Challenge BBVA Data & Analytics

Here, we answer the proposed questions and we also indicate which notebook contains the detailed code and analysis.

0. Notebooks descriptions

0_Descriptive.ipynb (basic descriptions of the data)
1_Merge_data_transformation (merge and data treatment)
2_Some_Graphs.ipynb (graphs)
3_Topic_Modelling.ipynb (to determine categories of content/recommendation engine -> simple code)
Notebooks with predictions
4_prediction_H20.ipynb (H2O automl)
4_prediction.ipynb (several models with interactions)
4_prediction-(without-interactions).ipynb (several models without interactions)

1. **How would you merge both datasets? What do you observe?**

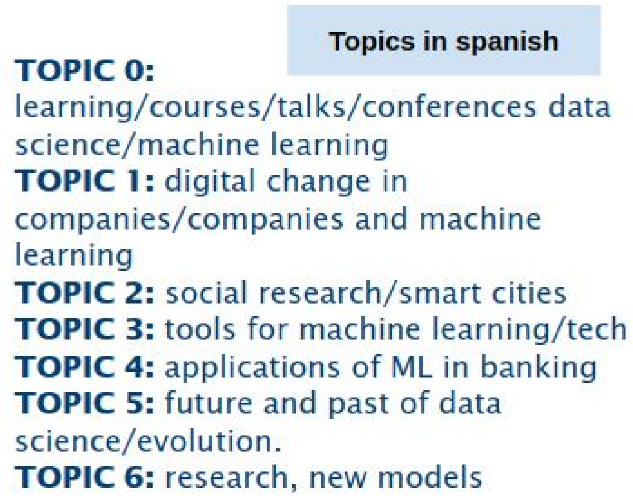
   We do some data text transformation to merge the data, because there is not id for the posts in any of the two datasets. We observe that in the blog_scraping dataset the url contains the title of the posts in english, and this coincides with the variable 'Page' in the blog_analytics dataset. So, after some cleaning we merged the datasets, to obtain a final set of 9.860 rows.

   This final set contains information of valid posts (here, we exclude search pages or visits to other web places not related with posts).

   This analysis is in the notebook_ 1_Merge_data_transformation.

2. **Identify the trending topics in our blog. Besides, are we missing any important topic (regarding Data Science, Product creation, etc.)?**

   Using topic modelling (see presentation) we identify seven different topics. We only consider spanish posts and the content to create the model. Looking at the key words we named the categories.

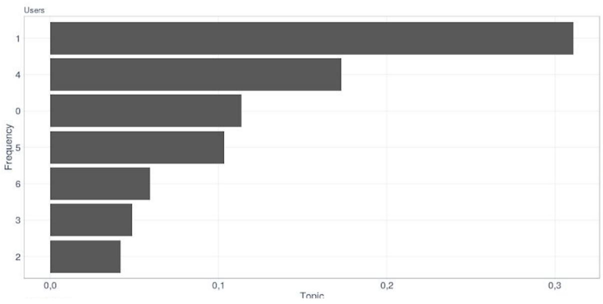


Topics in spanish

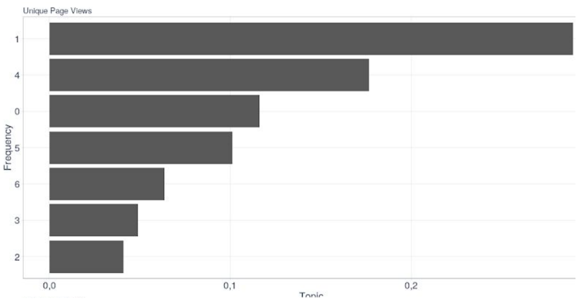
TOPIC 0: learning/courses/talks/conferences data science/machine learning
TOPIC 1: digital change in companies/companies and machine learning
TOPIC 2: social research/smart cities
TOPIC 3: tools for machine learning/tech
TOPIC 4: applications of ML in banking
TOPIC 5: future and past of data science/evolution.
TOPIC 6: research, new models

Looking at the features: users, page unique views, page views, Avg_Time_Page, we can observe que los posts with higher percentages are 1, 4 and 0.

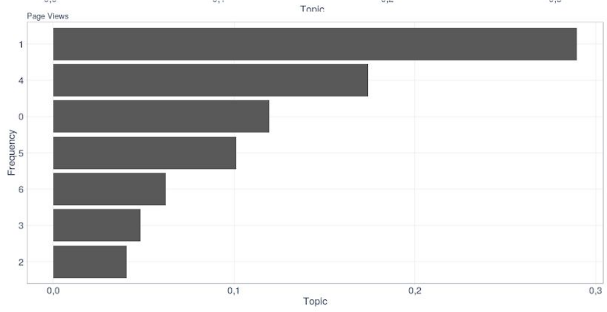
Users



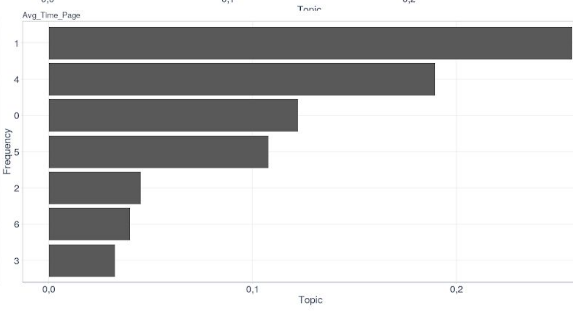
Unique Page View



Page Views



Avg_Time_Page



3. **Build an analytical model that predicts the average time on a given article.**

   Please check the presentation and notebooks 4 and 5 for details.

