Analysis of voice characteristics from entrepreneurs' pitches

Student's name: Panayiotis Christodoulou

Student's contact info(email): pchris19@ucy.ac.cy

Supervisor's name: Dimosthenis Stefanidis

Supervisor's contact info(email): dstefa02@cs.ucy.ac.cy

INTRODUCTION

This project's purpose was to predict if a team ,from the shark tank us tv show, will or will not make a deal based on static images and voice characteristics from entrepreneurs' pitches.

This project is divided into 4 parts. The first part is the creation of a dataset from Shark Tank(US) episodes, the second is the extraction of sound characteristics of entrepreneurs' tone of voice. The third part is the data analysis, and the last one is the machine learning.

Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

Creation of a dataset from Shark Tank(US) episodes

- I used the Database sheet of this excel file https://docs.google.com/spreadsheets/d/1tMeaiqG0Boy-gkNueT_bpVJAj2zvxoslP2m4gLHGBew/edit?usp=sharing as my initial database.

Season	No. in series	Episode	Company	Description	Deal	Industry	Entrepreneur Gender	Number of Entrepreneurs	State	Health Of Business(Sales:Y/N)	AGE	AGE	AGE	AGE	AGE	AGE	BEAUTY(M/F)	BEAUTY(M/F)	BEAUTY(M/F)	BEAUTY(M/F)	BEAUTY(M/F)	BEAUTY	ASK			DEAL			Royalty Deal?	Loan?	Barbara	Mark	Lori	Robert Herjavec	Daymond	Kevin	Guest	# Sharks	$ per shark	Details / Notes
											Entrepreneur_A	Entrepreneurs_B	Entrepreneurs_C	Entrepreneurs_D	Entrepreneurs_E	AVERAGE	Entrepreneur_A	Entrepreneurs_B	Entrepreneurs_C	Entrepreneurs_D	Entrepreneurs_E	AVERAGE	Amount	Equity	Valuation	Amount	Equity	Valuation			Corcoran	Cuban	Greiner		John	O'Leary				
2	1	2.01	Wurkin Stiffs		Yes	Fashion / Beauty	Male	1		Y	47					**47**	49.8					49.8				$100,000	40.00%	$250,000			1				1			2	$50,000	
2	1	2.01	Tippi Toes		Yes	Fitness / Sports / Outdoors	Female	2		Y	20	22				**21**	58.9	71.6				65.25				$100,000	30.00%	$333,333			1	1						2	$50,000	

- Variables Explanation:
  - Deal : is a Boolean that indicates if a team made (yes) or didn't make a deal(No).
  - Entrepreneur Gender : Indicate the gender of the team, it can be male, female, or mixed if a team contains men and women.
  - Health of Business : Boolean variable that shows if a team made(Y) or didn't make any sales(N).
  - There are 6 different ages and beauty variables one for each entrepreneur (the teams of seasons 2 and 3 were up to 5 member) and one for the average. Age is an integer with the age of the entrepreneur and beauty is a double variable (percentage) that indicates the beauty of a team member.
  - The other data describe the deal a team made but because my project was only to predict if a team made or didn't make a deal, I didn't use them.

Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

- I downloaded all the available episodes of the tv series (from the website https://yesmovies.mom/search/?keyword=shark-tank). Unfortunately, in season 7 only 9 (from 29) episodes were available.
  Also, some episodes were downloaded in low quality(640x360)

LOW QUALITY EPISODES (640x360)	
Season Number	Episodes Number
1	9
2	2
3	10,11
4	2,7,14,22
5	6,7,26
6	All episodes
7	All episodes
8	7,12,13,15,16,21
9	14 and every episode after episode 16
10	4,7,11,22
11	21,22,23
12	None

- I don't believe that the low quality affects the results as when I listened to those episodes seemed fine and the voice analysis program didn't show any errors.
- I cut each episode per team of entrepreneurs ,using the app wondershare Flimora(https://filmora.wondershare.com/), following these steps:
  - Import media to import the episode you want to cut.
  - Double click on the episode.
  - Add the bracket"{" when the team starts talking.
  - Add the bracket "}" when he ends his performance.
  - Drag it to the eye part(down left) of the app.
  - Check if the data from the excel matches up(name of the company, episode number, gender).
  - Export the episode in the form (3.02_Salespreneur).
  - NOTE: to make the export process faster go to file-preferences-performance-gpu acceleration
- I watched the episodes in fast forward(x 1.5) to :
  - Extract useful information (write it in the excel file) like number of entrepreneurs per pitch, health of business (if they did or did not make sales).
  - Take a screenshot of the Entrepreneurs, and name it like the episode.

  - Cut every team screenshot so there is only one person in every picture(ends with : _A, _B) .

- I used Face++ (https://www.faceplusplus.com/attributes/) to extract information from the static images of entrepreneurs (like age and Beauty).I used the program Faceplpl_API.py . I had to download and use mongoDB and robo3t.

Extraction of sound characteristics of entrepreneurs' tone of voice

- I cut the videos so there is only entrepreneurs' voice. I used Wondershare Flimora, like in the previous part, and Wisecut(https://app.wisecut.video/login). The names of the episodes are like 201-CBS-Foods
- I used my-voice-analysis (https://pypi.org/project/my-voice-analysis/#description) to extract information about the voice characteristics of the entrepreneurs following these steps.

  - To use this tool, I had to transform the videos into wav type, with tools like : https://cloudconvert.com
  - And then I run the program info.py which uses the functions mysptotal and mysppron of my-voice-analysis.
  - Then I added all the data into the excel file.

Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

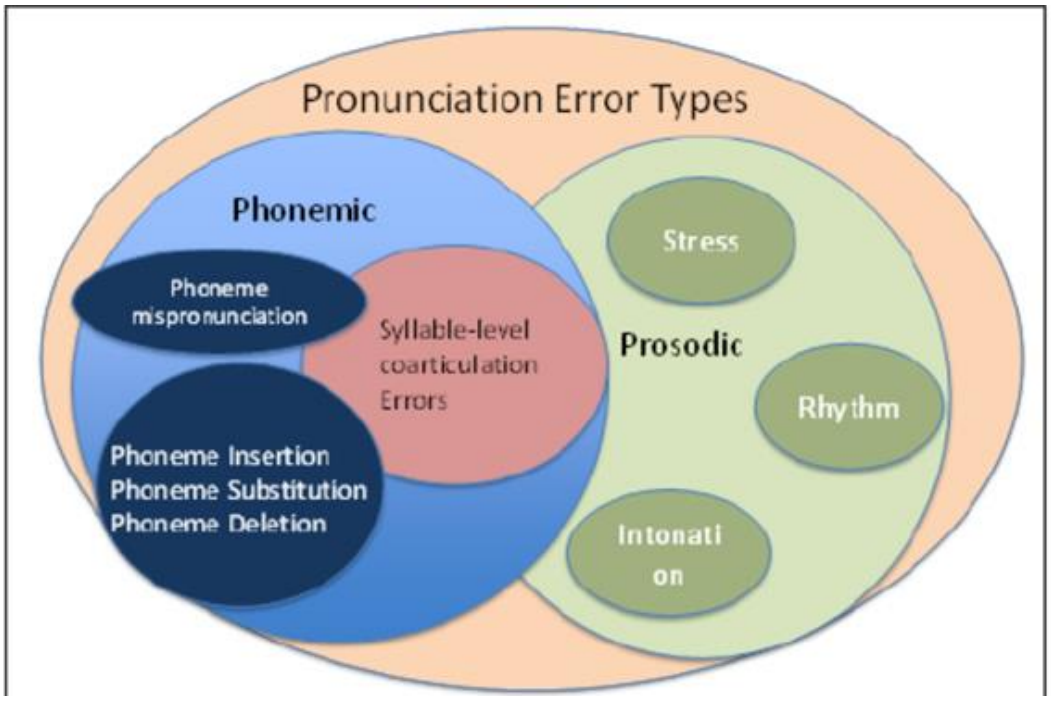
- It provided me the following data:

Data	Details
Pronunciation_posteriori_probability_score_percentage	Pronunciation posteriori probability score percentage
number_ of_syllables	Detect and count number of syllables
number_of_pauses	Detect and count number of fillers and pauses
rate_of_speech	Measure the rate of speech (speed)
articulation_rate	Measure the articulation (speed)
speaking_duration	Measure speaking time (excl. fillers and pause)
original_duration	Measure total speaking duration (inc. fillers and pauses)
balance	Measure ratio between speaking duration and total speaking duration
f0_mean	Measure fundamental frequency distribution mean
f0_std	Measure fundamental frequency distribution SD
f0_median	Measure fundamental frequency distribution median
f0_min	Measure fundamental frequency distribution minimum
f0_max	Measure fundamental frequency distribution maximum
f0_quantile25	Measure 25th quantile fundamental frequency distribution
f0_quan75	Measure 75th quantile fundamental frequency distribution

Analysis of voice characteristics from entrepreneurs' pitches

Data Analysis

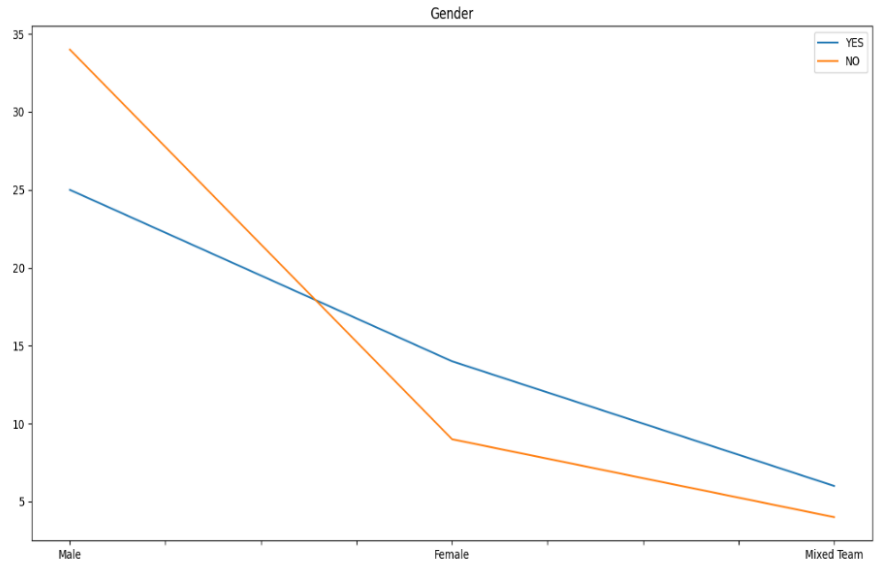
- The first part of my data analysis was to understand my variables. Most of the variables were easy to understand except the following. Speech ,rate of speech, which describes how fast you talk in words per minute. Articulation is the clarity of sounds and words we produce and its capable to hurt a speaker's credibility. Balance is the Speaking time divided with the total speaking time. Also, with f0 statistics we can realize if someone is talking too loudly, and they can help us to understand more about the speaker's credibility. Probability is the Pronunciation posteriori probability score percentage which is a speech scoring feature based on many pronunciations error types like stress, rhythm etc. More information in the schematic below.



