Visual Deep Choice: a visualization tool to understand the predictions from Deep Choice

Description

DeepChoice is a selection tool that uses machine learning systems to learn to select the best applicants in a selection process. It uses data from previous years to create a function that can give a score to indicate how likely is for a new applicant to be approved. Visual DeepChoice is a visualization tool designed to provide an interactive way for DeepChoice users to understand its predictions and performance, making it a more reliable tool to be integrated in their selection process.

We want the user to analyze how well the system can assign scores for good and bad applicants, understand the tradeoffs involved and control the number of applicants that are approved. We also want to show the features used by the system and how well they match with the features that users consider the most important ones. Finally, we want the users to be comfortable using DeepChoice in their selection process.

Andre Mendes da Silva
amd871@nyu.edu
Project Page on Github: https://github.com/NYU-CS6313-SPRING2016/Group-17-Brazilian-Fellowship-Data
Working demo: http://NYU-CS6313-SPRING2016.github.io/Group-17-Brazilian-Fellowship-Data

What is the problem you want to solve and who has this problem?

Background:

Estudar Foundation is an institution created 25 years ago that selects promising students from Brazil to receive scholarship, participate in mentoring and coaching programs and build networking with great leaders. From 2009 to 2015, the number of applicants has grown from 5,000 to 51,000, which is too many people to evaluate.

In order to scale the selection process, automatic tools are needed to aid the selection team. To solve this problem, Estudar is using a machine learning system called DeepChoice. DeepChoice uses data from previous processes to learn to predict the chances of applicant being approved. This information is very helpful to define in which applicants the selection team should focus their resources and perform more deep evaluations, for example using 1 on 1 interviews with the best applicants.

In order to use Deep Choice in the selection process, the HR team at Estudar wants to understand how the model makes the decisions and which are the most important features that it considers. Currently, DeepChoice uses machine learning methods that have clear predictors that help to understand the model but it is limited and not interactive. The predictions from the model would be much more understandable if the Estudar team could interact with a tool that shows the importance of the scores assigned by the program, and also how much each of them affects these scores. This is very important for the HR team to trust the system and completely integrate it in the decision process.

Problem: They HR team at Estudar needs to understand the decision, the most important features and interact with DeepChoice to use it in the selection process.

Who has this problem: The HR team at Estudar Foundation that uses the system and the Deep Choice company that provides the system for Estudar.

What questions do you want to answer with your visualization?

1. **What is the performance of the Deep Choice system?**
   The HR team should have access to a variety of metrics, such as precision, recall, accuracy in order to analyze how good Deep Choice is.

2. **How changes in the threshold for acceptance affects Deep Choice performance?**
   The machine learning method used by Deep Choice is able to predict a score for an applicant based on previous processes. If this score is above a certain threshold, the applicant is considered approved. We want to provide an interactive method to visualize how changing this threshold affect the system performance. Based on that, we want the HR team to select the type of system they prefer based on the tradeoffs shown in Table 1.

3. **What are the most important features on Deep Choice predictions?**
   Understand what are the features used by the system to make the predictions and how much they match with the most important features used by the user.

4. **What kind of system is the best for the HR team?**
   By analyzing changes in the threshold, we want the HR team to select the type of system they prefer based on the tradeoffs shown in Table 1.

5. **What impact does each feature have on an applicant's chance of approval?**
   How much each individual feature affects the overall chance for an applicant. The idea is to identify the features that matter the most and see how changing these features affect the final result for each applicant.

Aggressive	High	High	Moderate	Approve less applicants reducing the time to evaluate	Risk of missing good applicants is higher
Conservative	Low	Moderate	High	Avoid missing good applicants	Approve more candidates, increase time to evaluate
Moderate	Medium	Moderate	Moderate	Balanced	Balanced

Table 1 - Tradeoff between conservative and aggressive systems.

What is your data about?

The selection process at Estudar is composed of 2 major phases, system phase and personal phase, demonstrated in Figure 1. In the system phase, which is composed of 3 steps, applicants need to fill information about themselves. In the first step, they inform personal data such as age, gender, university and major. In the second step, they answer open questions and in the third, they record a video. All this information can be separated in 4 groups: numerical, text, video and audio.

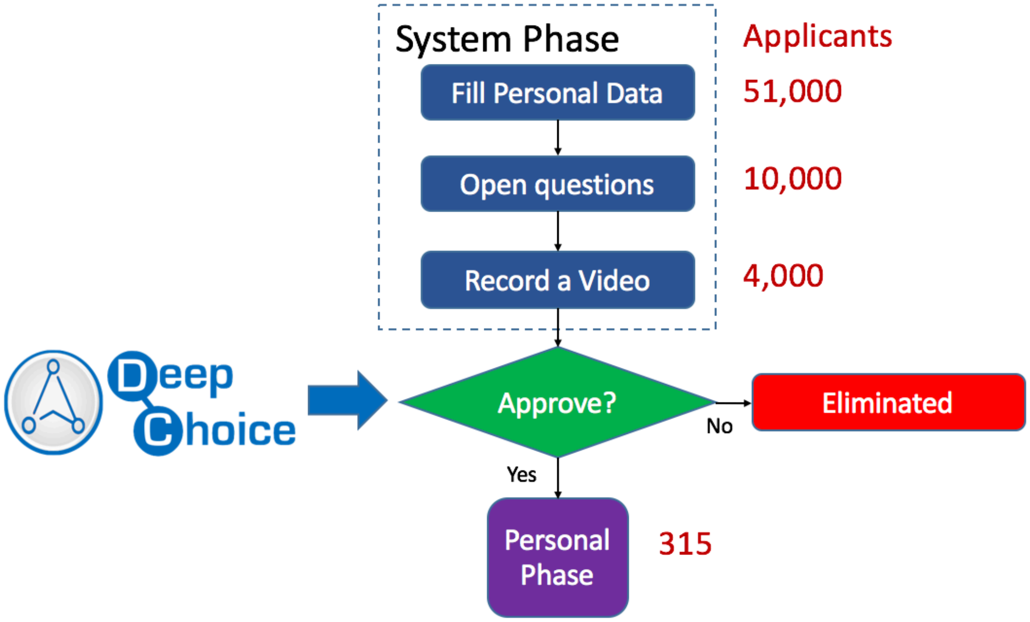


Figure 1 - Overview of selection process.

Numerical – Represent features that can be easily mapped to numbers.

feature	**age**	**gender**	**university**	**state**	**stability**
range	**18 to 35**	**male, female**	**Unlimited**	**28 options**	**0 to 1**
example	**18, 23, 34**	**male, female**	**Harvard, USP, NYU**	**SP, RJ, not from Brazil**	**0.58**
transformation	**no**	**binary**	**map to university**	**map to state**	**Already made**
After transformation	**18 to 35**	**0,1**	**0,1,2**	**0 to 28**	**0 to 1**

Table 2 – Examples of numerical features. There are approximately 55 numerical values that can be used.

Transformations

- The age is already a numerical value that does not need to be converted.
- Gender is a binary value that is converted to 1 if male and 0 if female.
- For universities we created 2 dictionaries. If the university is among the top 100 in the world, then the value for the feature is 1. If among the top 100 in Brazil then 2, if none, the value is 0.
- For the states, there are 27 states in Brazil. So we created a dictionary and mapped then to numbers. For example, SP is 13, so if the state of an applicant is SP, the value of the feature is 13. If applicant lives abroad, the value is 0.
- Stability is one the the psychological characteristics that Estudar used in the process. They already have a model that based on answers from the applicants, give a numerical value for all the characteristic they evaluate.

