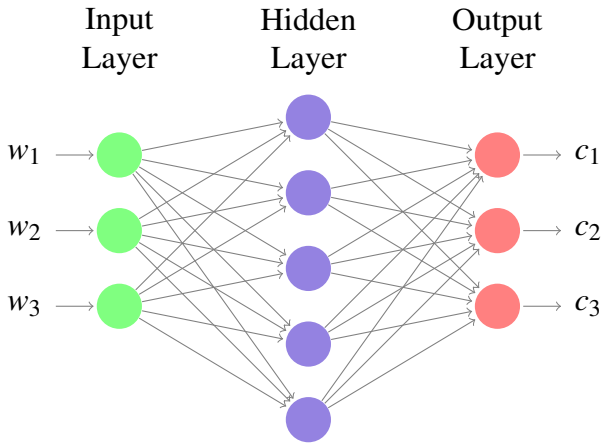

Figure 1 – Example of neural network with one hidden layer.

Oftentimes there is an interest to perform supervised learning on textual documents, that is, to predict a label based on a text. Having a way to represent a text, a collection of words, as a numerical vector would enable the use of common prediction models to achieve this objective. But W2V obtains vector representations of isolated words. We use the 8246-th instance from the Amazon Fine Food Reviews dataset presented in chapter 4 to better exemplify how the word embeddings are used to represent texts:

*This will be great with all that I use it for
 Thanks, Linda*

Only the words *great*, *use* and *thanks* from this example had their embeddings fitted by the corpus. The other words did not have their embeddings fitted because they were considered less informative: they had a low frequency on the corpus or were stop words, that is, words that are considered empty of meaning, such as prepositions, that generally have very high frequency. The removal of these words from the corpus affects the context window of many words and can result in better embeddings (GOLDBERG; LEVY, 2014; LEVY; GOLDBERG; DAGAN, 2015). The first two dimensions of this embeddings are presented at table 2.

words	Dim 1	Dim 2
great	-0.08	0.17
use	-0.06	-0.09
thanks	0.06	0.21
mean	-0.03	0.10
sd	0.06	0.13

Table 2 – Embeddings present at the 8246-th text from the Amazon Fine Food Reviews dataset.

The most straightforward way to represent a text would be through a matrix. Each word of the text is a row in this matrix and is represented by the embedding obtained by the presented method:

$$T^k = \begin{bmatrix} u(w_1^k) \\ \vdots \\ u(w_{n_k}^k) \end{bmatrix} = \left(T_{i,j}^k\right)_{n_k \times d} \tag{2.3}$$

where T^k is the k-th text in the dataset, n_k is the amount of words with an embedding in T^k, w_i^k is the i-th word of T^k with an embedding, $1 \leq i \leq n_k$, and u is the embedding given by Equation 2.2.

The text from the Amazon dataset would take the form:

$$T^{8246} = \begin{bmatrix} -0.08 & 0.17 \\ -0.06 & -0.09 \\ 0.06 & 0.21 \end{bmatrix}$$

However, all observations must have the same dimensions when fitting a prediction model. This is not guaranteed by Equation 2.3; the number of rows in T^k depends on the number of words with an embedding in that text, n_k. In order to solve this, it is standard in the literature that the matrix representation is converted to a vector through the column-wise mean of T^k (JOULIN *et al.*, 2016; LIU, 2017; BANSAL; SRIVASTAVA, 2018), as shown in Equation 2.4. This vector is then used as an input in the prediction models, that is, each element of this vector will be used as a regressor in the model.

$$S(T^k, g) = \begin{bmatrix} g(T_{\cdot,1}^k) & \dots & g(T_{\cdot,d}^k) \end{bmatrix} = \left(g(T_{\cdot,j}^k)\right)_{1 \times d} \tag{2.4}$$

where $T^k_{\cdot,j}$ is the j-th column of T^k and g is the mean function.

Using the Amazon text as an example:

$$S(T^{8246}, g) = \begin{bmatrix} -0.03 & 0.10 \end{bmatrix}$$

In this dissertation we propose the use of other summary statistics other the mean, such as the standard deviation, skewness, kurtosis and all quantiles. That is, we propose to represent a text as

$$S(T^k, \boldsymbol{g}) = \Big(S(T^k, g') \Big)_{1 \times d \cdot |\boldsymbol{g}|}$$

where $\boldsymbol{g}$ is a subset of summary statistics functions and $g' \in \boldsymbol{g}$.

If $\boldsymbol{g}$ contains only the mean and standard deviation functions, the Amazon text is represented by the vector:

$$S(T^{8246}, \boldsymbol{g}) = \begin{bmatrix} -0.03 & 0.10 & 0.06 & 0.13 \end{bmatrix}$$

In chapter 4 prediction models fitted from different $\boldsymbol{g}$ will be compared on real data.

2.0.1 Doc2Vec

An alternative approach to W2V for performing supervised learning is Doc2Vec (D2V), also known as Distributed Memory Model of Paragraph Vectors (PV-DM). It is an extension of the W2V model that fits a vector representation for an entire document besides the vector representations of each word. Thus, D2V is a natural competitor of our method, as it already provides a representation for each text. It uses the same neural network structure of W2V to achieve a vector representation of a collection of words. This is a newer method that has achieved good results in sentiment analysis (LE; MIKOLOV, 2014). We will also report the results of this method in chapter 4.

CHAPTER

3

PREDICTION MODELS

Once a vector representation has been obtained for a text, we can focus on the task of prediction.

In this section we present the prediction methods used throughout this dissertation. All of them treat each element of the vector representation as an explanatory variable of the text. More precisely, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be the observed data, where X_i is the vector representation and Y_i is the response variable of the i-th text. X_i is used as the input of the prediction models and Y_i as the output.

3.0.1 LASSO

The Standard Linear Regression Model (SLRM) is based on a simple concept and can have great results. Nonetheless, too many variables can lead to an overfit, meaning that the regression will be extremely accurate for the data used in training the model in detriment of prediction accuracy on new data (TIBSHIRANI, 1996).

The Least Absolute Shrinkage and Selection Operator (LASSO) method tackles this issue by shrinking or setting to zero some coefficients by constraining their L_1 norm:

$$\arg\min_{\beta_0,\beta} \frac{1}{n} \sum_{i=1}^{n} l(y_i, \beta_0 + x_i^T \beta) + \lambda ||\beta||_1 \tag{3.1}$$

where l is the negative log-likelihood and λ is a tuning parameter that controls the strength of the penalization. When $\lambda = 0$, we end up with the SLRM.

As a result of the constraint, the bias is slightly augmented in exchange of lowering the variance of the estimated coefficients. Therefore the prediction accuracy is enhanced.

All datasets used in this dissertation will be considered to have binary labels, so the

logistic regression will be used. Equation 3.1 will then assume the form:

$$\arg\min_{\beta_0,\beta} -\frac{1}{n}\sum_{i=1}^{n}\left[y_i * (\beta_0 + x_i^T\beta) - \log(1+\exp(\beta_0 + x_i^T\beta))\right] + \lambda||\beta||_1$$

and cross-validation will be used to select λ.

3.0.2 XGB

Decision trees are known for their easy interpretation and poor results in comparison with other methods such as LASSO. Ensemble models of trees have great results on that front (JAMES *et al.*, 2014).

Boosting is a ensemble technique used in different statistical models that is constructed through sequential fittings. Each step uses an updated value that captures the error of the previous fitting, $\tilde{y}$, instead of using the raw response each iteration.

This method forces the data to be learned slowly through weak models such as shallow trees. The chance of overfitting the data is diminished by doing so. Nevertheless, there is still a chance of that happening if the total amount of iterations, M, is too high (JAMES *et al.*, 2014). This tuning parameter is selected with cross-validation.

In particular, XGB is a boosting technique that gives state-of-the-art results while managing computational constraints well (CHEN; GUESTRIN, 2016).

