Mastering Machine Learning for the Data Science and Engineering Club UAB

Laura Planas Simón

Resumen – Existe una necesidad creciente en las empresas tecnológicas de encontrar profesionales especializados en el campo de los datos. Con la finalidad de solventar esta necesidad, desde la universidad se ha creado el "Data Science and Engineering Club UAB", que pretende proporcionar un puente de comunicación entre las empresas de la zona y los estudiantes de la UAB. Dado el nivel de madurez que posee el machine learning hoy en día, el objetivo primordial de este proyecto es dominar diferentes técnicas y adquirir conocimiento suficiente para afrontar cualquier tipo de problema de datos. El conocimiento adquirido en la realización de cada uno de estos proyectos es el que permite el crecimiento y divulgación del club. Por lo tanto, durante la realización del proyecto se han realizado diferentes problemas de machine learning, los cuales se han añadido como contenido didáctico al blog, y se ha participado en una competición real de Kaggle.

Palabras clave – Análisis de datos, Machine Learning, Deep Learning

Abstract – There is a growing need in technology companies to find specialised professionals in the data field. In order to solve this need, the "Data Science and Engineering Club UAB" has been created from the university with the aim of providing a communication bridge between the companies of the area and the UAB data students. Given the level of maturity that machine learning has nowadays, the main goal of this project is to master different techniques and acquire sufficient knowledge to deal with any type of data problem. The knowledge acquired in the realisation of each one of these projects is what allows the growth and disclosure of the club. Therefore, during the realisation of this project, different machine learning problems have been solved, which have been added as didactic content to the blog, and a real Kaggle competition has been attended.

Keywords – Data Analysis, Machine Learning, Deep Learning

1 Introduction

TRAINING machines with data is starting to become a popular way to solve complex business problems. Therefore, career opportunities in the data field are on the rise. As more companies are joining the data bandwagon, higher is the demand for skilled data professionals.

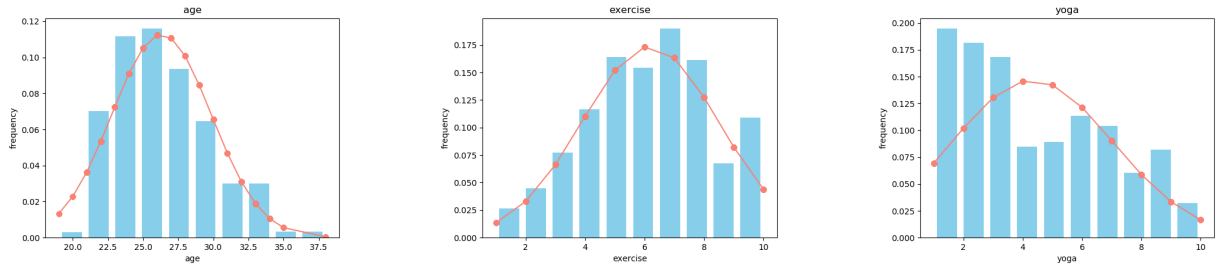
In our area this event is not different either. That is the main reason why the "Data Science and Engineering Club UAB" was created some months ago. This club is serving as a bridge between the area companies and the UAB students in the process of recruitment.

The club was created with the aim of launching data engineering competitions and of creating the DataUAB blog [1], where UAB students can gain visibility towards the companies who are looking for specialised employees. Given the level of maturity that the machine learning has these days, the data professionals need to prove their skills.

The blog is a learning tool where the students can find tutorials explaining theoretical machine learning concepts through practical and real-life data problems. The members of the club can either consult information from other students or publish posts presenting their own results on machine learning problems.

On the other side, the blog may also be very useful for the companies who are looking for qualified professionals in the field, because it can provide a solid prove that a student is capable of solving and explaining any type of data problem. The companies can also propose their own competitions to assess the students on a specific area of interest.

• E-mail de contacte: laura.planassi@e-campus.uab.cat
• Menció realitzada: Computació
• Treball tutoritzat per: Jordi González (Ciències de la computació)
• Curs 2018/19



(a) (b) (c)

Fig. 1: Histograms of some variables with its normal distributions

2 Database selection

A process of selection was performed to choose the datasets for each one of the new blog posts. These datasets were selected by looking for specific data that would allow the use of different machine learning techniques and the solving of some recurrent data problems.

The following datasets have been used:

1. **Speed Dating Experiment** [2]: To predict if a person will receive a call after a date or not. The data from this dataset was gathered from participants in a speed dating experiment conducted by two professors of the Columbia Business School. The aim for this experiment was to obtain experimental data to write a paper showing the gender differences in the mate selection [3]. This dataset has around 8,000 samples and 190 attributes.

2. **World Development Indicators** [4] **and World Happiness Report** [5]: Predicting the happiness score of a country using its economic, social and technological indicators. Two datasets are being used in this problem.

   The first one contains the economic development indicators from the World Bank [6]. There is a list of all the different present indicators in the dataset, along with the number of countries and number of years where the indicator has data [7].

   On the other hand, the second dataset contains a study about the happiness of the different countries on the world, assigning a happiness numerical score to each country.

3. **Cats vs Dogs** [8]: Identifying whether an image contains a cat or a dog. Being one of the most used datasets for beginners in deep learning problems, this dataset contains 12,500 images of cats and 12,500 images of dogs. We will give this problem a twist by trying to obtain the best results only using 1,000 images of each class.