Text – Represent features that contains text.

feature	situation	task	action	result
question	A situation you faced	What you had to do	What did you do	What was the result
example	Get into college	Do the SAT	Study a lot	Got into College
transformation	Use Natural Processing Language methods to stem, tokenize and transform each answer to a vector of words. Then created a vocabulary with reproved and approved candidates. Use a model to compare the answers to the vocabularies, define similarity and assigned a numerical score.			
after transformation	0 to 1	0 to 1	0 to 1	0 to 1

Table 3 - Examples of the text features. A total of 11 questions can be used.

Video – Represent features that contain images.

Use the image component of the video and train a deep network to learn to predict the best applicants based on their gestures and movements. This part of the project is not yet developed and would not be used in the tool proposed here.

Audio – Represent features that contain signals

feature	Subtitles - words	Signal
definition	Analysis of the words in the video	The analysis of the signal.
example	I want to reduce poverty in the word	high peak, low peak, frequency
transformation	the same one used for the Group text	Not implemented yet
after transformation	0 to 1	Not implemented yet

Table 4 - Examples of the audio features

The final Dataset

The final dataset used for the classification and for the visualization tool is composed of 2527 candidates (one per candidate per row), with 52 features (one feature per column). All these applicants were evaluated in 2015, so we have all the results as label data. This label data indicates if a candidate was approved or not in the first phase.

What have others done to solve this or related problems?

This proposed project is mostly inspired by the work of Krause et al. [2], where it is proposed a visualization tool called Prospector to visually inspect machine learning models. Although this would be the main reference for this work, we intend to explore Estudar dataset to answer other different questions than the ones used in the case study of the Prospector paper. Another important reference is the work proposed by Amershi et al. [1] where they present an interactive visualization that uses information contained in numerous statistics and graphs while displaying example-level performance. A more complete and broad analysis of the importance of make systems more understandable is discussed by Kuleska et al. [3]. When dealing with neural networks, Xu et al. [4] interprets the graph of a neural network used for image classification to retrieve which part of an image was responsible for a specific classification result. This can be very helpful when dealing with the video data.

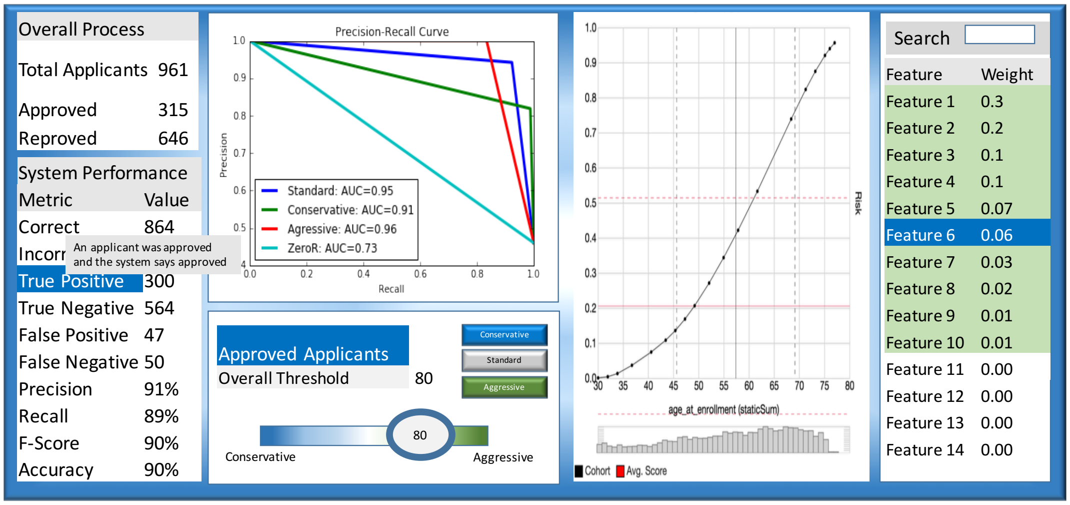
Initial Mockup

View 1: System

The visualization is divided in four columns. Starting from the column in the left side, the user can see the overall data about the process on top of the column. Below this, the user can see metrics to evaluate the performance of the system. If the user clicks on the metric it can se a description of what it represents.

How to read it:

On the bottom of the second column, the user can see the overall threshold and change it using a slider or by direct typing a value in the box. The user can also use the three buttons on the right to select the pre-defined thresholds for conservative or aggressive system, or click in the standard one to go back to the original threshold. On top of the column, a graphic shows the relation about precision and recall. **This part addresses question 2 because it allows the user to interact with the system and see how the threshold influence the metrics.**



Overall Process	
Total Applicants	961
Approved	315
Reproved	646

System Performance

Metric	Value
Correct	864
Incorr	
True Positive	300
True Negative	564
False Positive	47
False Negative	50
Precision	91%
Recall	89%
F-Score	90%
Accuracy	90%

An applicant was approved and the system says approved

Approved Applicants

Overall Threshold 80

Search

Feature	Weight
Feature 1	0.3
Feature 2	0.2
Feature 3	0.1
Feature 4	0.1
Feature 5	0.07
Feature 6	0.06
Feature 7	0.03
Feature 8	0.02
Feature 9	0.01
Feature 10	0.01
Feature 11	0.00
Feature 12	0.00
Feature 13	0.00
Feature 14	0.00

Figure 2 - System view.

On the column on right most side, the user can see all the features and the most important ones are highlighted in green. The user can also search for a feature and select it to visualize its details. A plot line on the top of third column shows the impact of the feature in the predictions and a histogram of the its distribution is shown on the bottom.

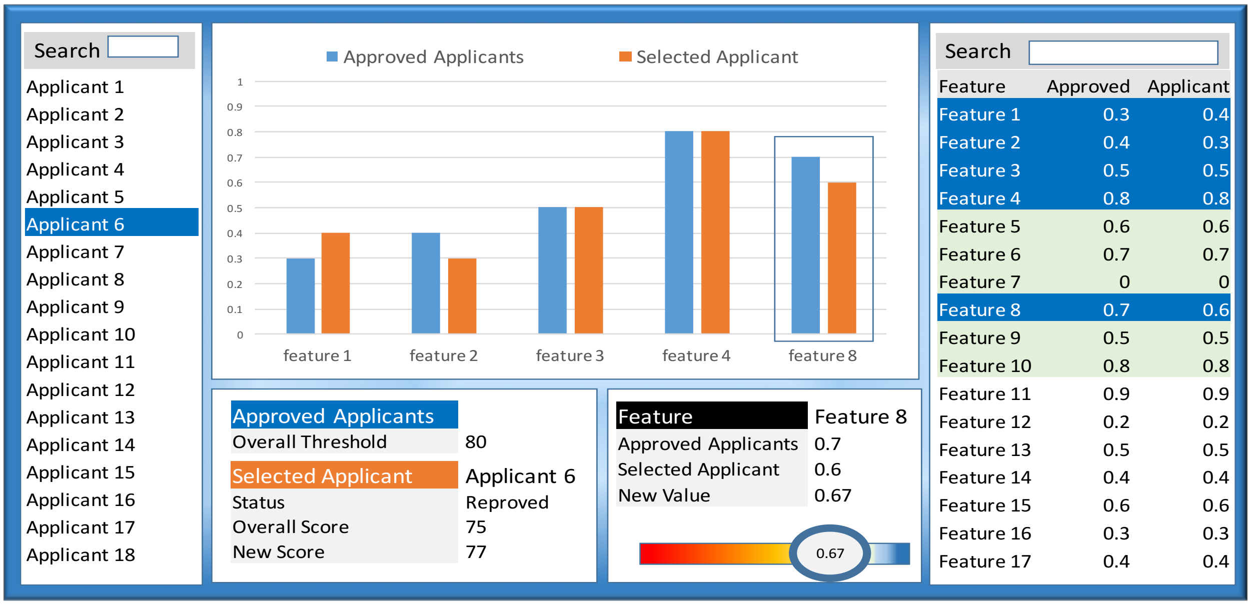
View 2: Applicants

How to read it:

The visualization is divided in 3 columns. Starting from the column in the left side, the user can see all the applicants and search and select one to see more details.