When dealing with classification problems, the XGB uses the negative binomial log-likelihood as the loss function:

$$L(y,F) = \log(1+\exp(-2yF(x))), \quad y \in -1,1$$

where the parameter $F(x)$ is a function of log odds ratio at each iteration,

$$F(x) = \frac{1}{2}\log\left[\frac{P(Y=1|x)}{P(Y=-1|x)}\right].$$

The conditional probabilities can then be estimated once $F(x)$ has been fitted by

$$p_+(\mathbf{x}) = \hat{P}(Y=1|\mathbf{x}) = (1+\exp(-2\hat{F}(\mathbf{x})))^{-1}$$
$$p_-(\mathbf{x}) = \hat{P}(Y=-1|\mathbf{x}) = (1+\exp(2\hat{F}(\mathbf{x})))^{-1}$$

as well as the probabilistic classifier

$$\hat{y}(\mathbf{x}) = 2\cdot 1[c(-1,1)p_+(\mathbf{x}) > c(1,-1)p_-(\mathbf{x})] - 1,$$

where $c(\hat{y},y)$ is the cost associated with predicting y as $\hat{y}$.

More details on this boosting algorithm can be read on Friedman (2001).

CHAPTER

4

APPLICATIONS

This chapter shows the results of the models when applied to different datasets:

- Judicial Process: prediction of the final ruling of the judge based of the judicial process
- Amazon Fine Food Reviews: prediction of the grade given to a product through the text review
- Initial Petition: prediction of the final ruling of the judge based of the initial petition

Most studies of vector representation of texts use in their prediction models only the mean statistic on W2V or extensions of W2V such as D2V (O'Sullivan; Beel, 2019) In the following sections it is shown that adding other simple summary statistics improves the performance of the models. All D2V feature vectors have dimension of 100 and the AUC are compared through the DeLong test ($\alpha = 5\%$) with Bonferroni Correction.

The entire corpus is used in fitting the word embeddings because there is no intention to validate the W2V results in this dissertation, only its usage in prediction models. The division between training and testing when fitting the models is done by sampling with a 50% probability that each text belongs to the training set. The same training/testing set division of a dataset is used for all models.

4.0.1 Judicial Process

This dataset has almost 22k observations. The texts are the summary of a judicial process with the final ruling given by the judge. The sample is reasonably balanced as shown in figure 2.

Besides the mean, the standard deviation (sd), skewness (skew), kurtosis (kurt) and all hundred quantiles (quant) of the corpus were used in fitting the models. In figures 3 and 4 we can see that the prediction models that include all statistics have their performances significantly improve.

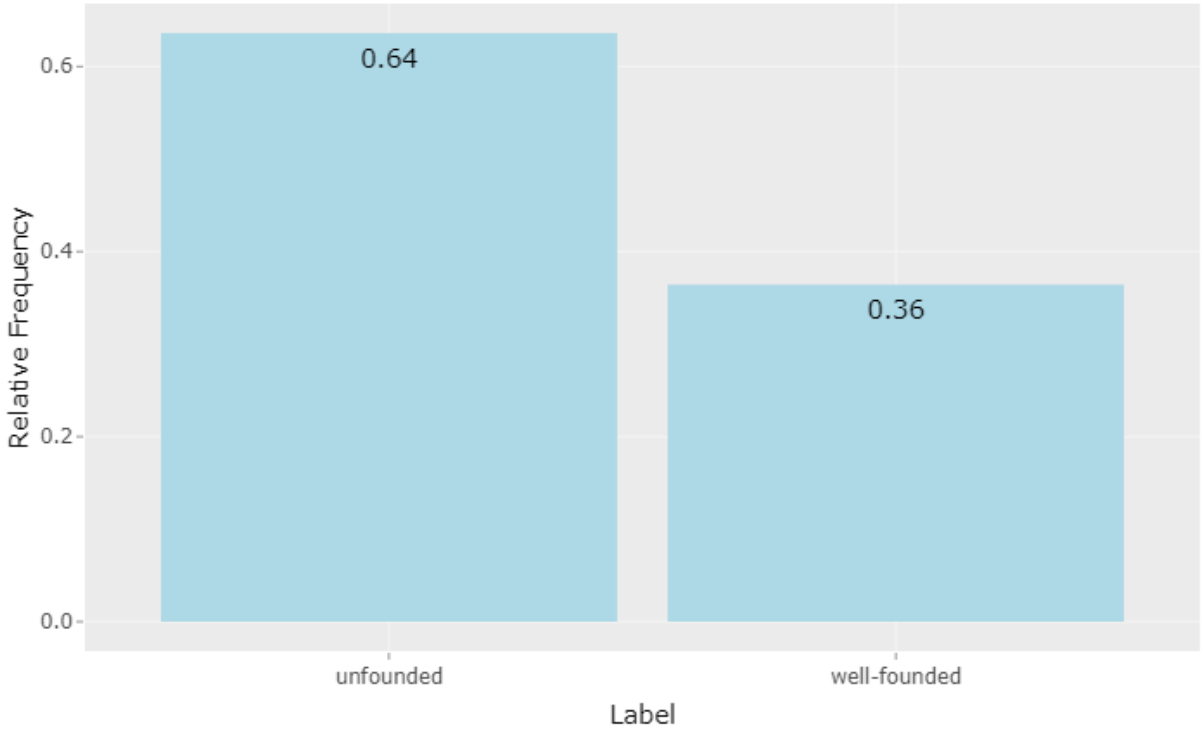

Figure 2 – Barplot of the relative frequencies of the labels on the Judicial Process dataset.

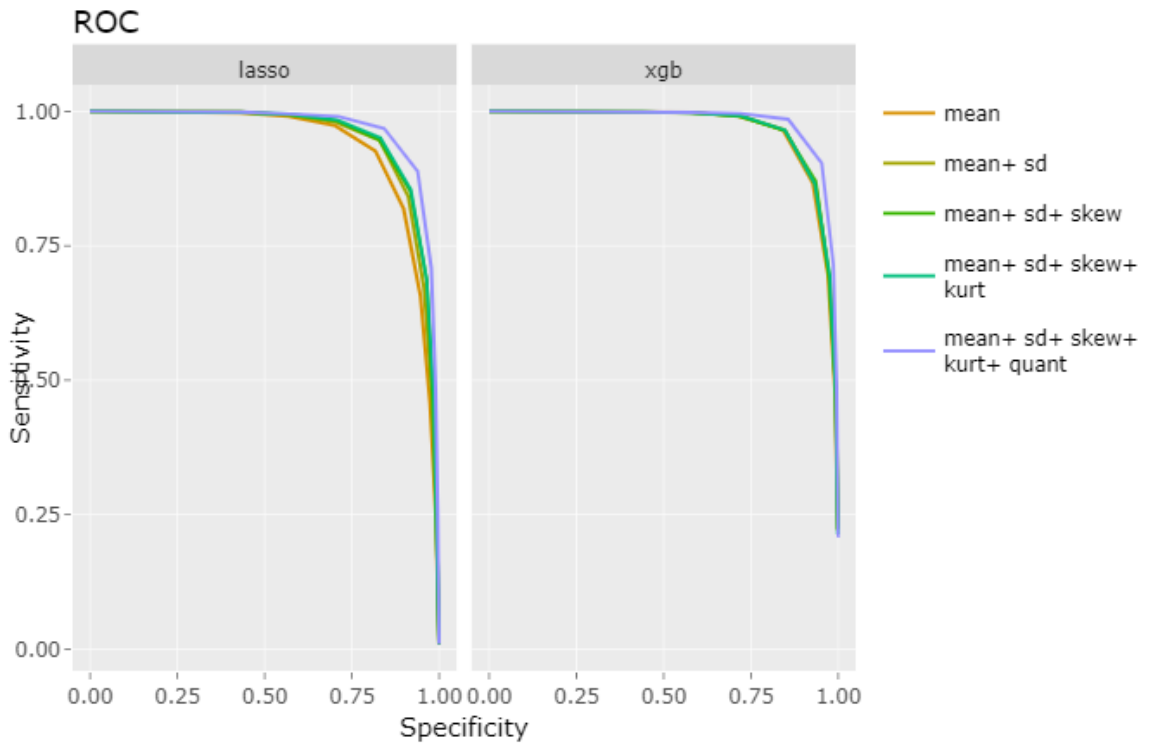

Figure 3 – ROC Curve of the lasso and xgb models fitted with different summary statistics on the Judicial Process dataset.