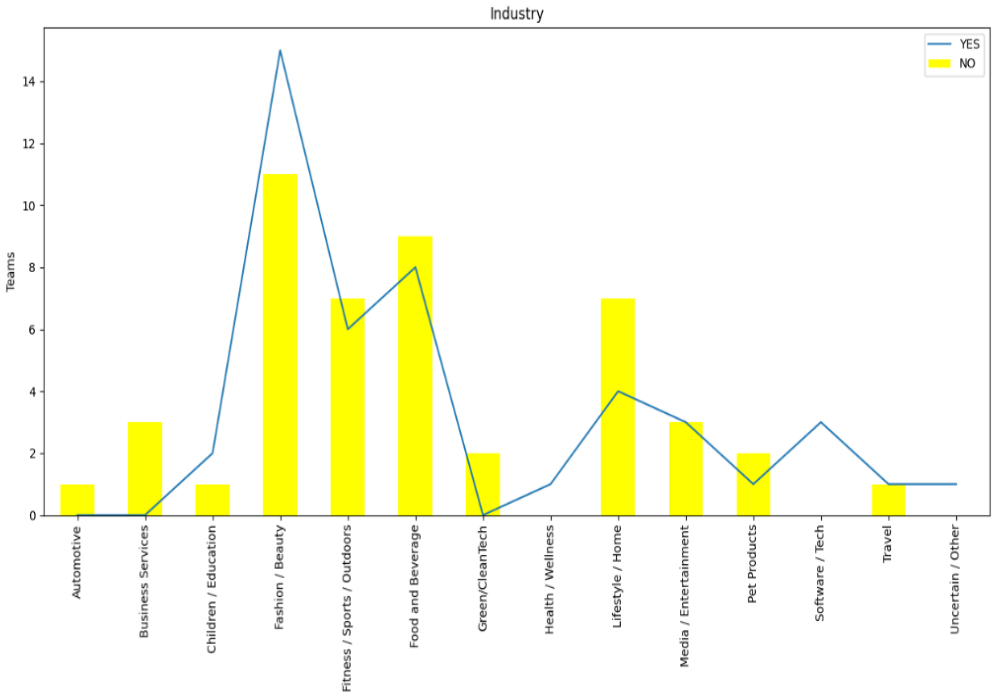
Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

- I created the program "DataAnalysis_MachineLearning.py" which I used to extract plenty of data for this section.
    - I created the following univariate plots and correlation matrices :



From this plot we can see exactly how many male/female/mixed teams got or did not get a deal.