In the column on the right most side, the user can see a list of features organized according to their importance. The most important features are highlighted in green. The user can then search and select 5 features to visually compare the scores from the selected applicant with the average score from approved applicants. This is presented in the graph area on top of the middle column. Finally, the user can also click in one of the bars to see more details about a specific feature.

On the left bottom of the middle column, the user can see the threshold for acceptance and the scores from the selected applicant. On the right bottom, the user sees the details of the selected feature in the graph, and can change its value using the slider or the input box, to see how it changes the overall score of the selected applicant.



Approved Applicants	
Overall Threshold	80
Selected Applicant	Applicant 6
Status	Reproved
Overall Score	75
New Score	77

Feature	Feature 8
Approved Applicants	0.7
Selected Applicant	0.6
New Value	0.67

Feature	Approved	Applicant
Feature 1	0.3	0.4
Feature 2	0.4	0.3
Feature 3	0.5	0.5
Feature 4	0.8	0.8
Feature 5	0.6	0.6
Feature 6	0.7	0.7
Feature 7	0	0
Feature 8	0.7	0.6
Feature 9	0.5	0.5
Feature 10	0.8	0.8
Feature 11	0.9	0.9
Feature 12	0.2	0.2
Feature 13	0.5	0.5
Feature 14	0.4	0.4
Feature 15	0.6	0.6
Feature 16	0.3	0.3
Feature 17	0.4	0.4

Figure 3 - Applicants view.

This second view would be used to answer a fifth question but it was changed during the project update. This was the question:

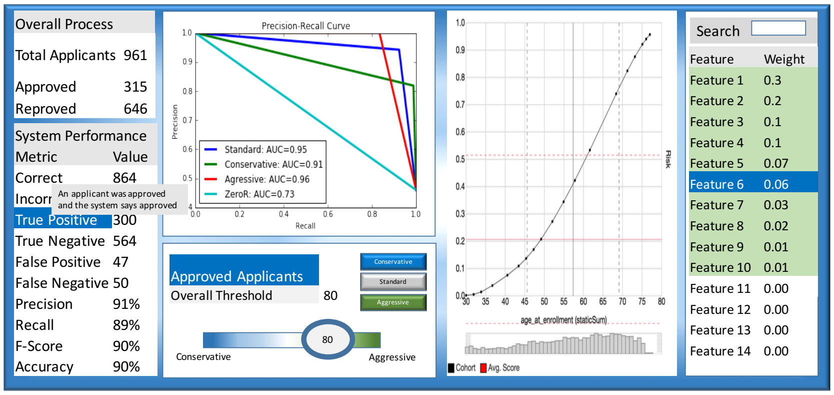
5 **What impact does each feature have on an applicant’s chance of approval?** How much each individual feature affects the overall chance for an applicant. The idea is to identify the features that matter the most and see how changing these features affect the final result for each applicant.

Project Update

The questions did not change but it was not possible to answer the 5th one with the visualization. After start implementing and hearing feedback, it was realized that it was better to do more things to better answer the first four questions.

Changes in the view

Mockup:



Overall Process	
Total Applicants	961
Approved	315
Reproved	646

System Performance

Metric	Value
Correct	864
Incorr	
True Positive	300
True Negative	564
False Positive	47
False Negative	50
Precision	91%
Recall	89%
F-Score	90%
Accuracy	90%

An applicant was approved and the system says approved

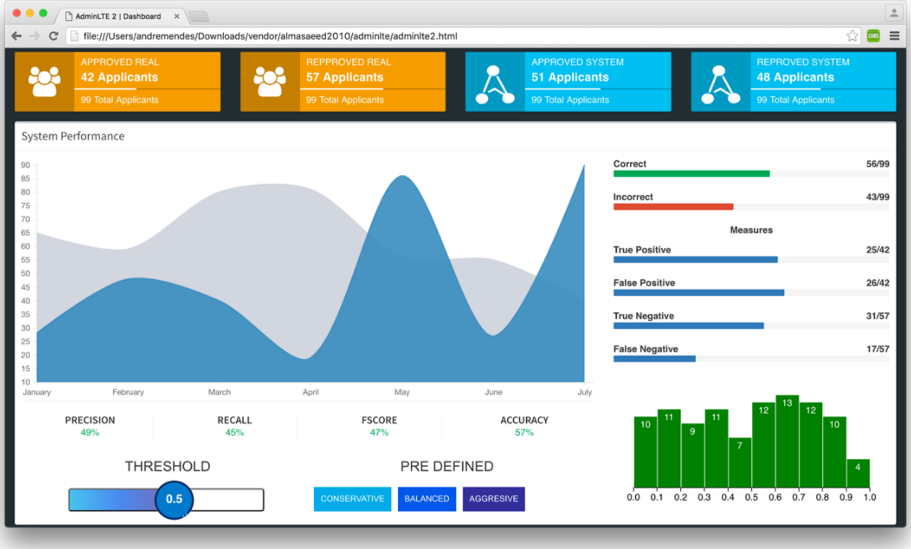
Approved Applicants

Overall Threshold 80

Search	
Feature	Weight
Feature 1	0.3
Feature 2	0.2
Feature 3	0.1
Feature 4	0.1
Feature 5	0.07
Feature 6	0.06
Feature 7	0.03
Feature 8	0.02
Feature 9	0.01
Feature 10	0.01
Feature 11	0.00
Feature 12	0.00
Feature 13	0.00
Feature 14	0.00

Figure 4 - Mockup of the tool in the proposal.

Project update:



AdminLTE 2 | Dashboard

file:///Users/andremendes/Downloads/vendor/almasaeed2010/adminlte/adminlte2.html

APPROVED REAL
42 Applicants
99 Total Applicants

REPPROVED REAL
57 Applicants
99 Total Applicants

APPROVED SYSTEM
51 Applicants
99 Total Applicants

REPROVED SYSTEM
48 Applicants
99 Total Applicants

System Performance

Correct	56/99
Incorrect	43/99

Measures

True Positive	25/42
False Positive	26/42
True Negative	31/57
False Negative	17/57

PRECISION	RECALL	FSCORE	ACCURACY
49%	45%	47%	57%

THRESHOLD 0.5

PRE DEFINED

Figure 5 - Version for the tool in the project update phase.

The visualization was divided in one row on top and two columns. The overall information about the process was showed in the top row. Here it was already possible to compare the results from humans and from the system.

In the left column on top, there is a graph that shows the AUC for different system configurations that are defined by changing the value for the threshold. In the left bottom, the user has two options to change the threshold, using the slider or by clicking in the bottoms with pre-defined values for the threshold. The metrics about the system were moved to the top of the right column, and instead of just numbers, bar plots were added to help the user to understand the information. Finally, in the right bottom, the user can see a histogram that shows the distribution of the applicants according to their score from DeepChoice.