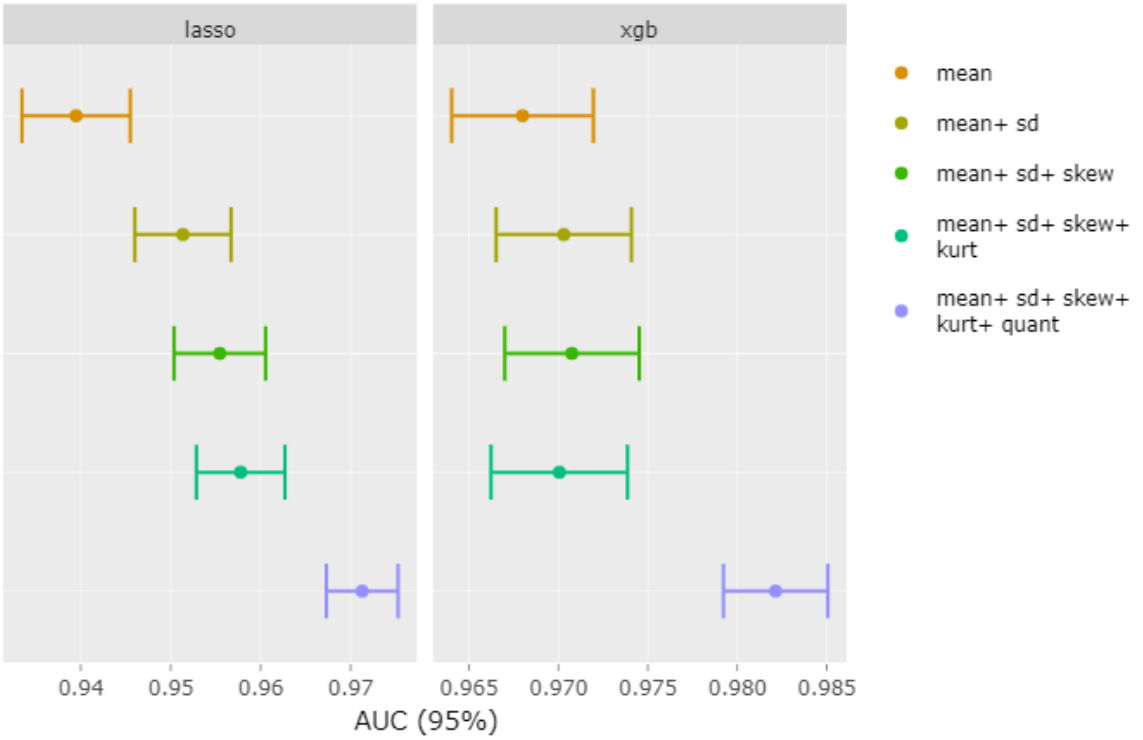

Figure 4 – AUC of the lasso and xgb models fitted with different summary statistics on the Judicial Process dataset.

Although the gain from using all quantiles is desirable, it is computationally expensive. The information matrix used as input in the models will have $101 * d$ columns added. Thus, we test three ways of overcoming this issue and their motivations follow in the text below:

- [tailquant] Filtering of the statistics in $\boldsymbol{g}$: only the moments and the quantiles below 0.05 and above 0.95 are kept in $\boldsymbol{g}$.
- [unifquant] Filtering of the statistics in $\boldsymbol{g}$: only quantiles multiple of .05 are kept in $\boldsymbol{g}$.
- [pca_dimx20] Application of principle components analysis to the information matrix and using only $20 * d$ most important axis.

Exploring the feature importance of the XGB model, tail quantiles, such as below 0.05 and above 0.95 seem to have a big impact on the model (Figure 5). Hence, we test how well the models do with only those [tailquant]. One should notice that tail quantiles capture information about outliers that the other moments do not.

Another possibility is to filter uniformly spaced quantiles, such as the multiples of .05 [unifquant], which reduces the input matrix to $21 * d$ columns. The reasoning behind this method is to keep information about the shape of the density function in the model. An argument in favor of using uniformly spaced quantiles instead of the tail quantiles is that the later was thought out based off the results of the XGB model while the first follows a more general principle, so it is

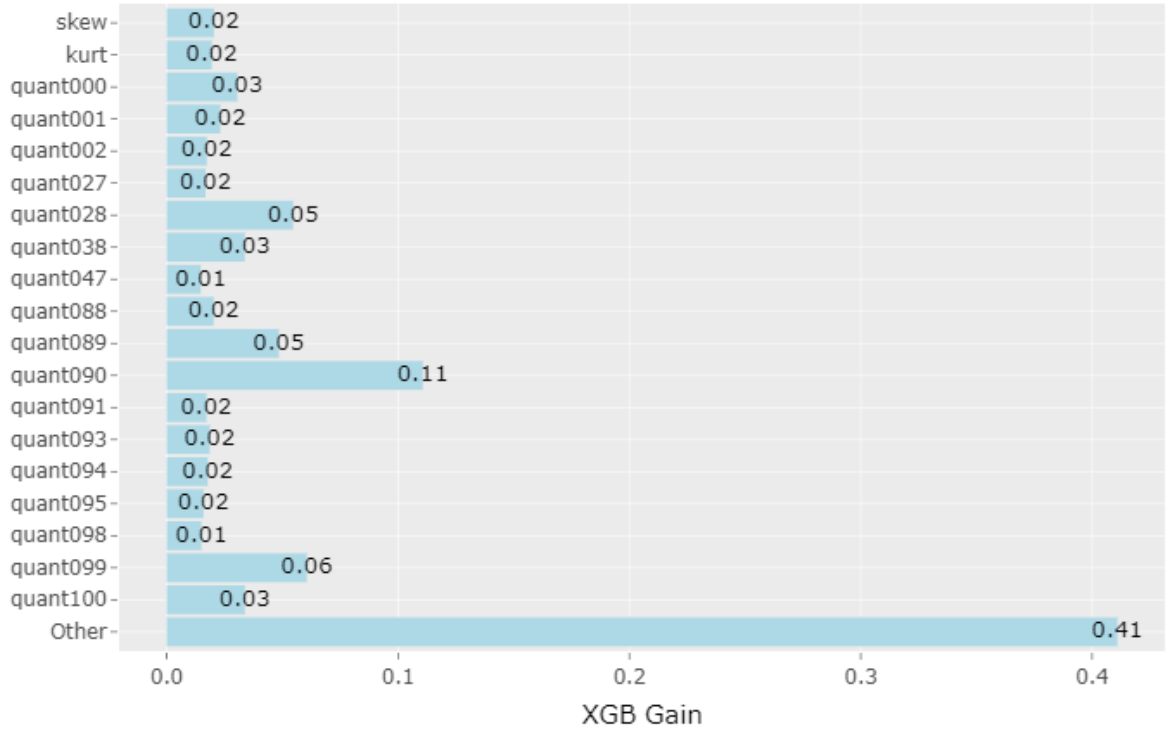

Figure 5 – Feature importance of the XGB method when all statistics are applied on the Judicial Process dataset. The category "Other" is the sum of all statistics that weren not in the top 15 most important statistics.

expected that uniformly spaced quantiles will have better performance than tail quantiles in most predicting models.

Moving away from the idea of feature selection, we apply Principle Components Analysis (PCA) to the information matrix as an alternative and keep only $20 * d$ dimensions in the model (pca_dimx20). This feature extraction technique applies a transformation to the information matrix that maximizes the variance of the resulting matrix, whose axes are also mutually uncorrelated.