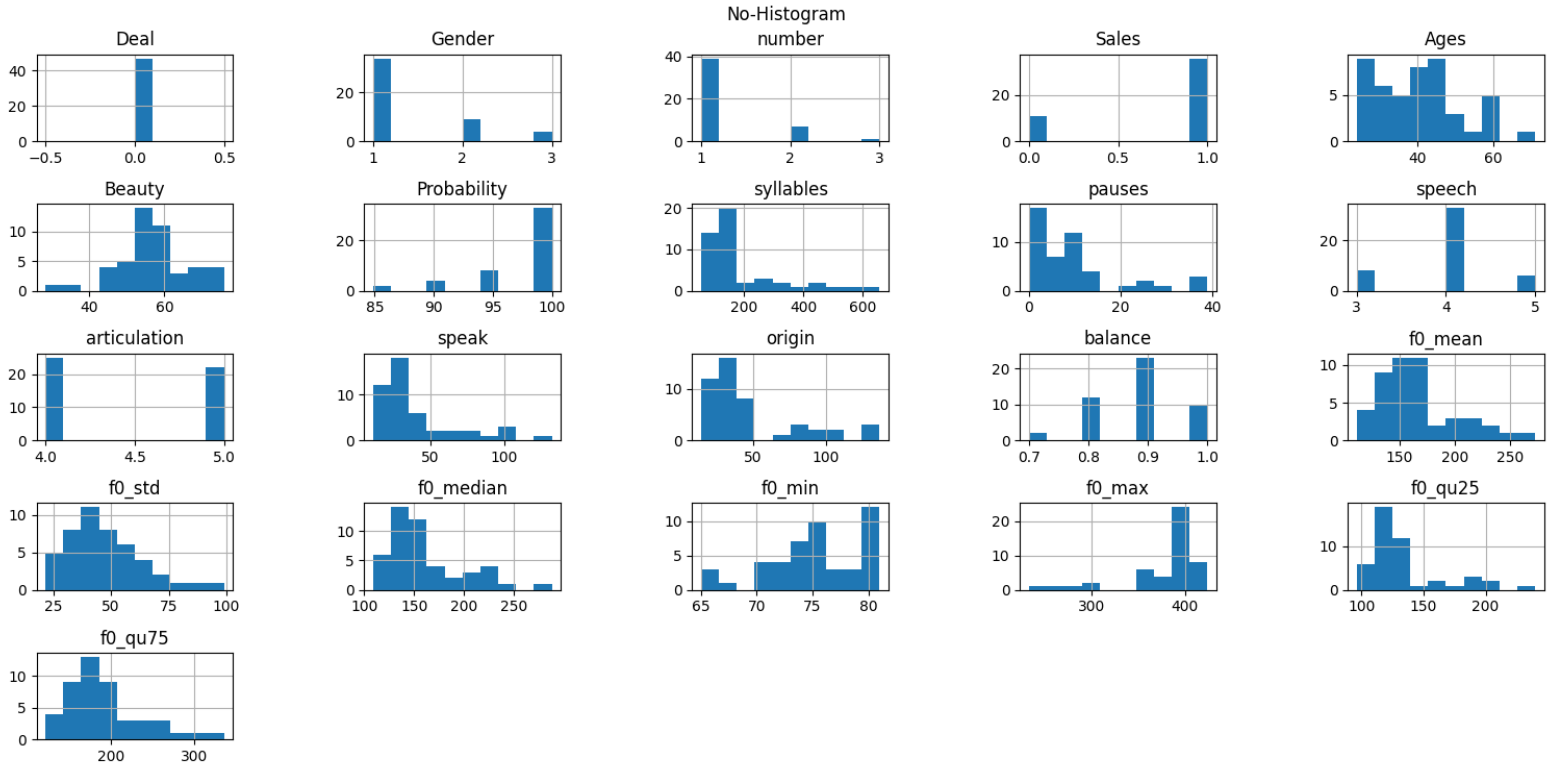


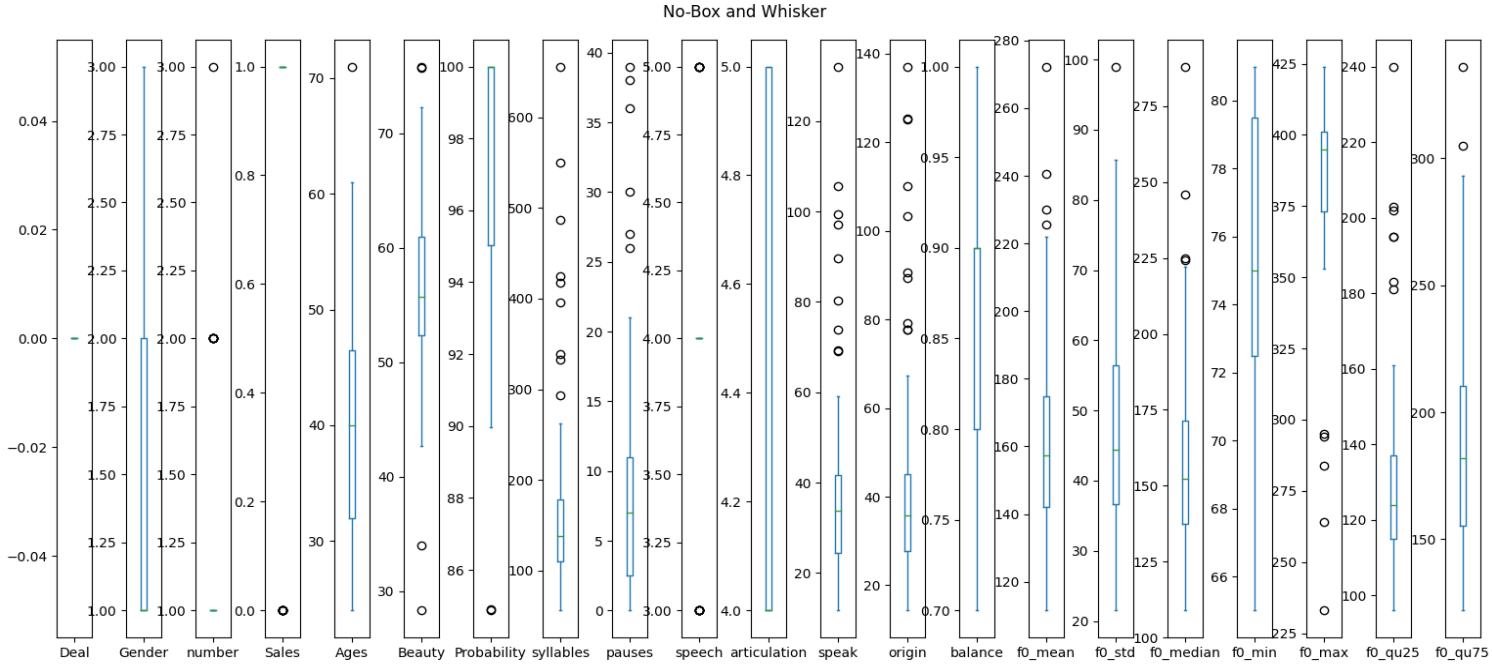
This plot describes exactly for all the industries how many teams did or did not get a deal.

Student's Report. Used for educational Purposes.

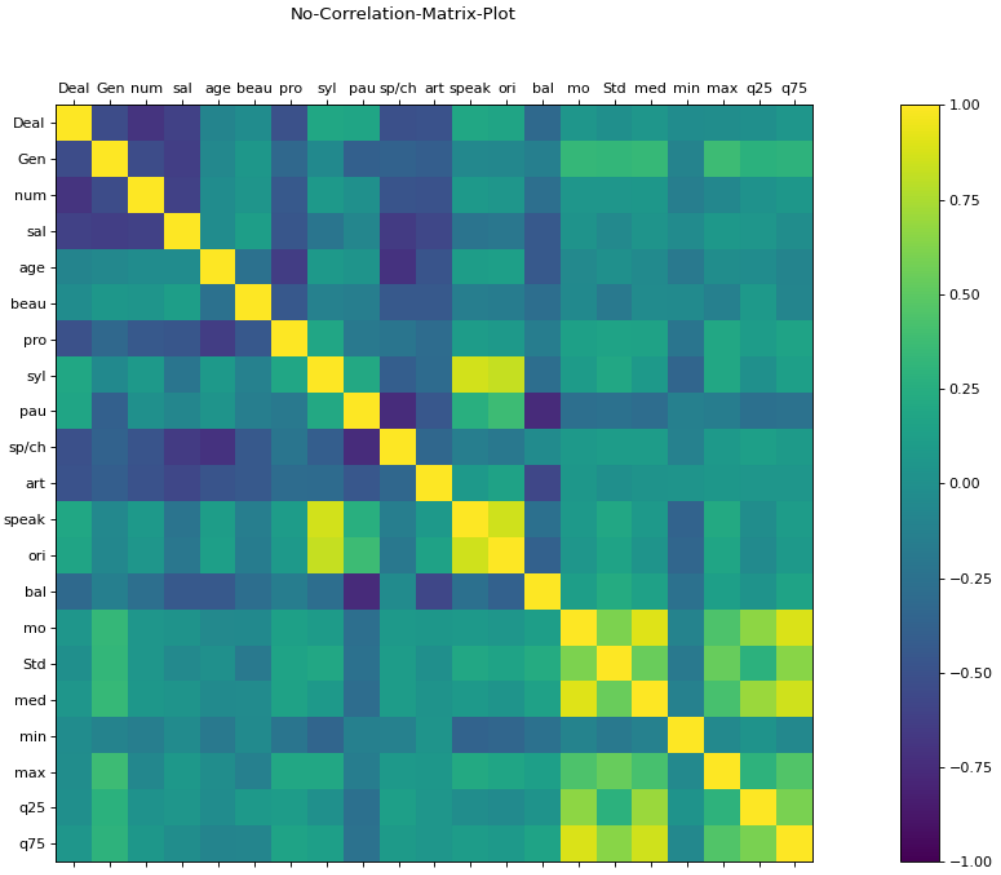
Analysis of voice characteristics from entrepreneurs' pitches

The following plots are the histogram, box and whisker and the correlation matrix for the teams that did not get a deal.