View 2: Other information about applicants – Scroll Down

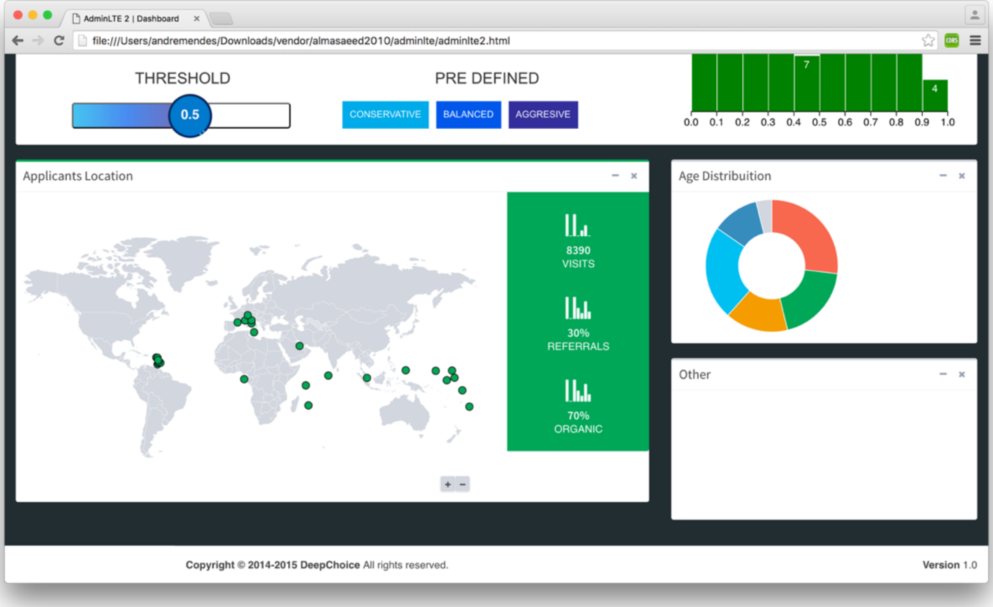


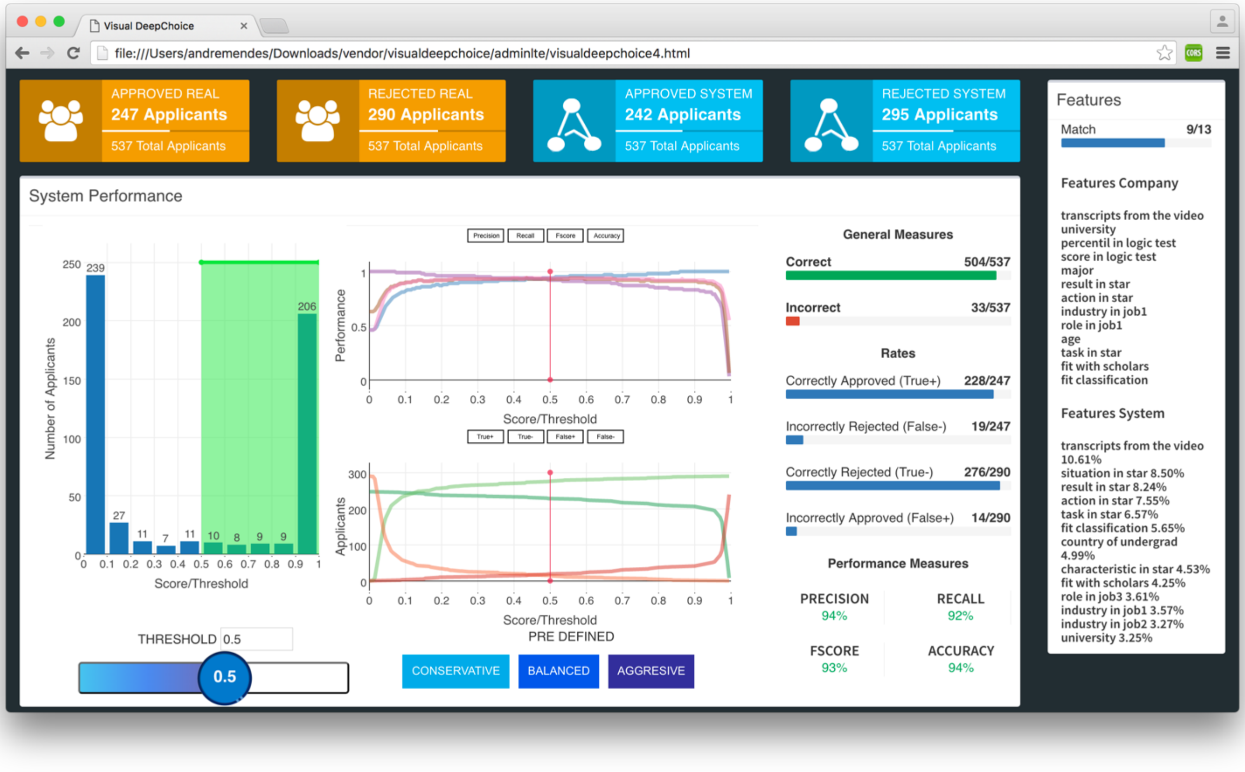
Figure 6 - View when scroll down in the tool in the project update version.

In this phase It was studied the possibility to provide more information about how changes in the threshold affects certain distributions such as the region where the applicants are from their age. Therefore, in the left there is a map that was intended to be more focused on Brazil. In this map, the user can see which regions are represented by the applicants approved in the process after the threshold is set. On the right column up, there is an example of a pie chart that will have the information about age distribution. On the bottom, I can add another graph to show another distribution, such as gender for example.

Visualization for User Study

After collecting important feedbacks from students, project supervisor and most important, from DeepChoice customers, some changes were made to better use the space and present more relevant information for the user. It was decided that, although very interesting, the information about applicant's distribution with regards to age, gender and localization was not a priority for this version. Instead, it was proposed a better way to analyze the metrics about the performance of the system and also a comparison between the most important features for users and system.

Figure 7 shows the version before the user study.



Visual DeepChoice

file:///Users/andremendes/Downloads/vendor/visualdeepchoice/adminlte/visualdeepchoice4.html

APPROVED REAL
247 Applicants
537 Total Applicants

REJECTED REAL
290 Applicants
537 Total Applicants

APPROVED SYSTEM
242 Applicants
537 Total Applicants

REJECTED SYSTEM
295 Applicants
537 Total Applicants

System Performance

THRESHOLD 0.5

General Measures

Correct 504/537

Incorrect 33/537

Rates

Correctly Approved (True+) 228/247

Incorrectly Rejected (False-) 19/247

Correctly Rejected (True-) 276/290

Incorrectly Approved (False+) 14/290

Performance Measures

PRECISION 94%

RECALL 92%

FSCORE 93%

ACCURACY 94%

Features

Match 9/13

Features Company

transcripts from the video
university
percentil in logic test
score in logic test
major
result in star
action in star
industry in job1
role in job1
age
task in star
fit with scholars
fit classification