Reducing the number of predictive variables did not significantly worsen the models performance (Figures 6 and 7) when compared to the models that use all quantiles. Only pca_-dimx20 had poor results when applied to XGB model, being as effective as the one that uses only the mean statistic. Surprisingly, the D2V method had the worst performance in both models.

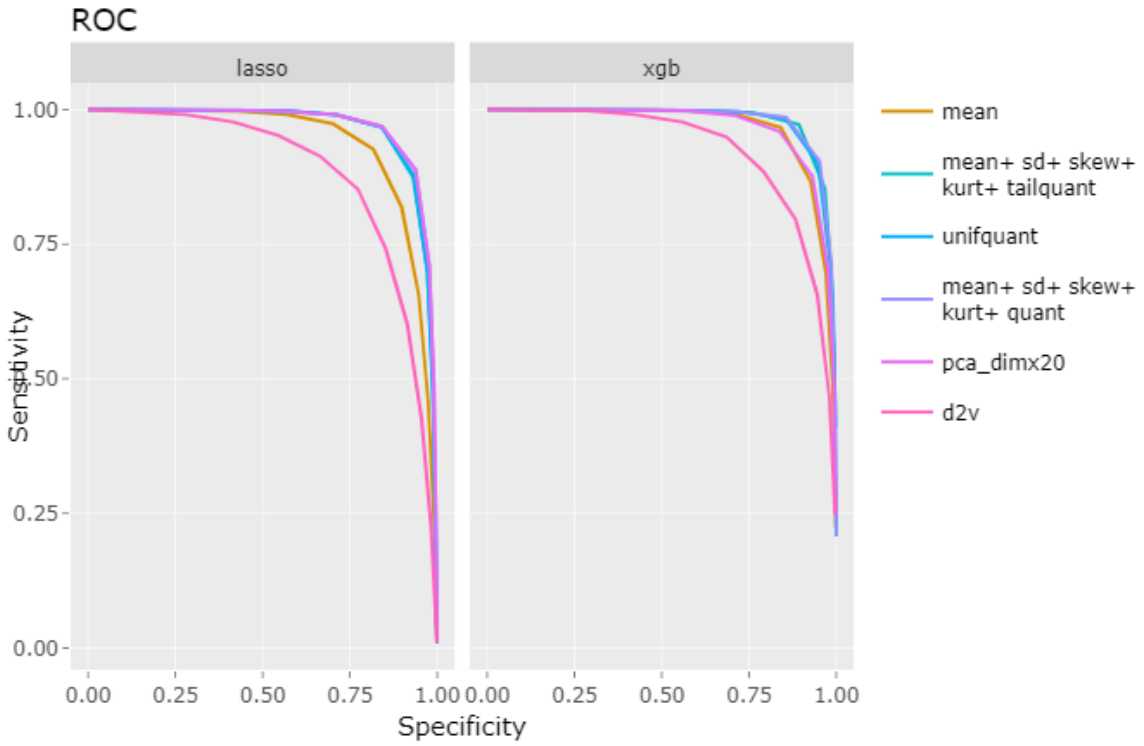

Figure 6 – ROC curves comparing models with mean statistic only, all statistics, the ones with filtered quantiles and doc2vec on the Judicial Process dataset.

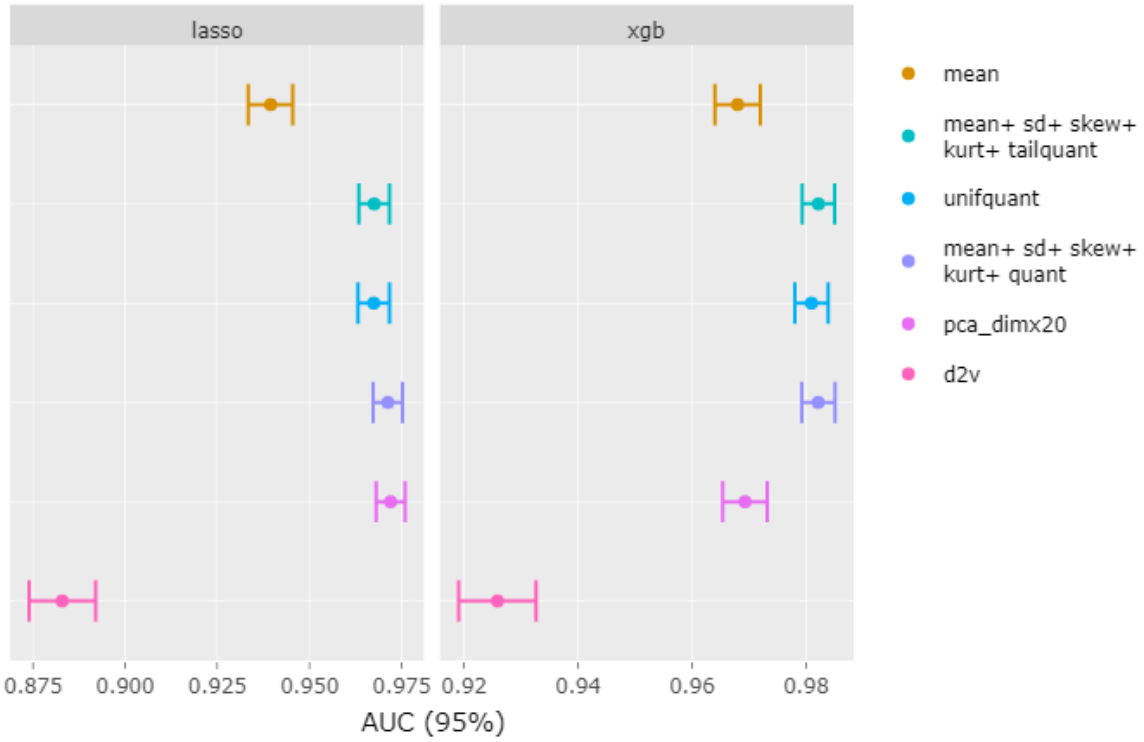

Figure 7 – AUC comparing models with mean statistic only, all statistics, the ones with filtered quantiles and doc2vec on the Judicial Process dataset.

4.0.2 Amazon Fine Food Reviews

The Amazon dataset has more than 500k observations. The texts are the reviews written by the buyers of a product. Each review is associated with a score between 1 and 5. In this analysis we consider that reviews with score greater or equal to 4 are classified as "good" and otherwise "bad".