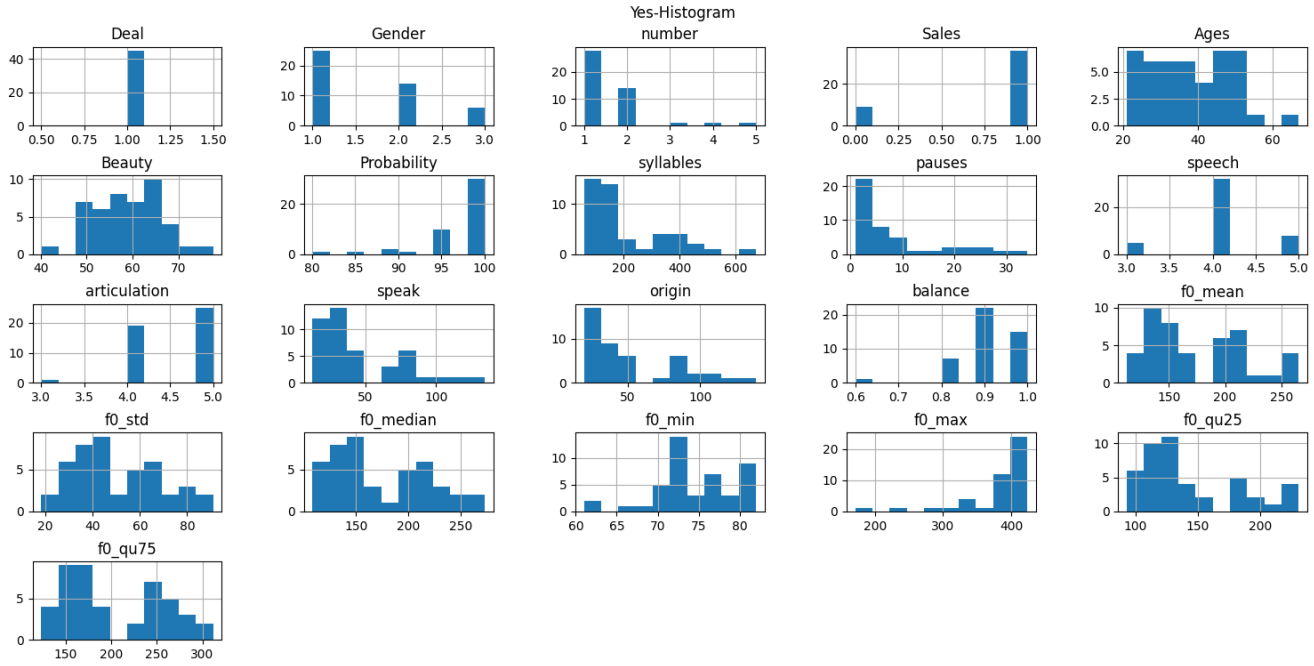
Analysis of voice characteristics from entrepreneurs' pitches

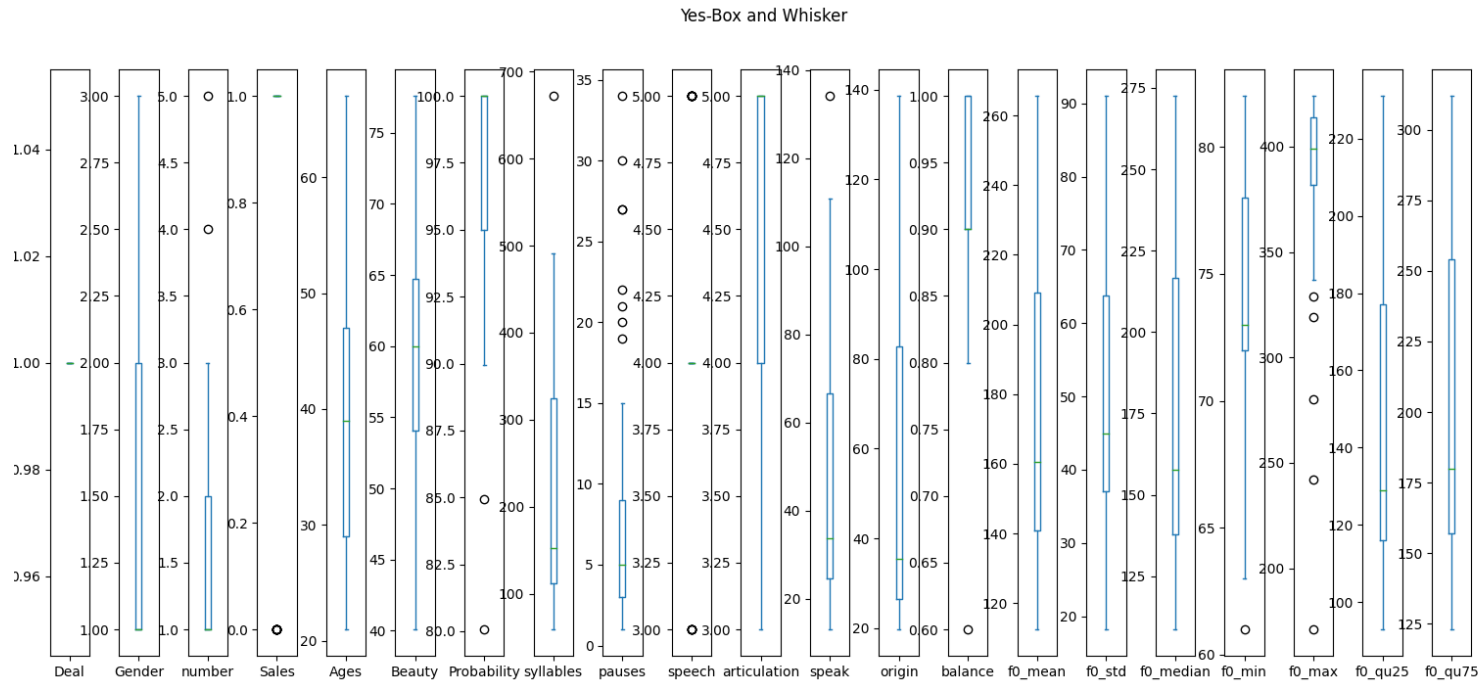


Student's Report. Used for educational Purposes.