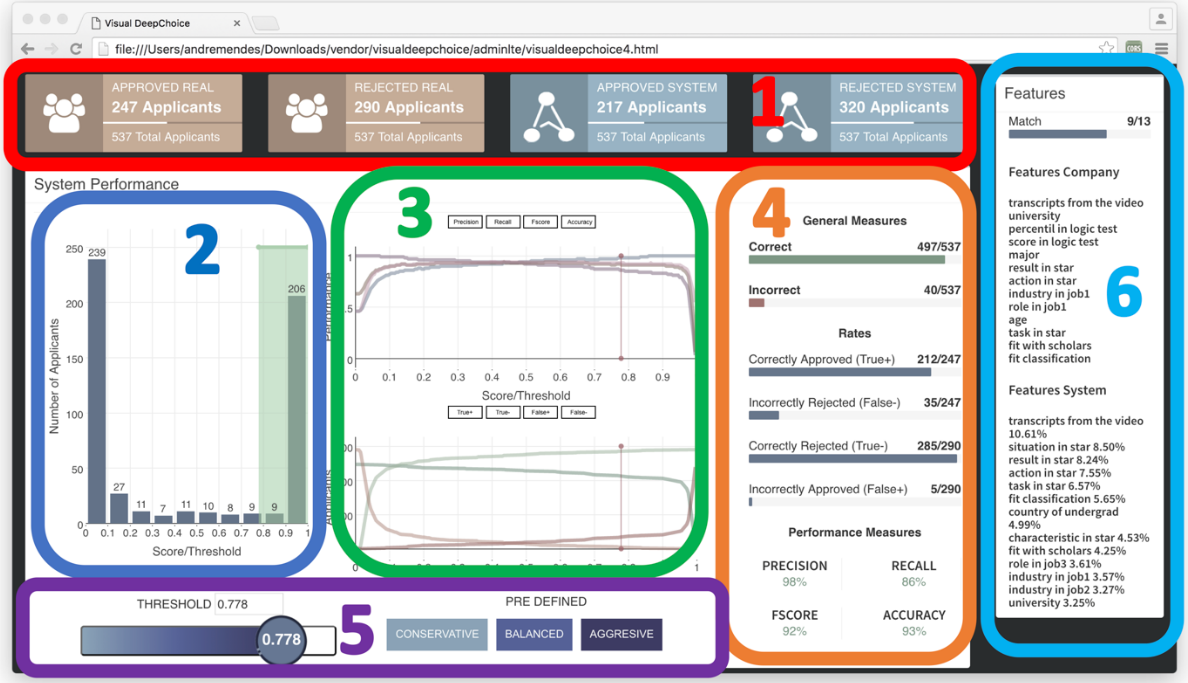
Features System

transcripts from the video 10.61%
situation in star 8.50%
result in star 8.24%
action in star 7.55%
task in star 6.57%
fit classification 5.65%
country of undergrad 4.99%
characteristic in star 4.53%
fit with scholars 4.25%
role in job3 3.61%
industry in job1 3.57%
industry in job2 3.27%
university 3.25%

Figure 7 - Final version before user study.

How to use it:

The general idea is to allow the user to interact with the tool so that he can analyze how the performance for the system changes when the user varies the threshold. This threshold is the boundary between approved and rejected applicants. Figure 8 shows the final version divided in 6 different regions that will be explained individually.



APPROVED REAL
247 Applicants
537 Total Applicants

REJECTED REAL
290 Applicants
537 Total Applicants

APPROVED SYSTEM
217 Applicants
537 Total Applicants

REJECTED SYSTEM
320 Applicants
537 Total Applicants

General Measures

Correct 497/537

Incorrect 40/537

Rates

Correctly Approved (True+) 212/247

Incorrectly Rejected (False-) 35/247

Correctly Rejected (True-) 285/290

Incorrectly Approved (False+) 5/290

Performance Measures

PRECISION 98%

RECALL 86%

FSCORE 92%

ACCURACY 93%

THRESHOLD 0.778

0.778

PRE DEFINED

CONSERVATIVE

BALANCED

AGGRESIVE

Features

Match 9/13

Features Company

transcripts from the video
university
percentil in logic test
score in logic test
major
result in star
action in star
industry in job1
role in job1
age
task in star
fit with scholars
fit classification

Features System

transcripts from the video 10.61%
situation in star 8.50%
result in star 8.24%
action in star 7.55%
task in star 6.57%
fit classification 5.65%
country of undergrad 4.99%
characteristic in star 4.53%
fit with scholars 4.25%
role in job3 3.61%
industry in job1 3.57%
industry in job2 3.27%
university 3.25%

Figure 8 - Final version divided in 6 different regions.

Region 1 – General Information bar

The region 1 highlighted in red on top of the screen is the general information bar. From the left, in the first two blocks in this bar, the user can see the original information from the process. He can see the number of applicants who were approved or rejected and compare it with he total number of applicants. In the last two blocks in this bar, the user can see the same information, but from the system.

Interactions:

In the blocks with system information, the number of approved and rejected applicants changes as the user changes the threshold.

Region 2 – Histogram area

The region 2 highlighted in blue is the histogram area. In this area, the user can see groups with similar scores assigned by DeepChoice. It is possible to see that it was able to separate applicants in two different groups, low score (<0.1) and high score (>0.9), that are represented by the high bars in the extremes of the graph.

Interactions:

As the user hover over the bars in the graph, he can see the number of applicants in each group. Besides that, a green region in the graph indicates which groups of applicants are going to be approved with the current threshold and this region changes as the threshold changes. For instance, in Figure 7, the threshold is set at 0.5. Therefore, all applicants with score above 0.5 are going to be approved. In the histogram data it represents the high bar in the extreme left and all the smaller bars on the left of the threshold that are inside the green area. Finally, the user can also zoom in and out areas in the graph by using the buttons in the right top of the plot.

Region 3 – General performance area

The region 3 highlighted in green represents the general performance area. In this area it is possible to see how the metrics for DeepChoice change in the whole range of possible values for the threshold. There are two plots. The first one shows precision, recall, accuracy and fscore, and the second one shows the number of true positives, true negatives, false positives and false negatives.

Interactions:

When the cursor hovers over the plot, the system shows the values for each active curve in the point where the cursor is (Figure 9), so that the user can compare the values for different thresholds. It is possible to active or deactivate a curve by clicking in its respective button in top of the plot. For example, if the user only wants to see the accuracy of the system, it can deactivate all the others and let only accuracy activated. In region 3 the user can also zoom in and out using the functions on the top left of each plot.

A vertical line in the graph shows the current value for the threshold. To change this value direct in the plot, the user can click in a point in the plot and the line will move, indicating the new threshold.

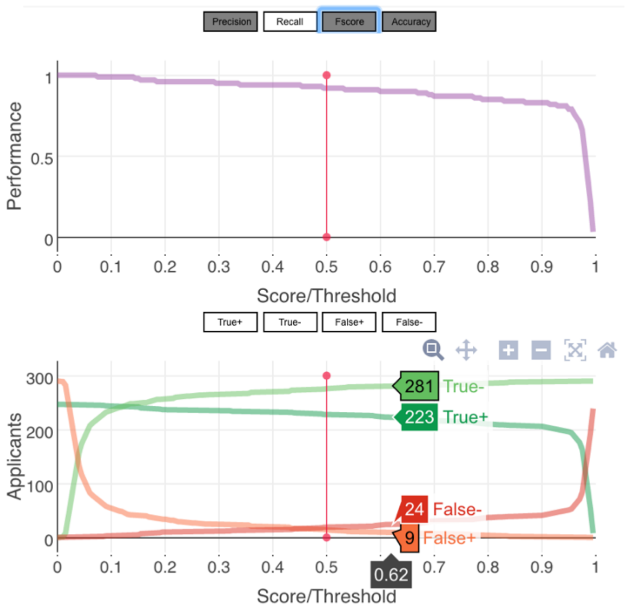


Figure 9 - Visualization of region 3 showing only one activated line and hover functions.

Region 4 – Numerical view for the measures

The region 4 highlighted in orange represents the numerical view for the measures area. In this part of the tool, the user can directly and quickly see how the system responds to changes in the threshold.

On the top, it can see the two most general measures, correct and incorrect. It indicates the number of applicants correctly or incorrectly classified by the system when using the current threshold. Below the general metrics, are the more specific metrics that represent:

Correctly Approved (True+): Applicants approved by the system and by the humans.
Correctly Rejected (True-): Applicants rejected by the system and by the humans.
Incorrectly Approved (False+): Applicants approved by the system but rejected by the humans.
Incorrectly Rejected (False-): Applicants rejected by the system but approved by the humans.

Finally, below the specific metrics, there are the general performance metrics:

Precision (P): TP/(TP+FP) - The fraction of retrieved applicants that are approved.
Recall (R): TP/(TP+FN) - The fraction of approved applicants that are retrieved.
Fscore: 2*P*R / (P+R) – The harmonic mean between precision and recall.
Accuracy: (TP+TN) / Total – The fraction of applicants correctly classified.

Interactions:

As the user changes the value for the current threshold, he can see how that affects all the measures in the system. Additionally, when the user hovers the cursor over the name of the metrics, he can see a more detailed description of what they mean.