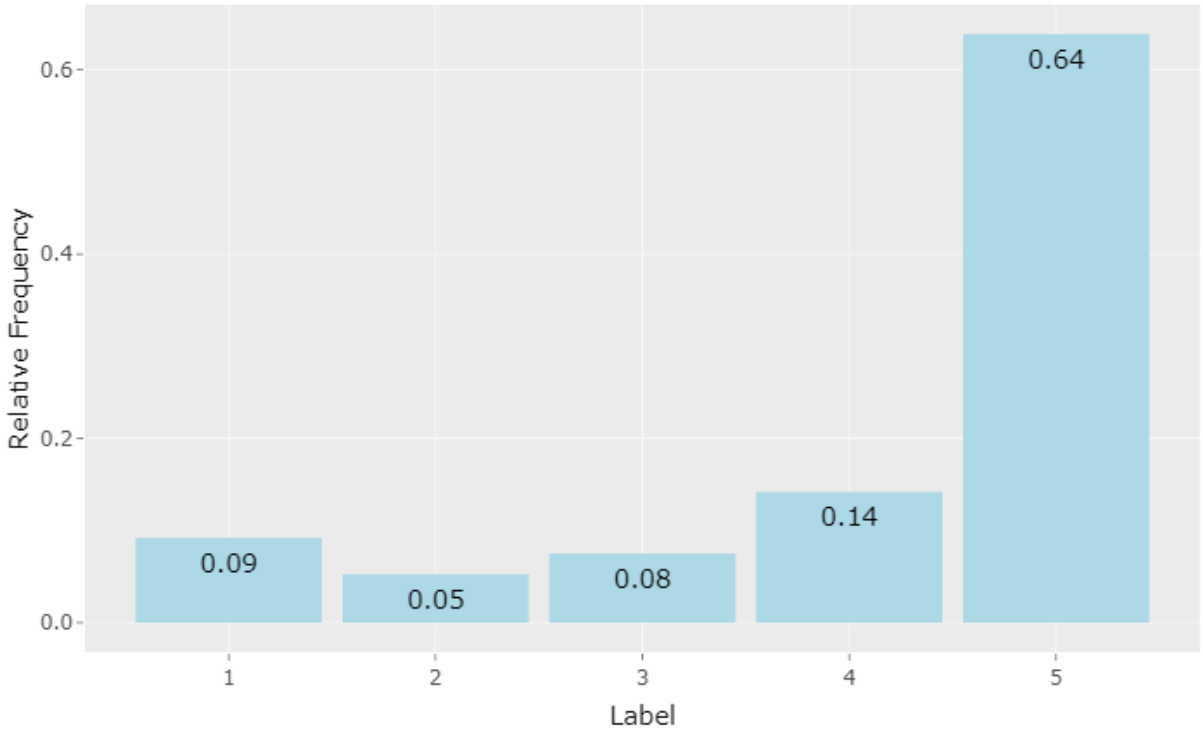

Figure 8 – Barplot of the relative frequencies of the labels on the Amazon dataset.

Classifying only the score 5 as "good" would be better in terms of balancing the sample, as seen in figure 8. But it is not uncommon in customer satisfaction analysis to group 4 and 5 as "good" and the fitted models had better performance like that.

Due to computational restraints, it was not possible to fit neither the model with all quantiles nor pca_dimx20 in this dataset. As seen in the previous section, using the tailquant or unifquant filters are ways to get models with better performance than using only the moments while shrinking the vector representation with all the summary statistics significantly. So in this dataset we only compare models using the mean statistic, filtered quantiles and D2V.

In Figures 9 and 10 we observe that adding statistics to the vector representation significantly enhances the performance of the prediction models and that even the mean statistic alone outperforms D2V. Unexpectedly, the unifquant method has better performance than tailquant with XGB model but not with the lasso model in this dataset.

Figures 13 and 14 show the number of non-zero coefficients of each statistic in the lasso model and Figures 11 and 12 show the feature importance of each statistic in the XGB model

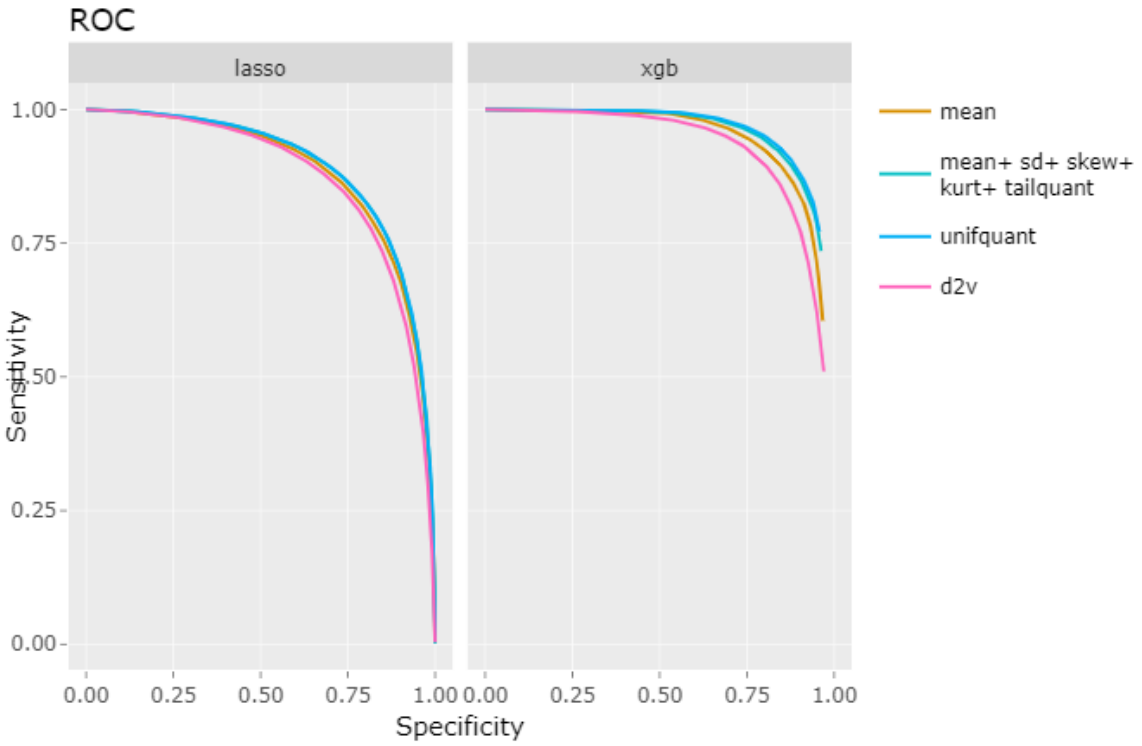

Figure 9 – ROC Curve of of the lasso and xgb models fitted with different summary statistics on the Amazon dataset.

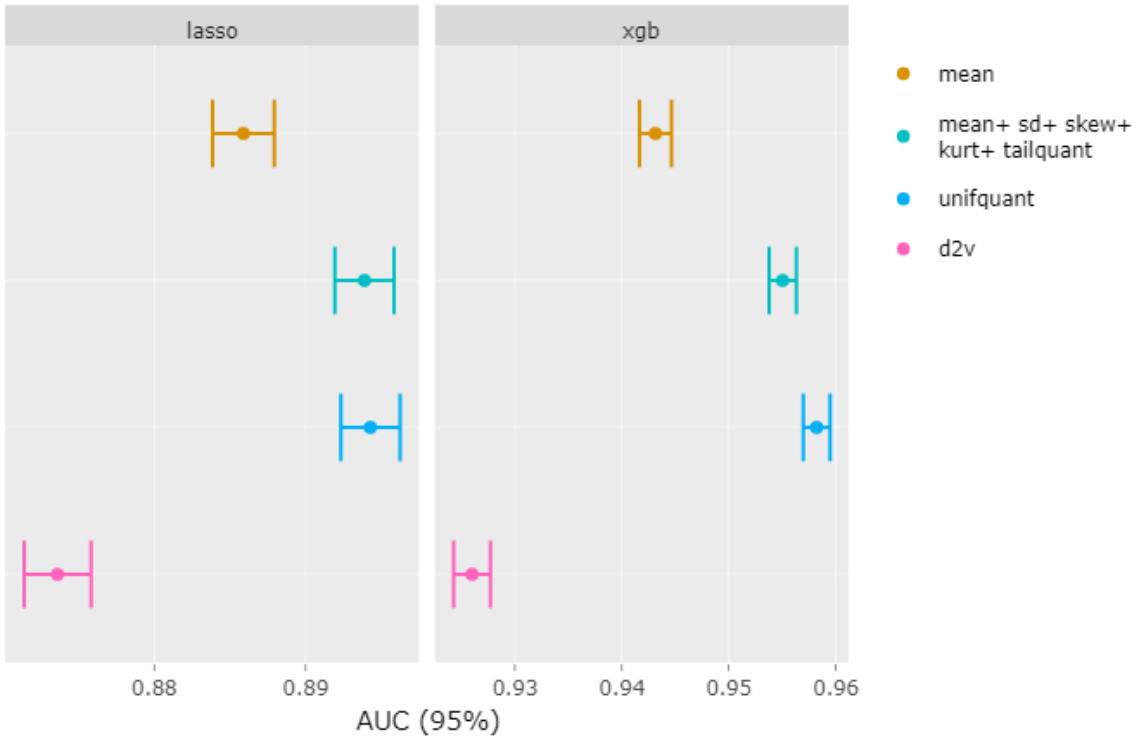

Figure 10 – AUC of the lasso and xgb models fitted with different summary statistics on the Amazon dataset.