Using the two first datasets, we will perform a classification and a regression problem respectively. These problems will act as an introduction for the majority of other problems of the same type. In the respective blog posts, there will be explained methods to clean and analyse the data, to use and optimize the models and to finally evaluate the results. On the other hand, the last dataset will be used to introduce the real-life problem of having few training samples in a deep learning problem and how to achieve the best results despite the limitation. In the blog we will give a brief introduction to Convolutional Neural Networks and to data augmentation techniques.

3 Blog posts methodology

The posts have been written to add more content, and therefore, spread the use of the blog among other students. These posts are composed by parts of Python code and theoretical descriptions. By showing small pieces of code, the reader can understand the theory behind machine learning concepts through practical examples. Also, the posts follow a very similar structure and are always based on the deep analysis of a specific dataset.

All the posts are written using Python 3.6 [9] in a Jupyter Notebook [10] environment, and then exported into HTML format to be added to the blog.

Down below the key points of the different posts written for the blog are presented, talking about the data analysis, the application of machine learning models and the obtained results.

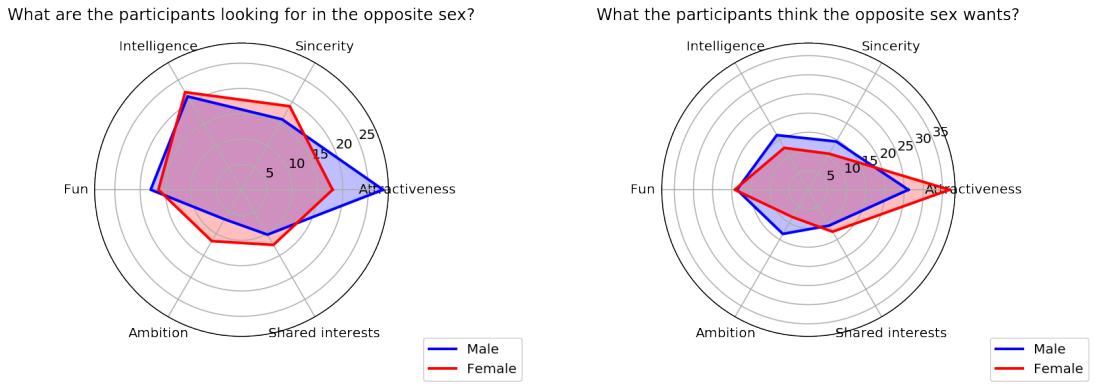
3.1 Speed Dating Experiment

In this post, the Speed Dating Experiment dataset [2] have been used to predict if a person will receive a call or not after a speed date using a classification model. Another of the goals of exploring this dataset is to discover which traits and personal interests make a person more appealing to another at first sight.

The data from this dataset has been analysed to find and understand events from the experiment. Then, a classification model has been applied and lastly, the obtained results have been evaluated.

3.1.1 Data analysis

The first steps explained in the post are related with the cleaning and preprocessing of the data. First of all, there is an explanation of why it is important to deal with the NaN values, either by eliminating them or filling them. The post also shows how to deal with the different types of non-numerical data, such as text attributes and categorical columns. In this specific problem, those variables are not giving information, so the columns containing these types of data have been deleted.



(a) (b)

Fig. 2: Preferences of both genders on the opposite, and general thought on what the opposite gender prefers

Since this dataset has a great number of attributes, it is very difficult to understand exactly how it is and how it is distributed. However, the distribution of the data has to be considered when building a machine learning model. The concept of distribution is, indeed, one of the key topics introduced on this section of the post. In particular, when building a classification model, the more normal the distribution of the data the better, because the model will not be biased to one of the classes.

The best way to observe the distribution of the different attributes is plotting them in bar charts. A line for the normal distribution has also been added to the plot to visually see how normal the data is. Some examples of these plots can be observed in the Figure 1.

Secondly, the post continues by explaining the concept of feature correlation, showing how it can help us detect redundant information that can hinder the classification process. The data contains 6 personal attributes of the participant: Attractiveness, Sincerity, Intelligence, Fun, Ambition and Shared Interests. The participants had to distribute 100 points between the 6 attributes. Since one of the goals is to discover how the personal attributes and the fact of receiving a call are related, the post shows how to calculate the correlation between these variables and how to plot it in a heatmap, as seen in Figure 3.

Lastly, the data has been analysed to get a sense of one of the topics why this experiment was performed: the differences between males and females when looking for a partner. It is important to explore the problem when working with data, because it allows to fully understand the information within the data.

In the dataset there are some columns which contain the importance that each participant gave to the attributes of the possible partner. There is also information about the attributes each gender thinks the opposite wants. The blog shows how to plot this information on a spider chart like the ones in Figure 2, which is a very good way to visualise the general gender preferences on the personal attributes.

3.1.2 Machine learning models

The following section in the post is based on showing different ways to build a classifier model and to evaluate its performance.