4. **The bounce rate (defined as https://goo.gl/fywvUr) is quite high in our website. Which actions (analytical or not) would you suggest to reduce it? Can you think of any model that would help us here?**

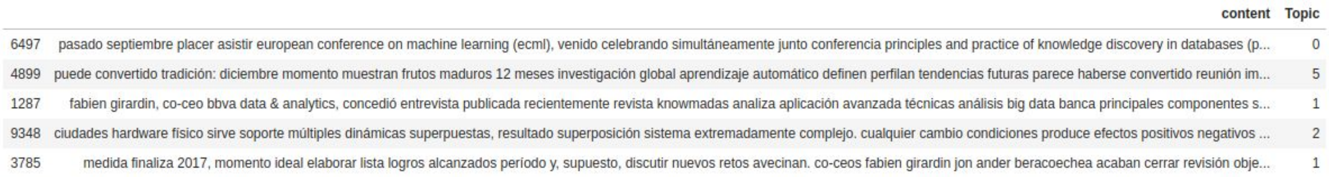
   There are many reasons that could explain why this ratio is high, for example:
   - **Unreadable content:** not attractive captions, lousy format, long and boring text, absence of bullet points, subtitles, highlighting of important ideas, pictures or any other element that catch the eye of the reader.
   - **Existence of pop-up pages:** Emerging pages can be considered by users as irrelevant and annoying, so they may harm user experience.
   - **Absence of a call to action:** the web page has a lot of content but does not send a message to click on, and maintain the navigation of the users in the web page.
   - **Lack of regular and fresh content:** By constantly adding fresh content, you will be gradually building momentum and building trust.
   - **Absent of customers interactions:** customer interactions with the web page can help to increment the engagement. This could be possible if the web page include the possibility of share content, give likes, allow comments, among others. These interactions may be transformed into data used to create recommendation engines based in collaborative filters.
   - **Do not use analytics to enhance content:** use analytics to understand what the users want can help to enhance content, for example: understand which are the top topics, the length and characteristics of the content can help to create more personalized posts. Also, the use of social media metrics (impressions, link clicks, engagement ratio in social networks) can help to prove how successful a content is, how well your social strategy is performing, and ultimately if you will have an impact on your overall business.

   - **Lack of page performance:** if a content takes too much time to appear may create disengagement.
   - **Content not optimized according to device.**
   - **Do not reach the correct target**

In order to improve content and learn what is the best way to take the attention of the user **we can use A/B testing and compare if some improvements in the way of present information** can help to create more engagement and lower the bounce rate. With this methodology it is possible to create several experiments that take into consideration all the points described above.

5. **Create an engine that, given an article, suggests the next one to read.**

We did this through a combination of two methodologies: Non-Negative Matrix Factorization (NMF) method and the calculation of similarities (cosine distances). For example, for the first article in the following list, the next 4 articles would be recommended to the user.



	content	Topic
6497	pasado septiembre placer asistir european conference on machine learning (ecml), venido celebrando simultáneamente junto conferencia principles and practice of knowledge discovery in databases (p...	0
4899	puede convertido tradición: diciembre momento muestran frutos maduros 12 meses investigación global aprendizaje automático definen perfilan tendencias futuras parece haberse convertido reunión im...	5
1287	fabien girardin, co-ceo bbva data & analytics, concedió entrevista publicada recientemente revista knowmadas analiza aplicación avanzada técnicas análisis big data banca principales componentes s...	1
9348	ciudades hardware físico sirve soporte múltiples dinámicas superpuestas, resultado superposición sistema extremadamente complejo. cualquier cambio condiciones produce efectos positivos negativos ...	2
3785	medida finaliza 2017, momento ideal elaborar lista logros alcanzados período y, supuesto, discutir nuevos retos avecinan. co-ceos fabien girardin jon ander beracoechea acaban cerrar revisión obje...	1

This methodology requires testing and validation, but we think this could be a good first approximation. **Check notebook 3 for more details and the presentation for more details.**

6. **Suggest actions to take (as a DS) in order to:**
   ○ maximize the number of readers
   ○ maximize the time spent on the site
   ○ maximize the number of CVs that we receive as a result of a post

   Most of the problems that may generate low engagement are described in point 5. The actions to handle this problems are several depending on the objective that you want to reach.

   - **maximize the number of readers:** create more content based on the success of past contents, share content in social networks and measure their impact, create recommendation engines, allow likes, comments and shares in content, create new content regularly and create experiments (A/B testing) to determine which changes may create a big impact.
   - **maximize the time spent on the site:** create recommendation engines based on collaborative filtering (based in likes) and based on similar content.
   - **maximize the number of CVs that we receive as a result of a post:** use A/B testing to create 'call to action' (links, bottoms, pictures) that highlight the possibility of send a cv or upload it.

7. **Suggest additional data that we should collect in order to measure and improve any of the above business objectives.**

   Metrics from social media: use information about likes, shares, comments, impresiones from social media like facebook, twitter, linkedin to understand with content going viral or have high engagement. Also, to understand which are the best moments to release new content, to do and analysis of followers/influences and detractors, among others.

   Information about events, holidays and weather: this information may be useful to create content that invites to assist to events of interest or talk about past events of interest, and also make recommendations according to weather and holidays. As for example which bootcamps do in summer or which book to read on a rainy day.

Google trends: with this tool we can analyse which topics are trending to create content related to that.

News related with the brand: take into consideration positive and negative news to create content that enhance the positive news and repair any reputational damage in the case of negative reviews.