Region 5 – Controls for the threshold

The region 5 highlighted in purple represents the controls area. In this area, the user can use the slider, the text input or the buttons to define the value for the threshold. The buttons represent pre-defined values, where each one contains a value for the threshold that characterizes a different type of system. These types are explained in Table 1.

Interactions:

In the input box, the user can directly set the value for the threshold. The possible values are always between 0 and 1, so if the user tries to set a value smaller or higher, they system will be set to the respective limit. When using the slider, the user can click and drag the round button to change the value for the threshold between 0 and 1. As the user does it, he can see in all the other regions (except in region 6), the impact of this new value in the system. When using the buttons, the user can click in each one of them and this will set the current threshold to the pre-defined values.

Region 6 – Features

The region 6 highlighted in light blue represents the features area. Here, the user can visualize the most important features used by the system and by the human as well as how much they match. With the dataset used, out of 13 of the most important features, 9 were both used by the system and by the humans. In the bottom of the column, in the features used by the system, the user can also see how much importance each feature has in the prediction. For this part of the visualization tool, there is no interaction in this version.

User Study

In order to validate the Visual DeepChoice tool, a user study was developed and applied. The goal is to show if the users can use the tool to answer the objective questions and if they can do that fast and easily.

The method

A google form with two sections was created. Section 1 contains 10 objective questions where the user has to use Visual DeepChoice to find the answers. Section 2 contains 8 subjective questions where the user evaluates the experience of using Visual DeepChoice. The questions were:

Objective Questions

1. What is the original threshold/Score of the system?
2. What is the original values for precision and recall, respectively, of the Deep Choice system when no changes are made in the threshold?
3. What are the values for fscore and accuracy, respectively when the threshold is 0.783?
4. What are the 3 most important features used by DeepChoice system to make the selections?
5. What are the 3 most important features used by the company to make the selections?
6. How many applicants are approved in a conservative system?
7. How many applicants are rejected in an aggressive system?
8. For which range of values (for example, 0.1 to 0.5) for threshold/score, the system presents the higher accuracy?
9. If the threshold/score goes up from 0 to 1, what happens with the number of Incorrectly rejected?
10. If the threshold goes down from 1 to 0, what happens with the number of correctly rejected?

Subjective Questions

1. How easy or difficult was it to find information to answer the questions?
2. Which threshold do you think is the best for the system? Why?
3. How intuitive you think the visualization is (For a very intuitive tool, you don't need manual)?
4. How much was the visualization helpful to understand how Deep Choice select applicants?
5. What is missing in the tool?

6. What is your overall grade for the visualization of the system?
7. Based on the performance of the system and the visualization, would you use DeepChoice in your selection process?
8. Reason for answer and question 7?

The assumptions

Visual DeepChoice:

- Provide the answers for analytical questions
- Is easy to use
- Is intuitive
- Is fast to use
- Is helpful to understand how DeepChoice works
- Helps people to want to use DeepChoice

The results

First we analyze the objective questions. With regards to accuracy, when users had to answer questions about a specific number such as in questions 1 and 2, all the answers were correct. However, when they were required to find a determined range, such as in questions 8 and 9, only 30% of the users found the right answer. This suggests that the general plot area was not correctly used. Changes were made to the final version to address this issue, giving more visual emphasis to the general plot area. With regards to time to answer the questions, the results show that it takes 10 seconds in average to answer about a specific number and around 40 seconds to answer about ranges.

In the subjective questions, 4 metrics were considered, intuitiveness, easiness, helpfulness and and if the user would use the DeepChoice. Figure 10 shows the summary for the answers.

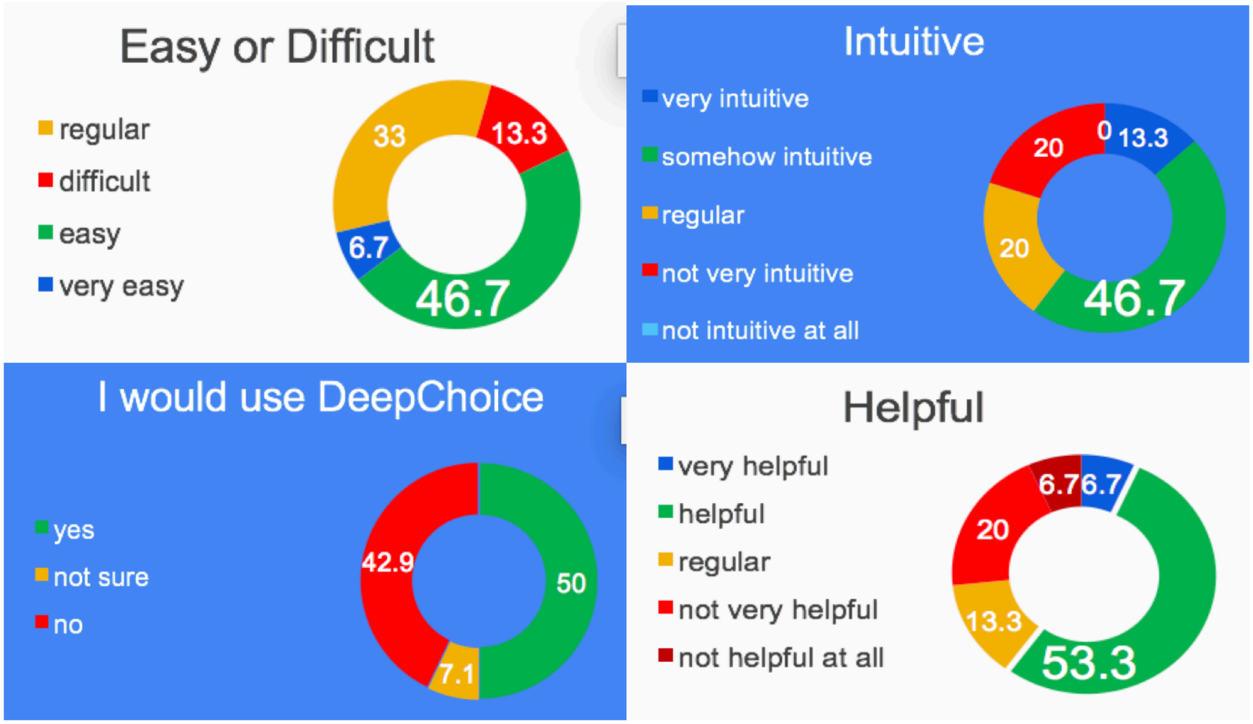


Figure 10 - Results for the User Study.

Considering these results, the following can be state about the assumptions:

Visual DeepChoice:

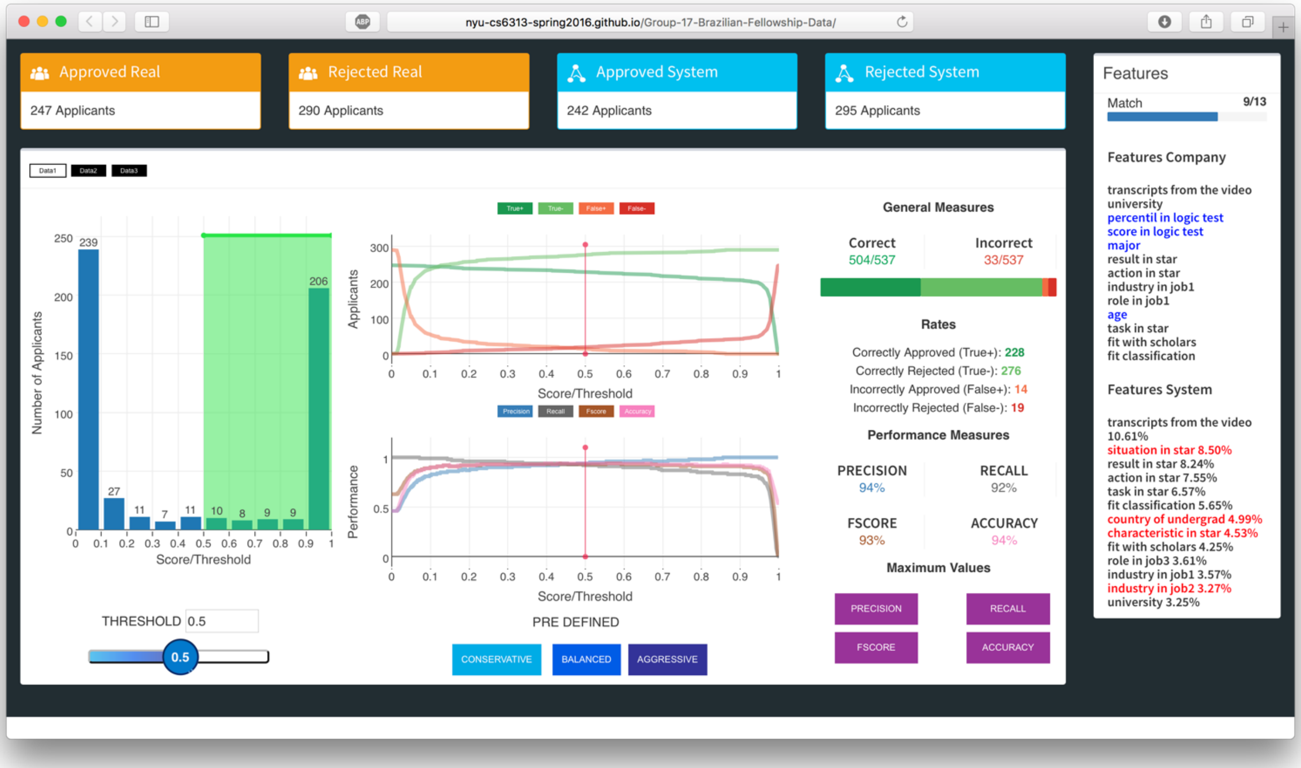
- Provide the answers for analytical questions – Yes.
- Is easy to use – Relatively yes, but it needs to be easier.
- Is intuitive - Yes.
- Is fast to use – Yes.
- Is helpful to understand how DeepChoice works – Relatively yes. This can be improved with a better explanation about the features.
- Helps people to want to use DeepChoice – The results are not strong enough to say. It is expected that it would be better with a better explanation about the features.

Final Visualization after User Study

Based on the feedback obtained from the user study and from the supervisor in the project, changes were made in the tool. The first one was in the Region 1. Basically, the information in this area is now shown using less color and smaller elements. In Region 3, the buttons to activate the curves were mapped to the respective curves using colors.

The biggest change is in region 4, where 1 new bar chart was designed to show the information that was previous shown using 6. In order to achieve that, it was first used a color hue channel that indicates the type of wrong (false positive and false negative) and right (true positive, true negative) prediction. Another channel used was area in the bar to show the proportion of right and wrong results. Also in region 4, for buttons were added to represent the maximum values for the measures (precision, recall,

In Region 5 it was added an input button so that the user can directly type a value for the threshold and in Region 6, a color hue channel was added to better identify the features that are used by both humans and system. If the color of the feature is blue, it means it was only used by humans. If the color is red, it was only used by DeepChoice. The other features in black were used by both. All these changes can be seen in Figure 11.



Approved Real
247 Applicants

Rejected Real
290 Applicants

Approved System
242 Applicants

Rejected System
295 Applicants

THRESHOLD 0.5

General Measures

Correct 504/537
Incorrect 33/537

Rates

Correctly Approved (True+): 228
Correctly Rejected (True-): 276
Incorrectly Approved (False+): 14
Incorrectly Rejected (False-): 19

Performance Measures

PRECISION 94%
RECALL 92%
FSCORE 93%
ACCURACY 94%

Maximum Values

Features

Match 9/13

Features Company

transcripts from the video
university
percentil in logic test
score in logic test
major
result in star
action in star
industry in job1
role in job1
age
task in star
fit with scholars
fit classification

Features System

transcripts from the video 10.61%
situation in star 8.50%
result in star 8.24%
action in star 7.55%
task in star 6.57%
fit classification 5.65%
country of undergrad 4.99%
characteristic in star 4.53%
fit with scholars 4.25%
role in job3 3.61%
industry in job1 3.57%
industry in job2 3.27%
university 3.25%

Figure 11 - Final Visualization after user study.

Data Analysis

One of the most important things that can directly seen in the visualization is that DeepChoice separates the data really well. As it can be observed in the histogram in Region 2, roughly 45% of the applicants received a score below 0.1 whereas other 40% received a score above 0.9. This is an important characteristic because it makes the selection more reliable since it makes easier and even more intuitive to define a good threshold. If the heights of the bars were more uniformly distribute, it could harder to define a good threshold because there is no clear separation in the data. Examples of this case can be seen in Figure 12 that shows Region 2 for other datasets.

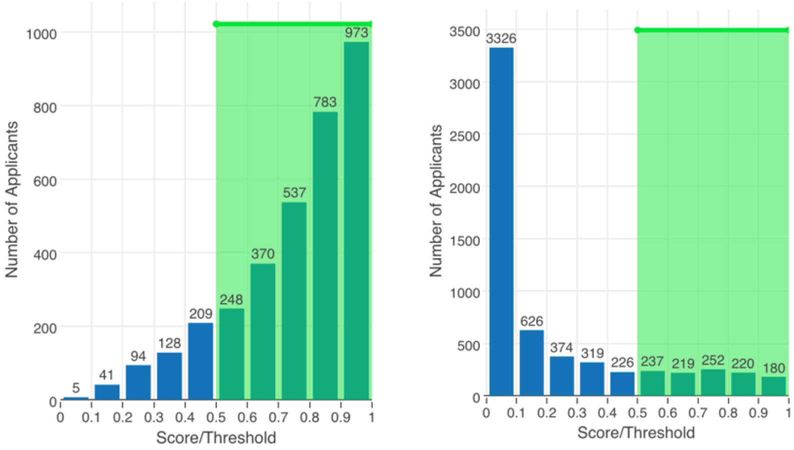
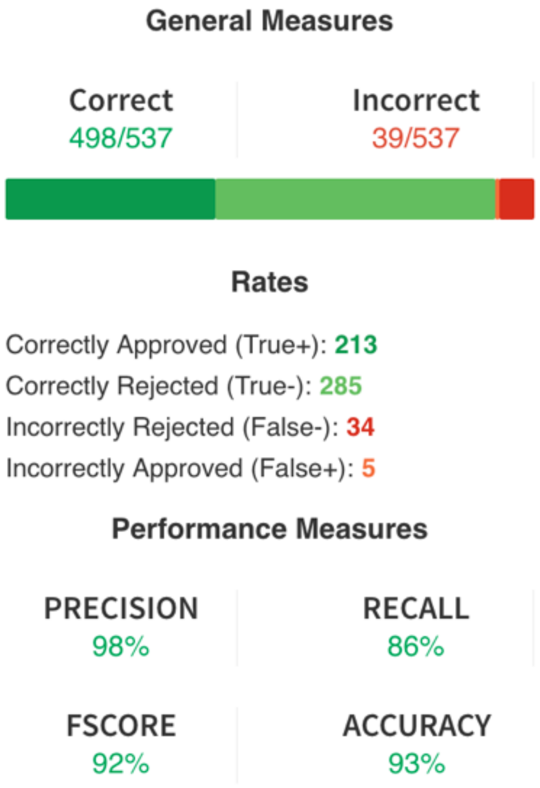


Figure 12 - Histograms for different datasets.