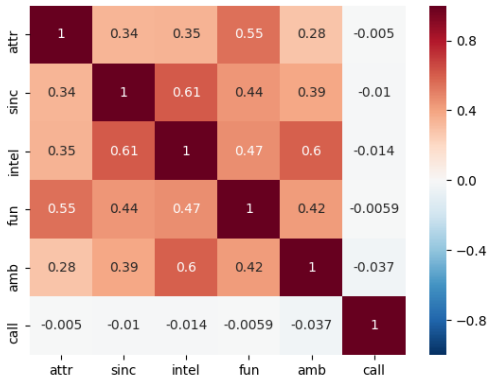


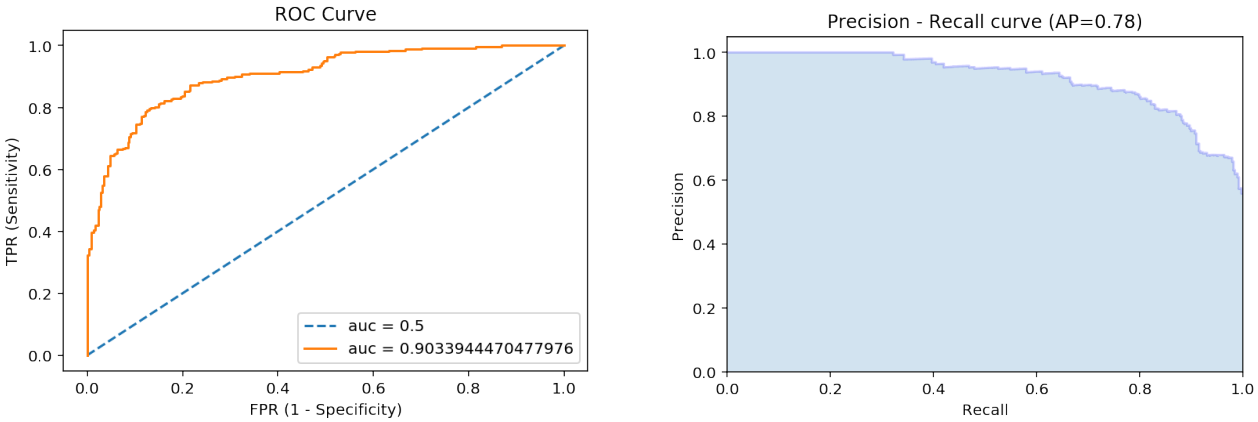
Fig. 3: Correlation heatmap of the personal attributes and the fact of receiving a call

The section begins by introducing the concept of train and test data division. The method used to split the data is very influential on the performance of the model. A good classification model has to ensure the correct performance regardless of the splitting of the data. Therefore, some methods to split the data are explained, specifically the Holdout validation and the K-fold Cross validation.

As well, training a model with consistent data partitions can help to detect and prevent the model from having overfitting. Overfitting happens when a model corresponds so closely to a particular set of data that it fails when fitting additional data or when predicting new observations. This is a crucial concept when building a model, so that's why it is briefly explained in the blog.

After that, a Random Forest model is used, to first fit the training data into the model, and then to use it to predict the class of the validation data. With the obtained predictions the model performance can be evaluated, and precisely, this is the next topic covered in the post.

Beyond the calculation of the accuracy, there are different ways in which the performance of a classifier can be measured. Specifically, the post explains how to calculate and plot the ROC curve and the Precision - Recall curve. With the ROC curve the AUC (Area Under the Curve) has been also calculated because it says the probability of



(a) ROC Curve

(b) Precision - Recall curve

Fig. 4: Evaluation results of the regression model

correct ranking of a random "positive"-"negative" pair. In the Precision - Recall curve the Average Precision (AP) has been also calculated. These are the most common methods used to evaluate a classification model, and that's why the post gives a little insight on what they represent. The evaluation plots can be seen in the Figure 4.

3.1.3 Results

Finally, the obtained plots can answer some of the raised questions explored in the post.

The average age of the participants of the experiment is between 20 and 30 years old, as seen in the Figure 1a.

Looking at the Figure 2a, a clear conclusion can be extracted. The male participants of the experiment are strongly conditioned by the psychical attractiveness of the female participants. This personal attribute seems to be the most important for men when looking for a partner. On the other side, the female participants seem to be aware of this fact, as it can be seen in the Figure 2b. Women seem to expect the attractiveness being the most desirable attribute in the eyes of men.

On the other hand, the attribute that women assess the most is the Intelligence, although the 6 attributes are even.

However, analysing the Figure 3, there can be observed that at the moment of truth, the decision of calling or not the date is not highly correlated with its personal attributes. In the same figure there can be seen that the attributes are correlated with each other, like for example the Intelligence with the Ambition, or the Attractiveness with the Fun.

On a different note, the performance of the model has been evaluated by using the plots from the Figure 4. Specifically, the ROC curve from the Figure 4a and its 0.9 AUC (Area Under the Curve) are saying that the model distinguishes between a positive and a negative sample the 90% of times.

On the other hand, the Precision - Recall curve from the Figure 4b can be used to analyse the precision-recall trade-off. The obtained Average Precision (AP) is 0.78, so the model is right the 78% of times when predicting a sample as positive.

3.2 World Development Indicators and World Happiness Report

For this post, the two datasets, World Development Indicators [4] and World Happiness Report [5], have been combined to predict the happiness score of each country by only using its economic, social and technological indicators. The prediction has been made with a regression model.

In this problem, the post shows an analysis of the differences between countries in the used indicators, the different applied regression models and their evaluations, visualised in plots.

3.2.1 Data analysis

Just as the previous post, this one starts by explaining good practices in the cleaning and preprocessing of data.

The first step of the problem is the selection of the indicators that are used in the prediction. This dataset contains a vast quantity of indicators, but the selected few ones can give a global vision of a country:

- GDP per capita: measure of a country's economic output that accounts for its number of population. It is obtained dividing the gross domestic product by the total population of a country, and can tell how prosperous a country feels for each of its citizens.
- Infant mortality rate (per 1,000 live births): number of children that does not survive of every 1,000 live births. This indicator can reflect the level of technological and medical development of a country.
- Internet users (of every 100 people): number of people that has access to Internet in a daily basis.
- Mobile phone users (of every 100 people): number of mobile phone subscriptions. This and the previous indicator can reflect the level of technological development of a population.