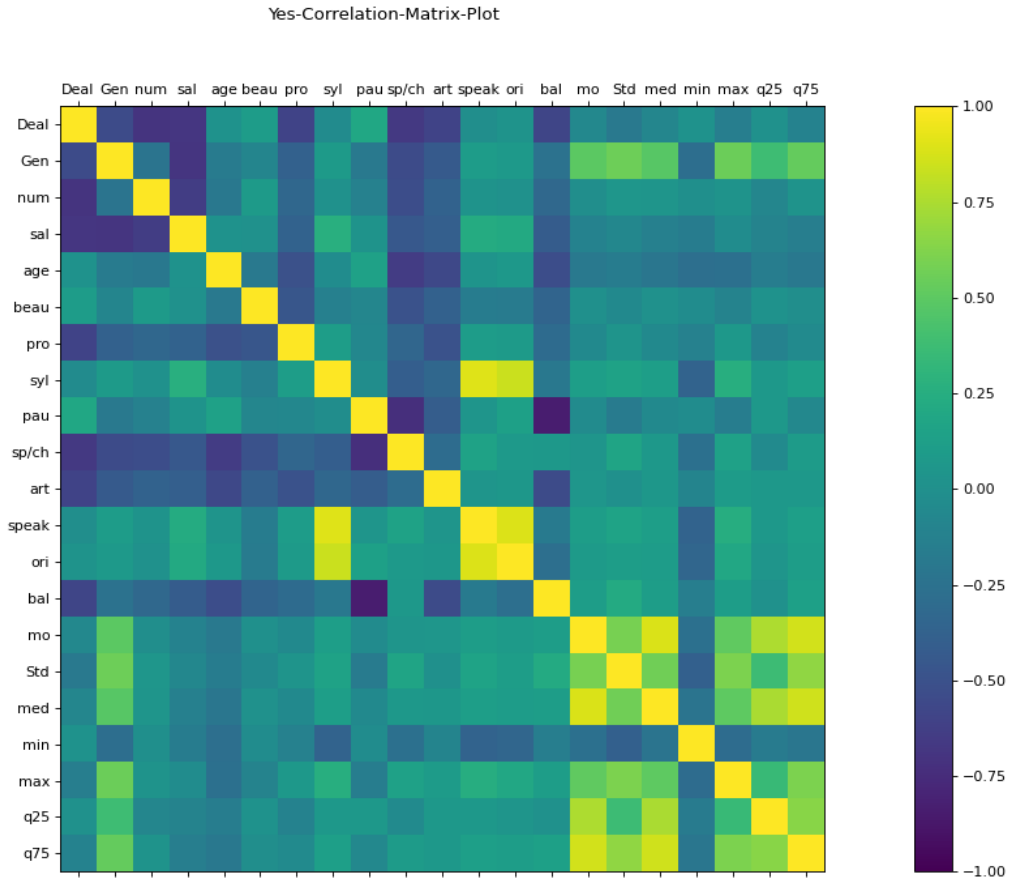
Analysis of voice characteristics from entrepreneurs' pitches

The following plots are the histogram, box and whisker and the correlation matrix for the teams that got a deal.





Analysis of voice characteristics from entrepreneurs' pitches

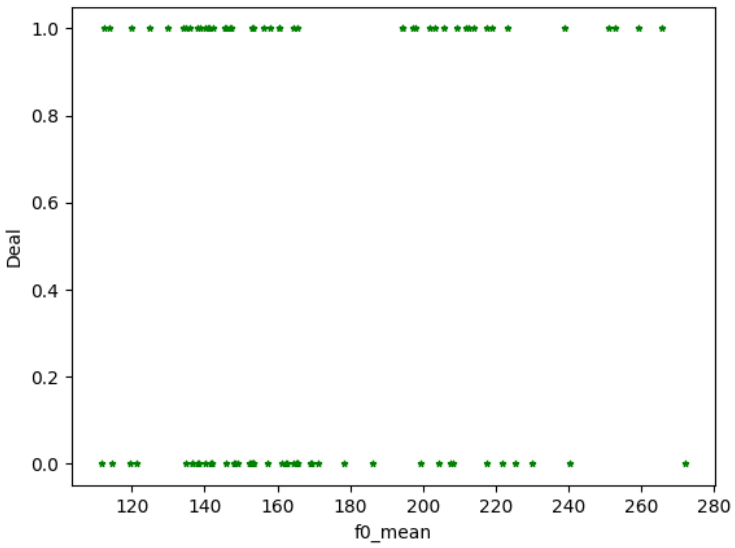


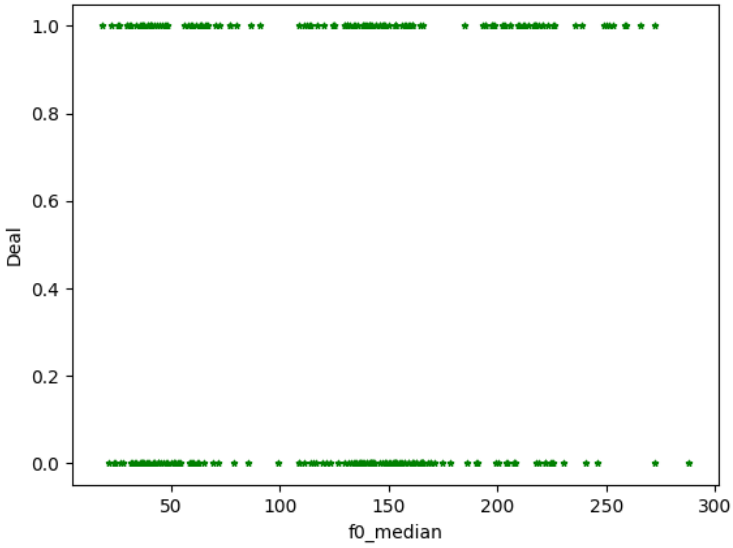
Student's Report. Used for educational Purposes.