This well defined separation also suggests that the features most used by DeepChoice provide insightful information about the applicants. It suggests that there are groups of features that strongly indicates good applicants and bad applicants. In the datasets in Figure 12 it does not happen. For example, in the dataset on the right side, there is probably a group of features that indicates bad applicants, since it can classify a high number of applicants with a score lower than 0.1 (high bar in the left). However, it appears that there is not a good group of features that indicates good applicants, since there is not a high bar in the right side of the graph.

This open important questions about how much information is gained with each feature or with a specific group of features. Also, would these features be general enough for other datasets or they were specifically good for the predictions in this dataset (Even though it was performed cross-validation). In a future step, it would be very interesting to analyze the features and identify which of them individually or as group provide more information and how changes in the most important features affect how well the system is able to separate the applicants.

Even though the separation is important and it shows interesting information, this is not enough to say that DeepChoice performs well. It could be the case that many of the applicants that receive a lower score were actually approved and vice-versa, for this region is important to see how much of the applicants the system actually got right. For each value of the threshold, this information can be directly checked in Region 4. This region contains the values for the metrics in the system and proportional bar that shows the proportion of correct and incorrect classifications as shown in Figure 13.



General Measures

Correct	Incorrect
498/537	39/537

Rates

Correctly Approved (True+): **213**
Correctly Rejected (True-): **285**
Incorrectly Rejected (False-): **34**
Incorrectly Approved (False+): **5**

Performance Measures

PRECISION	RECALL
98%	86%
FSCORE	**ACCURACY**
92%	93%

Figure 13 - Region 4 with emphasis in the proportional bar indicating correct and incorrect classifications.

By changing the threshold, it can be seen that the system has a consistent performance in a large range of values, as the metrics seem to stay above 89% and the green area is much higher than the red area in the proportional bar. However, by using Region 3, the general plot area, it is much easier to confirm the consistency in the results. In plot 1 it is clear that if the threshold is set between 0.2 and 0.8, the values for the metrics are closer to 1 (around 0.9). This reinforces the idea that DeepChoice is very good in splitting the candidates but at this time, it also shows that it does that in the correct way, which is giving higher scores to approved applicants and lower scores to rejected ones. This can be better visualize in Figure 14 and Figure 15 that shows Region 2 and Region 3 for the original and a different dataset, respectively.

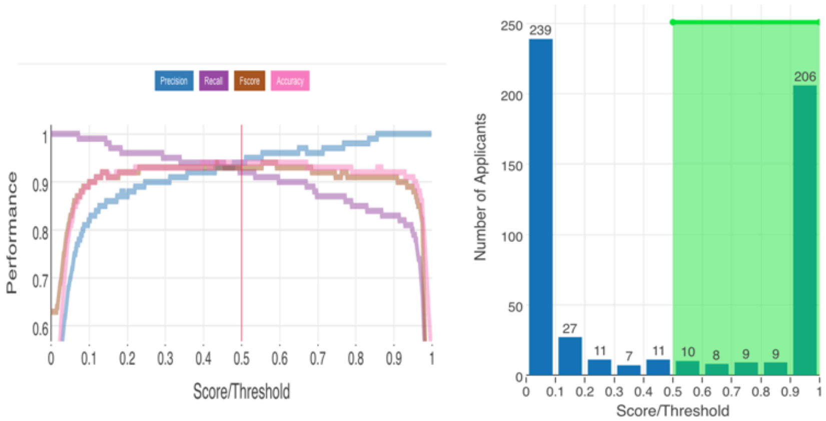


Figure 14 - Left: Plot for performance measures using original dataset. Right: Histogram of the distribution of applicants scores using original dataset.

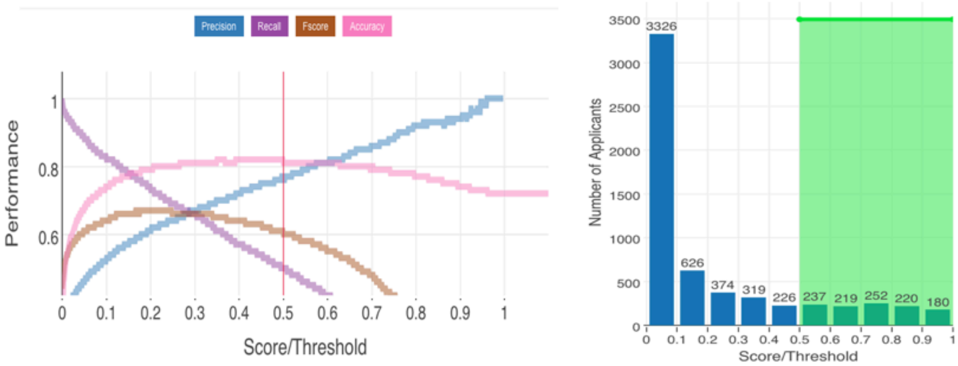


Figure 15 - Left: Plot for performance measures using different dataset. Right: Histogram of the distribution of applicants scores using different dataset.

Finally, Visual DeepChoice also presents the features used to make the predictions. This provide initial information about how DeepChoice makes the selections and it is important to confirm that DeepChoice is working correctly. For our dataset, the features from DeepChoice match with the features selected by the company in 70%. This is the final piece to show that DeepChoice works well because it confirms that the system is able to separate the data, this separation is correct, very accurate and precise, and it uses the right information (features) to achieve this separation.

The meaning of each feature was not encoded in the visualization tool but all of them are very familiar for the company responsible for the selection process.

Although this initial step towards understanding the model was made, this is perhaps the most interesting part of the problem and there is much to advance. By analyzing the features that DeepChoice uses with the its performance, different insights can be obtained from the system and also from the original selection process.

To illustrate this situation, consider an example. DeepChoice is a data-driven approach to select applicants. The system is fed with information but it does not know what information is the most important. In order to select well, the system has to learn what is an important feature and what is a bad one when selecting an applicant. The method used will try to find the features that provide more information so that it can separate as good as possible the approved from the rejected ones. So consider this case:

- DeepChoice is able to achieve a very high performance;
- The most important features for the company are logic, university.
- The most important features for the system is age and gender.

The company says that race, gender and age are just collected information but they should be not considered in the process. However, in this case, is the DeepChoice system completely wrong or there is a bias towards race and age? This kind of question is very important and the analysis of the performance of the system and the features be used to answer them.

For this and many other reasons, the next step in this work is to improve the analysis of the features to understand their impact in the selection and what kind of changes we have in the system when we change the features it has access to.

Conclusion

The development of the Visual DeepChoice was substantially important to better understand the results from the DeepChoice because it provides important information in a quick, easy and interactive way.

During the development, many concepts learned in the classroom could be applied, especially with regards to which channels to use to represent different types of information. One of the main difficulties during the implementation was the lack of experience with JavaScript and the D3 library. Many things that were proposed in the beginning could not be implemented because it was realized during the development that they would take more time and practice than expected. However, the results obtained are satisfactory considering the time frame given to make the project.

About the goals that were set in the beginning and the results obtained, the Visual DeepChoice is an important step towards understanding DeepChoice and the user study helped to prove it. We consider that the visualization tool useful to answer the main questions and it does it in an easy way.

As said in previous sections, the next steps involve making a visualization for the features in order to better understand how the impact in the performance of the system. We believe this will finally give rich information to completely understand how DeepChoice works and make it more reliable.

Reference

[1] Amershi, Saleema, et al. "Modeltracker: Redesigning performance analysis tools for machine learning." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[2] Krause, Josua, Adam Perer, and Kenney Ng. "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models."

[3] Kulesza, Todd, et al. "Principles of explanatory debugging to personalize interactive machine learning." *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 2015.

[4] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *arXiv preprint arXiv:1502.03044* (2015).