Secondly, in this problem two different datasets are being joined, so the post explains the process of selecting the necessary columns from each dataset and then how to join

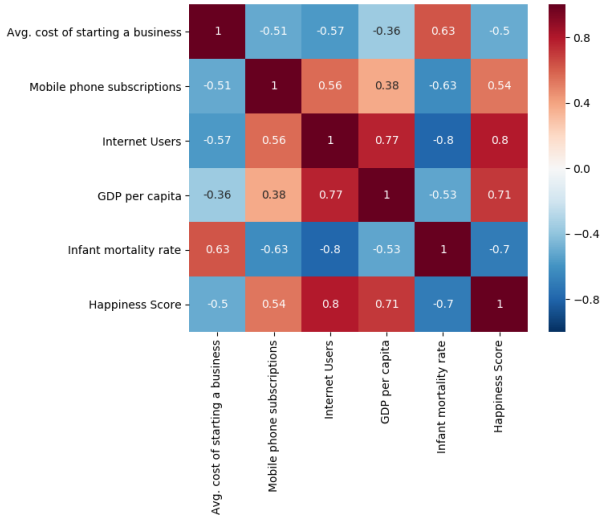


Fig. 5: Correlation heatmap between the indicators and the happiness score

them by the name of the country, that is the common value in the two datasets.

The next analysis performed in the post is the search of which indicators are more correlated with the happiness score. Using the same correlation heat map as the previous post, the Figure 5 is obtained.

Another key point from the post is the fact that the data is highly related to economy and financial concepts. Because a data specialist may not have a lot of knowledge about this or another topics, it is important to perform a deep analysis of the data. It is a good practise to create some plots to visually see the data and better understand it.

For this reason, the post contains a little explanation of how to represent the data from this problem on top of interactive maps created with Python and visualised with HTML. An example of a map showing the indicator "Internet Users" can be seen in the Figure 6. A map is the perfect representation for this problem, because it allows a very easy understanding of the values of the indicators and the happiness scores of the countries.

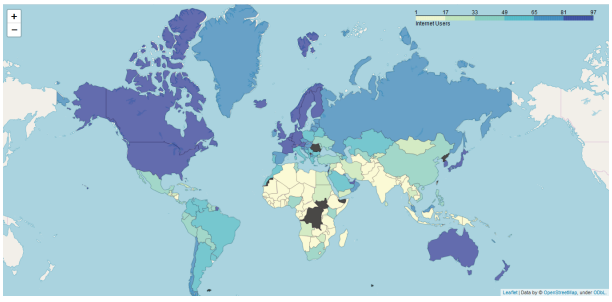


Fig. 6: Interactive HTML map capture showing the Internet Users world distribution

3.2.2 Machine learning models

In this section, different regression models are applied to the data to predict the happiness score of each country. The used models are a linear regression and some polynomial regressions of different degrees. For the polynomial regressions, the function degrees goes from 2 to 4.

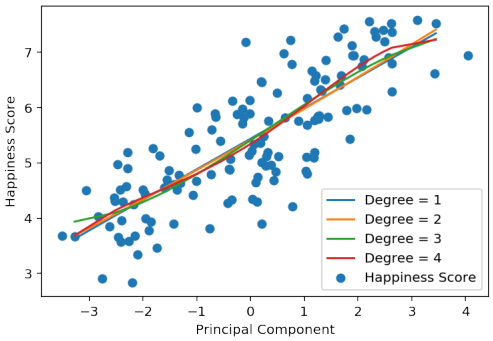


Fig. 7: Regression model visualisation using the 4 indicators at once and after applying a PCA to the data

The 80% of the data has been used to train the model, and the remaining 20% has been used to make predictions and to evaluate its performance.

To show how accurate the predictions are, all the regression functions are plotted in the same chart, on top of the points representing the data. This way, it's easy to observe if the prediction is fitting the data. In the Figure 8, a chart for each one of the used indicators can be observed.

To obtain a visualisation of the regression model using the 4 indicators at once, a PCA had to be applied to the data to reduce the dimensionality. The PCA (Principal Components Analysis) is a statistical procedure used to find the linearly uncorrelated variables from a set of data. Usually, PCA is applied to find the principal components of the data, but can also be used to reduce its dimensions. The visualisation of the data points and the regression models after a PCA being applied can be observed in the Figure 7.

Also, the post shows how to evaluate the performance of a regression model by using the metrics RMSE and R-square. The formula, the piece of code and the theoretical explanation is provided for the reader to fully understand the use of these metrics.

3.2.3 Results

The Figure 5 provides some interesting conclusions about the relation between some aspects of a society with its happiness.

The first noticeable fact is that the number of Internet Users is highly correlated with the happiness score of the country. The GDP per capita, in turn, is also highly correlated with the number of Internet Users. Knowing these two facts, it's reasonable to confirm that the technology level of a population is directly related with its wealth. In the Figure 8c the distribution of the data shows that the happiness and the number of internet users increase together. This fact is the same as the one observed in Figure 8d, where the number of mobile subscriptions goes hand in hand with the happiness.

Furthermore, the GDP per capita indicator itself is also very correlated with the level of happiness, so we can also confirm that the richness of a country is strongly conditioning the happiness of the population. Even knowing that, observing the Figure 8a there can be seen that the regression function stabilises with bigger values of GDP per capita. That means that the GDP per capita reaches a point where can't increase the happiness of the country.