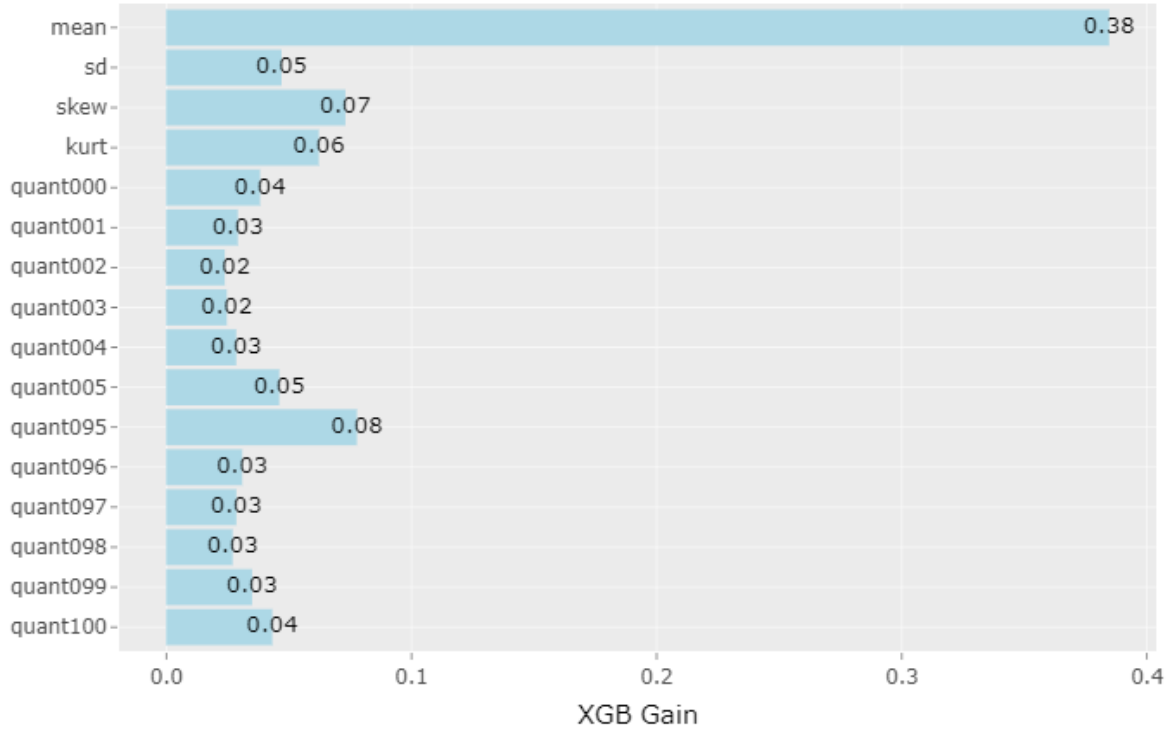

Figure 11 – Feature importance of the XGB method when all moments and the tail quantiles are applied on the Amazon dataset.

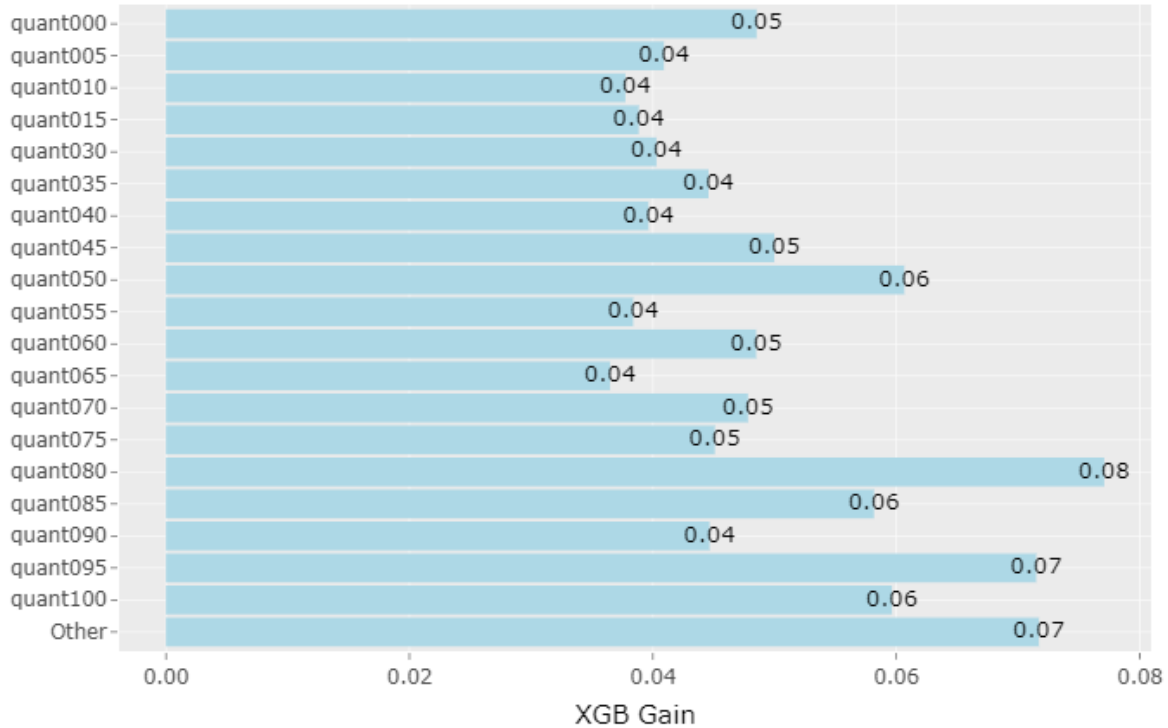

Figure 12 – Feature importance of the XGB method when uniformly spaced quantiles are applied on the Amazon dataset.

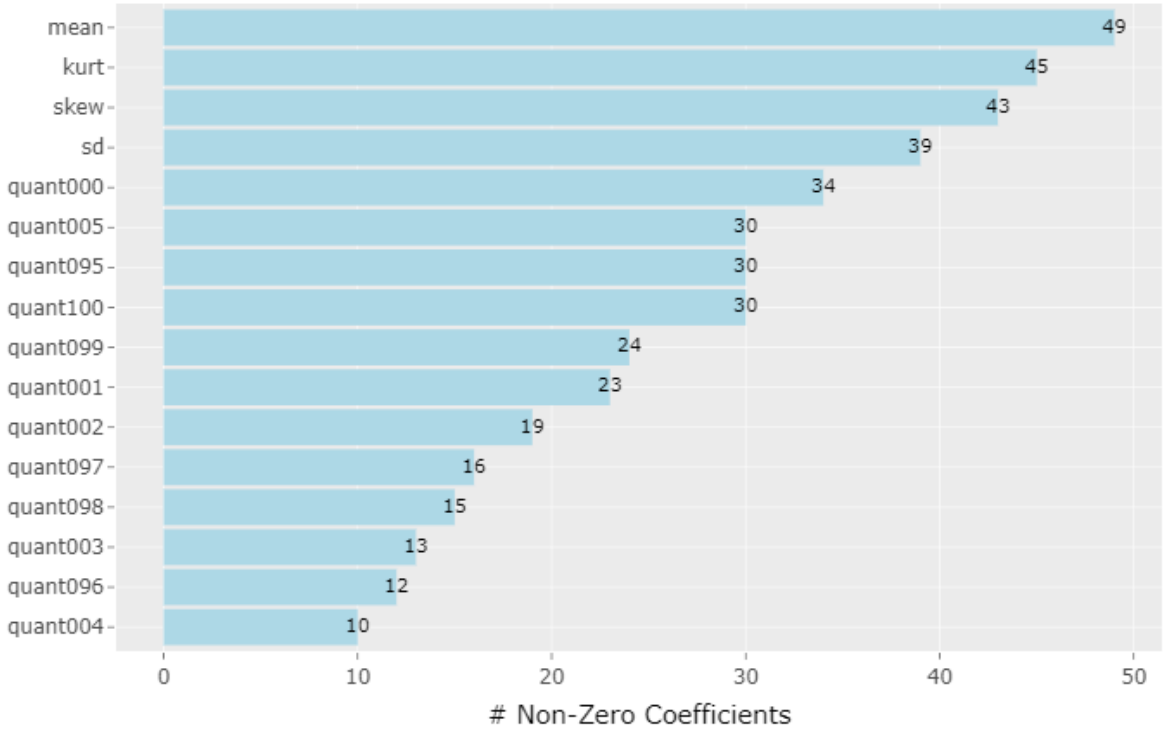

Figure 13 – Number of non-zero coefficients of the lasso method when tail quantiles are applied on the Amazon dataset.