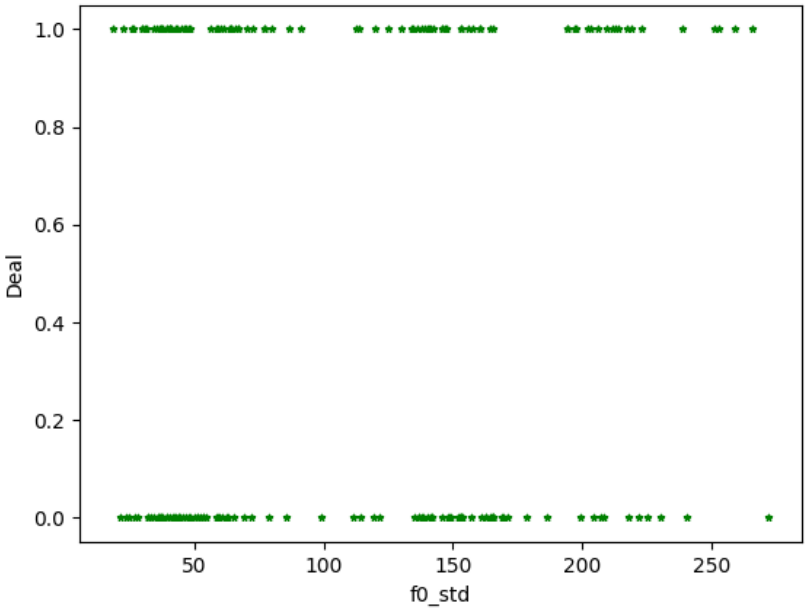
Analysis of voice characteristics from entrepreneurs' pitches

- I created a google sheet with all the correlation matrices.(https://docs.google.com/spreadsheets/d/1jcHFJiroFR9WGLWFPxMVQwAsh2pguXP0JDQRP8tGl_w/edit?usp=sharing) . It contains the correlation among Deal and all the others features(Corr_DEAL), the correlation among all the features only for teams that got a deal(Corr_Yes) ,did not take a deal(Corr_No) and the correlation among all the features for all the teams , that did and did not make a deal(Corr_ALL).

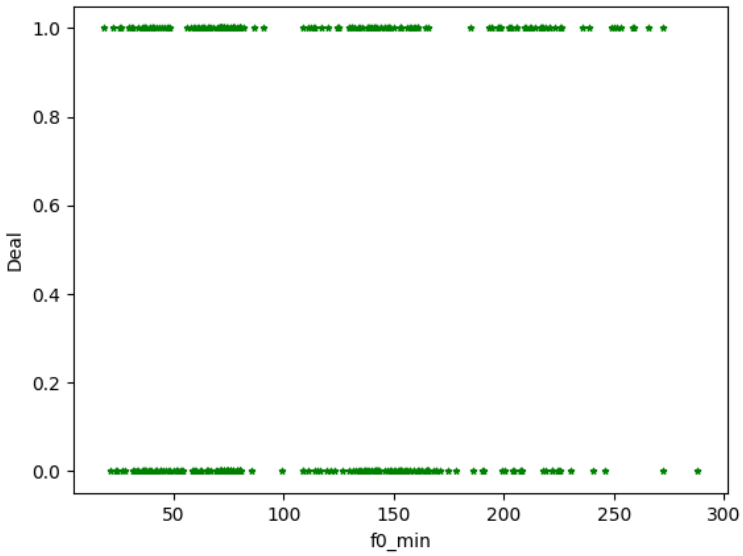
- I created plots for all the f0 statistics to check if there were specific values for which there were only deals or no deals, but I couldn't come to that conclusion.