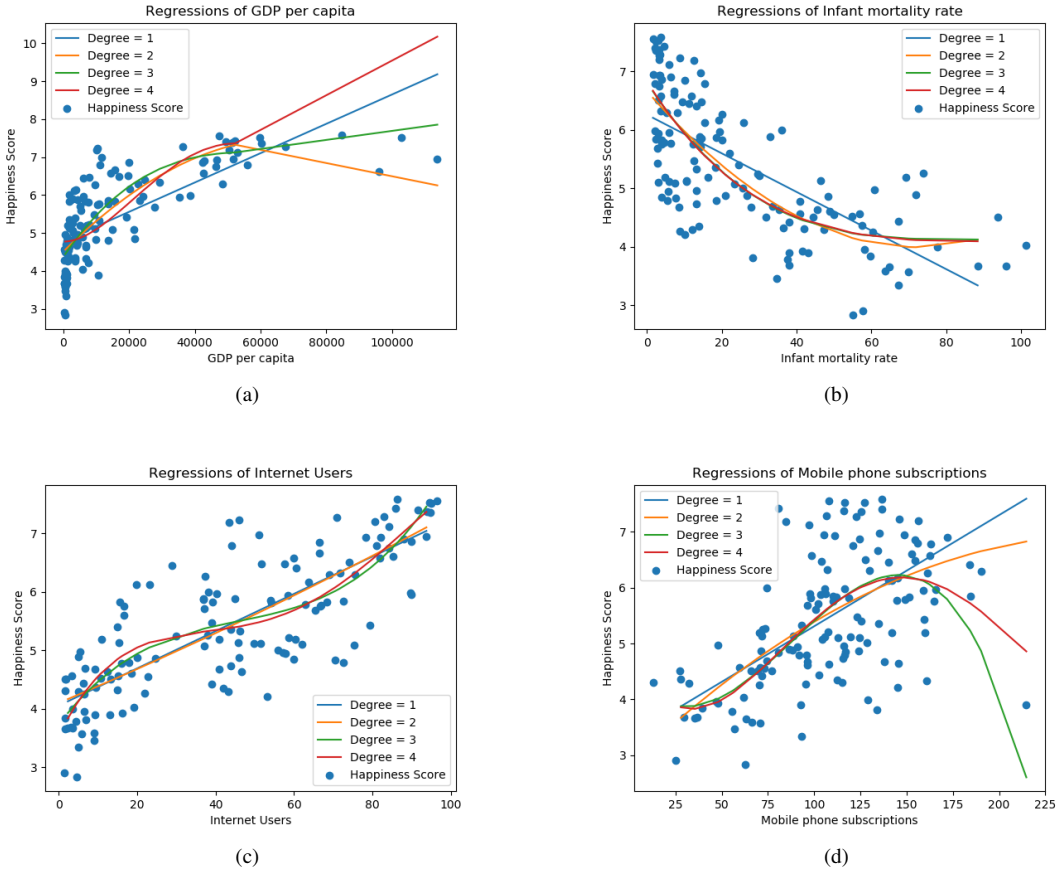


Fig. 8: Regression models visualisation using only 1 indicator at once

On the other hand, the indicator Infant mortality rate is inversely correlated, in general with all the other indicators, and with the happiness score. Also, in the Figure 8b the level of happiness decreases at the same time as the infant mortality rate increases. It's logic to confirm, then, that the impoverished countries have a higher Infant mortality rate caused by a lower technological development, and this situation is causing a decrease in the happiness of the population.

These conclusions can also be extracted in a different way by observing the information in the Figure 6. This map, which shows the Internet Users indicator, can be very useful to quickly see which are the countries where the Internet is widely used and, on the other hand, to see the countries where its use is not widespread. All the indicators can be plotted on the top of a map to see the information in a visual way.

Talking about the performance of the regression models, the R-square of the showed models keeps around a 0.7 value. By looking at the distribution of the data on the charts from the Figure 8, it can be observed that the regression functions are fitting the data properly. There are difficult data distributions for a regression model to fit, as for example the one observed in the Figure 8d.

The more high the degree of the regression polynomial, the better fitted the data. Nonetheless, the increase of the degree of the model can cause overfitting, so this is an aspect that has to be treated carefully to build a correct regression model.

3.3 Dogs vs Cats

In this post the dogs and cats images from the Dogs vs Cats dataset [8] have been used to train a Convolutional Neural Network, to later use the obtained model to predict if the animal from an image is a cat or a dog.

Only 1,000 images have been used for each class, a quantity that is usually insufficient to train an image classification model. This twist in the problem will allow the simulation of a very common real-life situation, which is the lack of training data.

To resolve this problem, this post will speak about data augmentation techniques and other methods to obtain good results although having very few images. The post will also be covering how to build from scratch a CNN with the Keras library.

3.3.1 Data analysis

In this problem the dataset consists of images, so the analysis and preprocessing is completely different.

The first concept introduced in this section of the post is a Generator Function. When working with images, loading the whole dataset into a single variable is not an option, so a generator function is used. A generator function is like a normal Python function, but it behaves as an iterator. The generator functions are very important because they allow a fast access and the possibility of resizing the images to match with the input size of the CNN.