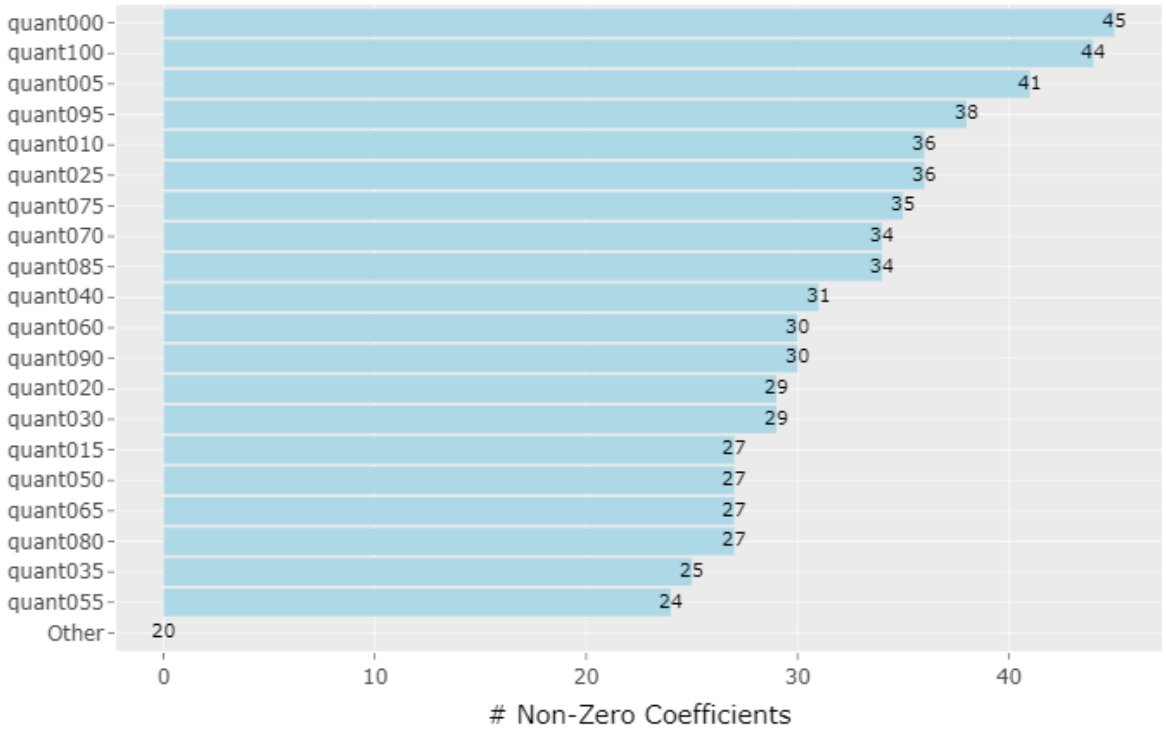

Figure 14 – Number of non-zero coefficients of the lasso method when uniformly spaced quantiles are applied on the Amazon dataset.

when filtering methods are applied. Coherently with the observations made about the ROC and AUC measures, these figures indicate that other statistics besides the mean were relevant in fitting the models. And the feature importance seems to be more homogeneous among the statistics when unifquant is used rather than tailquant.

4.0.3 Initial Petition

This dataset has over 147k observations of different kinds from a judicial standpoint. Only processes filed by private individuals against legal entities whose final decision is known and from the categories "denied", "accepted" or "partially accepted" were selected for this analysis. Decisions classified as "denied" will be considered as "unfounded" in this analysis and "well-founded" otherwise. The final dataset has almost 85k observations and is reasonably balanced. The frequency of the decisions can be seen in Figure 15.

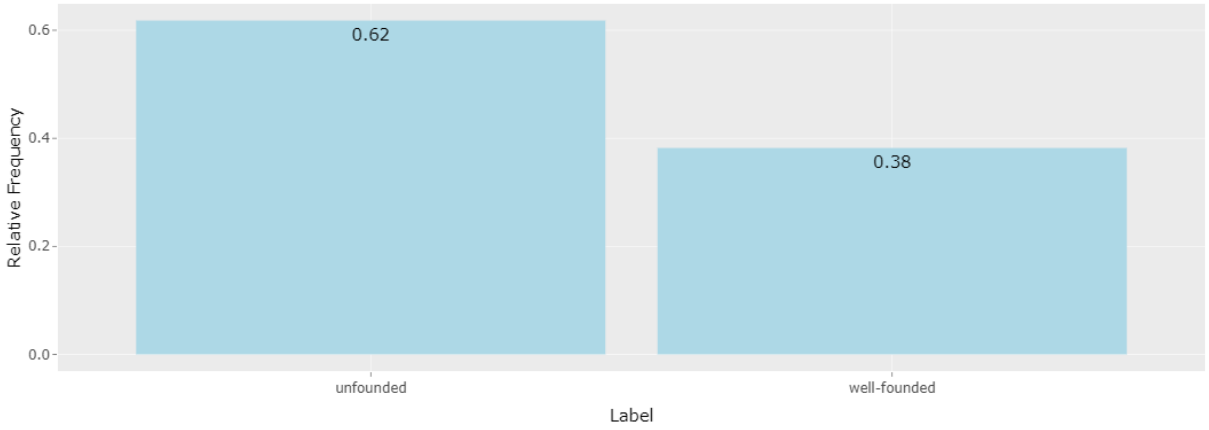

Figure 15 – Barplot of the relative frequencies of the labels on the Initial Petition dataset.

In Figures 16 and 17 we observe once again that adding statistics to the vector representation significantly enhances the performance of the prediction models while D2V has poor results in comparison. The pca_dimx20 has good results with the lasso model but has the same performance as only using the mean statistic with the XGB model. The filtering quantiles methods seem to have better results overall, their AUC confidence intervals overlap with the one that only uses the mean statistic but they also overlap with the one that uses all quantiles when the XGB model is applied.

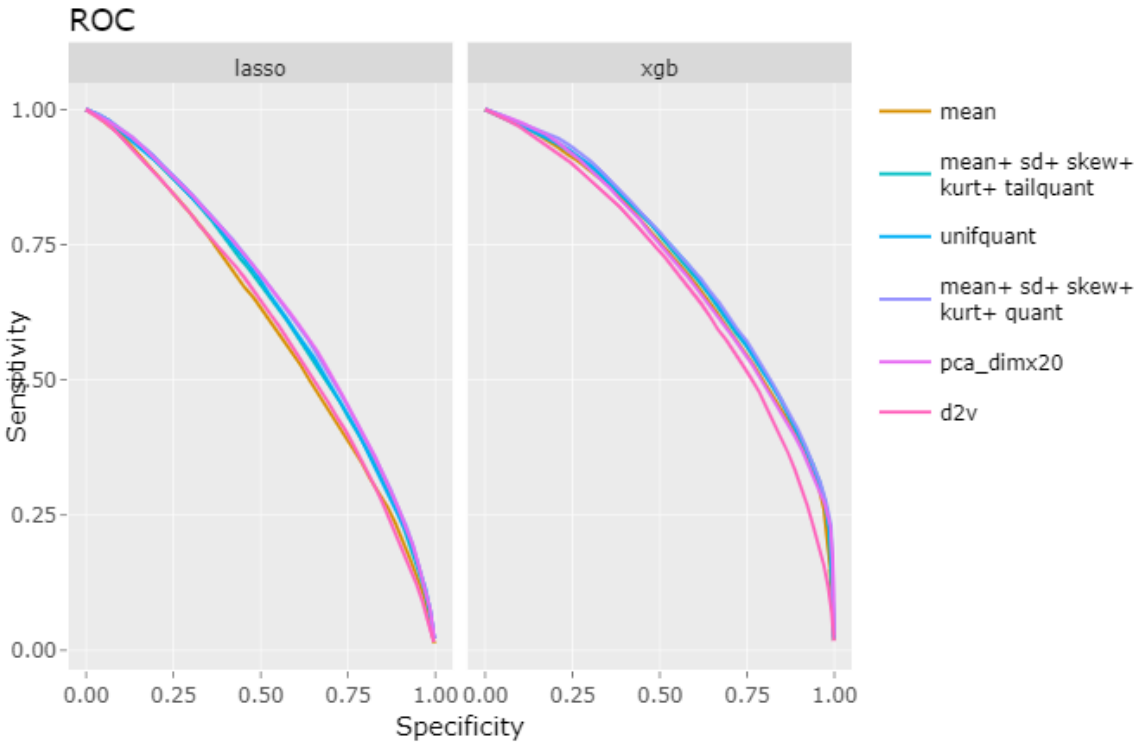

Figure 16 – ROC Curve of of the lasso and xgb models fitted with different summary statistics and the Doc2Vec model on the Initial Petition dataset.

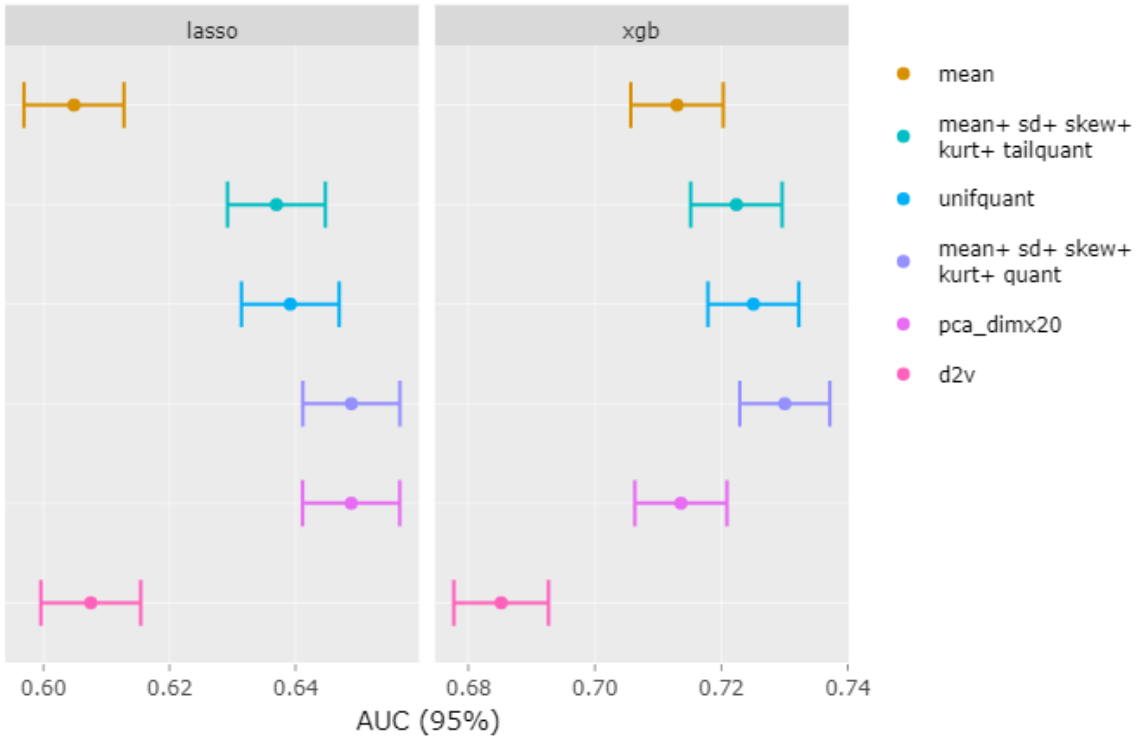

Figure 17 – AUC of the lasso and xgb models fitted with different summary statistics and the Doc2Vec model on the Initial Petition dataset.

CHAPTER

5

CONCLUSION

In this dissertation we proposed to improve prediction models for texts by using other summary statistics of word2vec embeddings besides the mean, such as other moments and quantiles. We observed that adding other summary statistics to the vector representations of texts enhances the AUC of the fitted models. Nonetheless, each function added to the conversion of the matrix representation of the text augments the dimension of the input matrix in $d \cdot n$. Consequently more time and computer power is demanded as functions are added to the prediction model.

We noticed that filtering the quantiles, such as using only the tail or uniformly spaced quantiles instead of all quantiles, shrinks the size in approximately 85% of vector representation while maintaining a significant improvement of the prediction models. This lead to better results than the state-of-the-art doc2vec in all applications.

Future work encouraged by these findings include:

- Use neural networks to deal with the computational constraints met when fitting the prediction models with all quantiles on the Amazon dataset.
- Test other datasets, with numeric response variables where regression models are needed.
- Take into account the embedding's order, such as weighting the embeddings according to the order they appear in the document when applying the summary statistics.

BIBLIOGRAPHY

BANSAL, B.; SRIVASTAVA, S. Sentiment classification of online consumer reviews using word vector representations. **Procedia Computer Science**, v. 132, p. 1147 – 1153, 2018. ISSN 1877-0509. International Conference on Computational Intelligence and Data Science. Available: <http://www.sciencedirect.com/science/article/pii/S1877050918307610>. Citation on page 23.

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. **CoRR**, abs/1603.02754, 2016. Available: <http://arxiv.org/abs/1603.02754>. Citation on page 26.

FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001. Citation on page 26.

GOLDBERG, Y.; LEVY, O. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. **CoRR**, abs/1402.3722, 2014. Available: <http://arxiv.org/abs/1402.3722>. Citations on pages 21, 22, and 23.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning: With Applications in R**. [S.l.]: Springer Publishing Company, Incorporated, 2014. ISBN 1461471370, 9781461471370. Citation on page 26.

JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; MIKOLOV, T. Bag of tricks for efficient text classification. **CoRR**, abs/1607.01759, 2016. Available: <http://arxiv.org/abs/1607.01759>. Citation on page 23.

LE, Q. V.; MIKOLOV, T. Distributed representations of sentences and documents. **CoRR**, abs/1405.4053, 2014. Available: <http://arxiv.org/abs/1405.4053>. Citation on page 24.

LEVY, O.; GOLDBERG, Y.; DAGAN, I. Improving distributional similarity with lessons learned from word embeddings. **TACL**, v. 3, p. 211–225, 2015. Citation on page 23.

LIU, H. Sentiment analysis of citations using word2vec. **CoRR**, abs/1704.00177, 2017. Available: <http://arxiv.org/abs/1704.00177>. Citation on page 23.

O'Sullivan, C.; Beel, J. Predicting the Outcome of Judicial Decisions made by the European Court of Human Rights. **arXiv e-prints**, p. arXiv:1912.10819, Dec 2019. Citation on page 27.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society. Series B (Methodological)**, [Royal Statistical Society, Wiley], v. 58, n. 1, p. 267–288, 1996. ISSN 00359246. Available: <http://www.jstor.org/stable/2346178>. Citation on page 25.


ICMC
USP
SÃO CARLOS