Secondly, to resolve the problem of the lack of data,

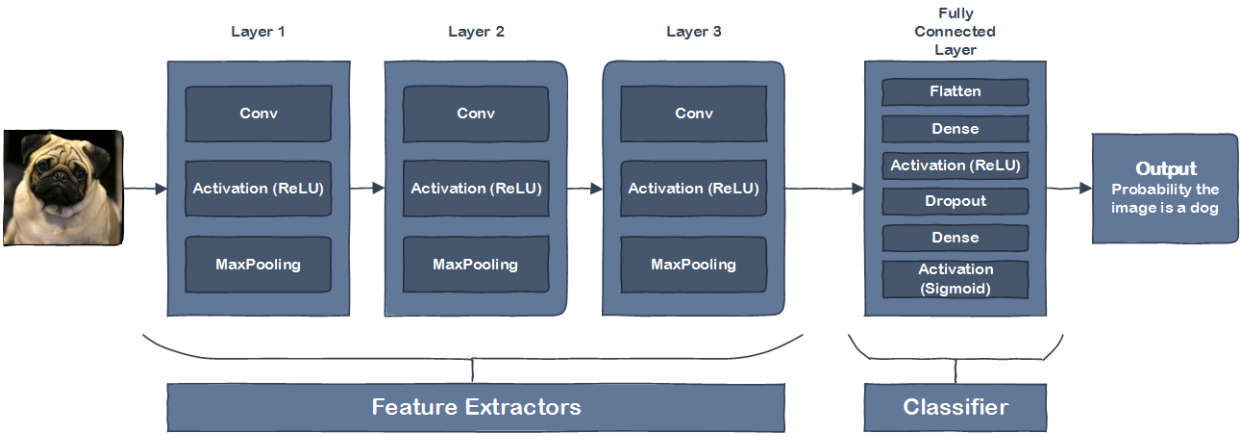
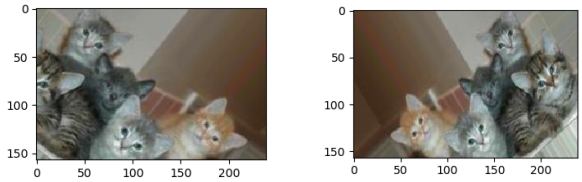


Fig. 9: CNN Architecture



(a) (b)

Fig. 10: Generated images using basic operations

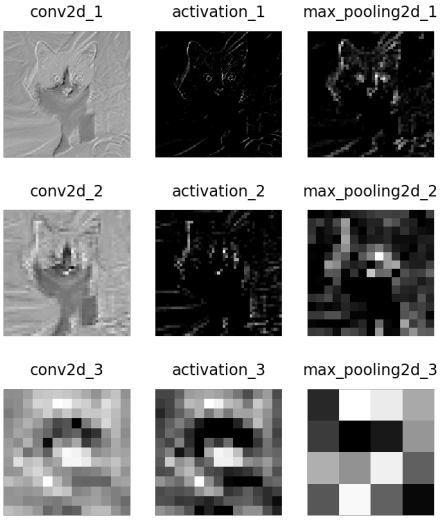


Fig. 11: Convolution layers outputs

image augmentation techniques are introduced. These are methods to artificially increase the number of images of a dataset by altering the original ones with basic operations, such as rotations, zooms, translations and so on. For example, by only flipping all the images we can double the quantity of training data. In the Figure 10 there can be observed some of the generated images by the Image Generator from the Keras library.

3.3.2 Deep learning models

Since in this problem deep learning has been used, this blog contains a theoretical introduction about what is a Convolutional Neural Network and a little tutorial of how build one from scratch with the Keras library.

CNNs are a type of Neural Networks that try to understand images the way the humans do and to extract features from them. A convolution consists in sliding a filter over an image. Instead of looking at an entire image at once, it is more effective to look at smaller pieces of the image [11].

A very basic structure for an image classifier is shown in the Figure 9, which is the CNN that has been built from scratch in this problem. This CNN consists of 3 Convolutional Layers with a ReLU activation function and a MaxPooling filter, and finally a Fully Connected Layer for the classification. The activation function of the output of the net is a Sigmoid function, because it returns a probability between 0 and 1. This probability is the probability of the image being a dog.

After training the net with the 2000 images, the Figure 11 shows the output of the different convolution layers for a specific image. As it can be seen, the two first convolution layers are extracting the face shape of the cat from the image. The convolution from the third layer becomes abstract and less visually interpretable.

On another note, during the evaluation of the model a research was done to find which attributes are the ones that the CNN detects as cat features and dog features. These characteristics can be seen in the Figure 12.

3.3.3 Results

The state of the art for this problem suggests machine classifiers can score above 80% accuracy on this task [12]. This score was beated in the Kaggle competition, since nowadays there are models with an accuracy of a 99%. These models have achieved this accuracy by using more training data not provided in Kaggle and much more complex CNN architectures.

In the own solution the model performs with an 80% of accuracy. Therefore, the state of the art has been achieved by only using the 7% of the available data. For being such a simplistic approach to this typical deep learning problem, the image augmentation techniques have allowed obtaining a very acceptable result.

As said before, the Figure 12 is giving a clear conclusion about what are the features that the model identify as cat or

(a) (b)

Fig. 12: Cats and dogs predictions

dogs characteristics.

In first place, we can see that the good classified cats usually appear looking directly to the camera, with a very bright eyes and with a very characteristic cat shaped ears. These 3 characteristics are precisely the ones that don't appear in the bad classified cat photos. In this photos there usually are sideways cats not looking to the camera, and cats with flatted ears.

On the other hand, in the good classified dog images the dogs have a very different ear shape than cats. The ear shape and the big dog body shape seem to be determinant features for the model to determine the image is a dog. In the bad classified dog images we can observe that the model fails when the dog is more little or hairy, being then more similar to a cat.

4 KAGGLE COMPETITION: TITANIC

The selected Kaggle competition is "Titanic: Machine Learning from Disaster" [13]. By participating in this competition, the final purpose is to create a little guide to be published in the blog explaining the basics of participating in a real data competition. In the post we will explain the own solution to the competition's problem, but most importantly, we will explain the techniques and models used by the users that have achieved the best results.

This competition consists in predicting if a person survived the Titanic disaster knowing some of its attributes, such as the gender, the ticket class or the age. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive the tragedy. Therefore, the challenge on this competition is to find, using machine learning tools, the groups of people that were more likely to live.

4.1 Data Analysis

The first step to understand the problem has been the analysis of the attributes from the database. After plotting a correlation map between the attributes and the fact of surviving, the 3 most important attributes have been analysed: Age, Sex and Ticket Class. The distribution of these variables gives very important information about which was the group of people more likely to live.

The Figure 13 is showing some clear facts about the disaster. First, passengers under 15 years old had a greater chance of survival (seen in Figure 13a). The 15–35 age band had much worse odds and the survival rate was essentially 50:50.

Also, we can affirm that women had priority in the rescue, because in Figure 13b can be seen that survived twice as many women than men.

Finally, the Ticket class was also related with the fact of surviving. The majority of passengers were travelling in 3rd class, but the people who survived the most belonged to the 1st and 2nd class (seen in Figure 13c). We can affirm, then, that richest people had priority during the evacuation.

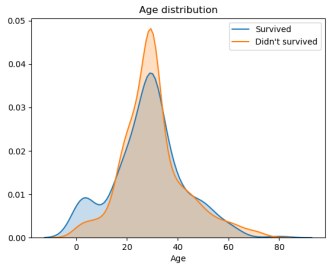
4.2 Machine Learning models

Basically, the own attempt has consisted in preprocessing the data by filling the NaN values with the mean of the columns, and then trying 4 different popular classification models. The 4 used models have been a Decision Tree classifier, a Random Forest classifier, an AdaBoost classifier and a XGBoost classifier.

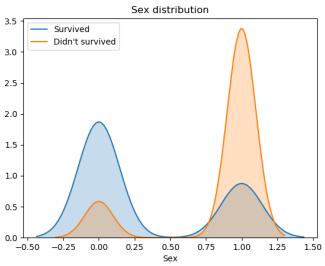
The first 2 models are Ensemble learning methods. The goal of ensemble algorithms is to combine the predictions of several base estimators built with a given learning algorithm in order to improve the robustness over a single estimator. The Random Forest classifier also return the feature importance of the different variables.

In the Figure 14 we can observe that the Fare, the Sex and the Age are the variables that are giving more information to the Random Forest model. This plot can also be useful to drop the variables with less importance, as for example the Embarked attribute.

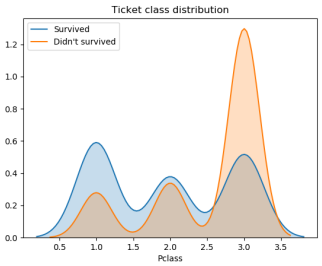
On the other hand, the 2 last models use Boosting techniques, which have recently been rising in Kaggle competitions and other predictive analysis tasks. Boosting trains a series a low performing algorithms, called weak learners, by adjusting the error metric over time. Weak learners are algorithms whose error rate is slightly under 50%.



(a)



(b) 0: women, 1:men



(c) 1: 1st class, 2: 2nd class, 3: 3rd class

Fig. 13: Age, Sex and Ticket class distributions of those who survive and those who didn't survive

4.3 Results

The obtained results in the competition using these models can be seen in the Table 1.

TABLE 1: TITANIC COMPETITION ACCURACY

Model	Score
Decision Tree	0.71770
Random Forest	0.74641
AdaBoost	0.75119
XGBoost	**0.77511**

This approach for the competition is very simple because the final purpose of it is the creation of a beginner's guide for people who have never been in a data competition before. The most basic concepts of a competition are explained, like for example the importance of following the submission format.

Even with such a simple solution for the problem, a good accuracy have been achieved with the different models and we obtained a position in the second quartile of the ranking.

The most common solution used by people on the top of the ranking consists in a better preprocessing of the data, like for example filling the NaN values with a more representative value than the mean, and then doing a Grid Search to obtain the best parameters for the used models.

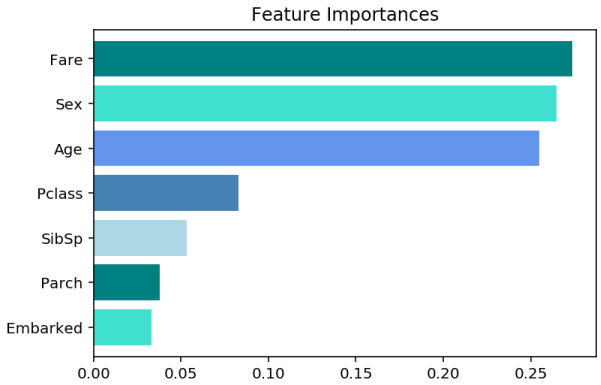


Fig. 14: Titanic feature importances

5 THE DATAUAB BLOG

The DataUAB platform [1] was created some months ago by a previous member of the club. Even if the platform was already created, there is always room for improvement. That's the reason why some aesthetic enhancements have been made to the blog, to make it more appealing to the future students that will use it.

The Jupyter Notebooks and HTML files generated for the blog can be found in the DataUAB Github page [14].

The aesthetic of the blog and the published posts can be observed in the Figure 15, and are the following:

- How to get a second date [15]: post explaining the Speed Dating Experiment problem.
- The secret of happiness [16]: post explaining the World Development Indicators and Happiness report problem
- How to build a powerful image classifier with few images [17]: post explaining the Dogs vs Cats problem.
- A beginner's guide to data competitions [18]: post explaining the Titanic Kaggle competition.

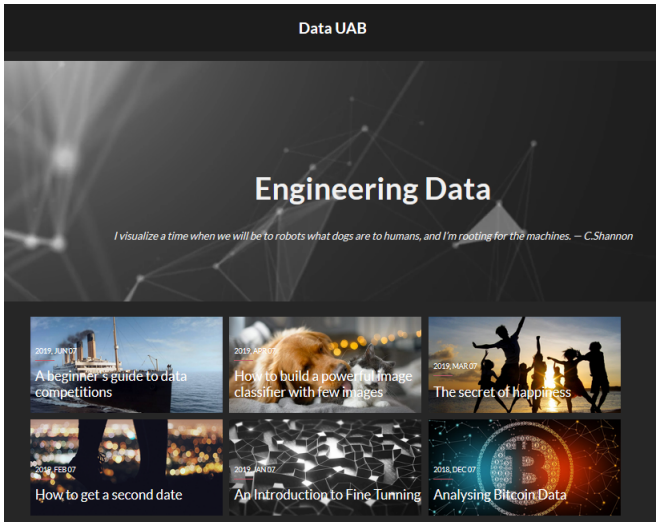


Fig. 15: DataUAB platform and published posts

6 Conclusions

The main goal of this project has been mastering different machine learning techniques and acquire enough knowledge to deal with any kind of data problem. By publishing the acquired knowledge on the blog, the disclosure of the existing "Data Science and Engineering Club UAB" was also achieved.

To summarise all the results obtained in this bachelor project, down below there are some of the conclusions of each problem.

1. In the Speed Dating Experiment, the most desirable attribute for men is the Attractiveness and for women the Intelligence. Even there are clear preferences in the two genders, the psychical attributes are not correlated with the fact of receiving a call after the date.

2. In the World Development Indicators and Happines Report we found that the happiness is highly correlated with the GDP per capita and the Internet Users. Nevertheless, the GDP per capita reaches a point where even increasing, the happiness stabilises. Also, the happiness is inversely correlated with the Infant Mortality.

3. In the Dogs vs Cats problem, the model can predict clearly a cat when it looks directly to the camera and when you can see its ears. The model fails when the cats are sideways. In addition, the model can predict clearly a dog when it's big and has the ears flatted. It fails when the dog is hairy or little.

4. In the Titanic competition, children, woman and people with first class tickets were the ones with more probabilities of surviving the disaster.

Looking into the future, the club will continue growing if we continue promoting it among the students. Each new member of the club will contribute by publishing new posts, participating in new competitions and making new improvements.

Acknowledgements

First of all I would like to thank my tutor, Jordi González, for the guidance he gave me during the last months and for the trust he placed in me since the very beginning of this project. The finalisation of this project and, in fact, this stage of my life, would not have been possible without the unconditional support of my loved ones. Thanks to my parents and my sister for always supporting me in my decisions and helping me to wisely choose the next steps of my life. Thanks to my best friend Clara, for always being there for me no matter what. And thanks to Jaume, my soulmate, for all the unconditional love, the disinterested help and the motivation you give me everyday.

References

[1] U. A. de Barcelona, "DataUAB Blog," https://datauab.github.io/, [Online; accessed June 2019].

[2] Kaggle, "Speed Dating Experiment," https://www.kaggle.com/annavictoria/speed-dating-experiment, 2016, [Online; accessed June 2019].

[3] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson, "Gender differences in mate selection: Evidence from a Speed Dating Experiment," *Quarterly Journal of Economics*, 2006.

[4] Kaggle, "World Development Indicators," https://www.kaggle.com/worldbank/world-development-indicators, 2017, [Online; accessed June 2019].

[5] Kaggle, "World Happiness Report," https://www.kaggle.com/unsdsn/world-happiness, 2017, [Online; accessed June 2019].

[6] W. Bank, "World Bank Indicators," https://data.worldbank.org/indicator, [Online; accessed June 2019].

[7] B. Hamner, "Indicators in WDI Data," https://www.kaggle.com/benhamner/indicators-in-data, [Online; accessed June 2019].

[8] Kaggle, "Dogs vs Cats," https://www.kaggle.com/c/dogs-vs-cats/data, 2014, [Online; accessed June 2019].

[9] P. S. Foundation, "Python 3.6 Documentation," https://docs.python.org/3.6/, [Online; accessed June 2019].

[10] Jupyter, "Project Jupyter Notebook," https://jupyter.org/, [Online; accessed June 2019].

[11] S. A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features off-the-shelf: an Astounding Baseline for Recognition," 2014.

[12] P. Golle, "Machine Learning Attacks Against the Asirra CAPTCHA," 2008.

[13] Kaggle, "Titanic: Machine Learning from Disaster," https://www.kaggle.com/c/titanic, [Online; accessed June 2019].

[14] Github, "DataUAB Github," https://github.com/DataUAB/, [Online; accessed June 2019].

[15] L. Planas Simón, "How to get a second date," https://datauab.github.io/speed_dating/, [Online; accessed June 2019].

[16] L. Planas Simón, "The secret of happiness," https://datauab.github.io/happiness/, [Online; accessed June 2019].

[17] L. Planas Simón, "How to build a powerful image classifier with few images," https://datauab.github.io/catsvsdogs/, [Online; accessed June 2019].

[18] L. Planas Simón, "A beginner's guide to data competitions," https://datauab.github.io/titanic/, [Online; accessed June 2019].