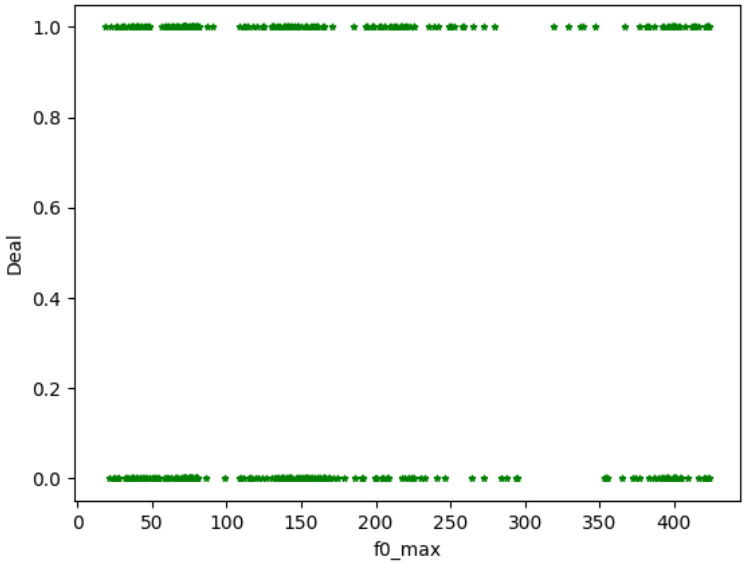


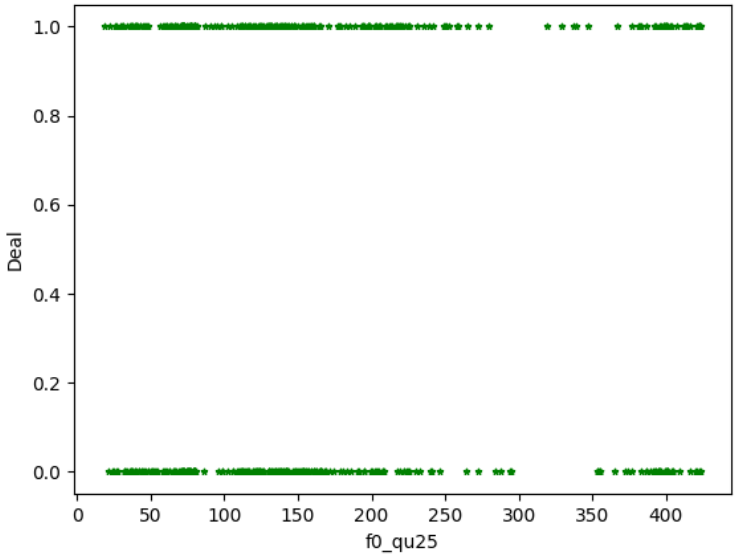


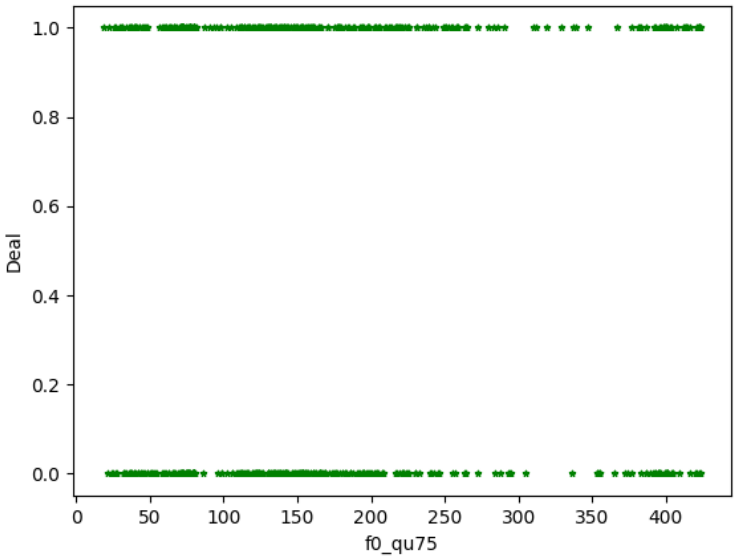


Analysis of voice characteristics from entrepreneurs' pitches









Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

- I detected Multicollinearity with VIF.

- I decided to delete the features syllables, pauses, speak, origin, balance because they are based on the way I cut the videos so they couldn't really help a model to predict if someone made or didn't make a deal.

- I made t-test analysis for the f0 statistics variables with the deal(yes/no) variable

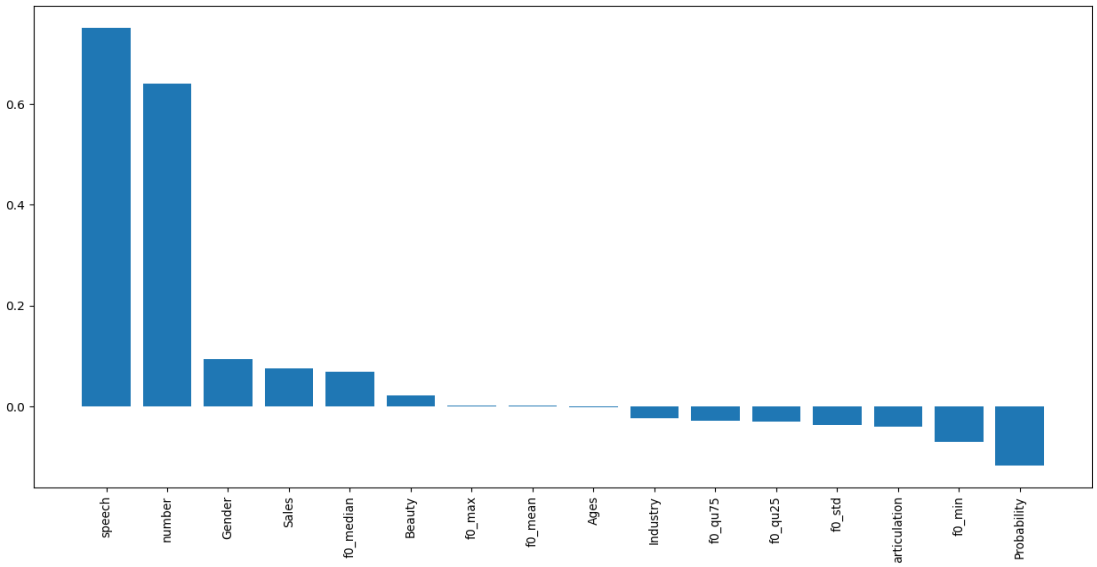
f0 statistics	p-value
f0 mean	0.27
f0 std	0.57
f0 median	0.19
f0 min	0.35
f0 max	0.63
f0 qu25	0.29
f0 qu75	0.30

- I calculated the importance of the variables(scores) with the model logistic regression.

Features	Scores
speech	0.750
number	0.640
Gender	0.094
Sales	0.075
f0 median	0.070
Beauty	0.023
f0 max	0.002
f0 mean	0.001
Ages	-0.002
Industry	-0.023
f0 qu75	-0.028
f0 qu25	-0.031
f0 std	-0.037
articulation	-0.041
f0 min	-0.070
Probability	-0.118

Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches



Analysis of voice characteristics from entrepreneurs' pitches

- I also calculated the importance of the features with logistic regression using statsModel which provided me a larger data volume.

Current function value: 0.617150
Iterations 6

Logit Regression Results

Dep. Variable	Deal	No. Observations	92
Model	Logit	Df Residuals	76
Method	MLE	Df Model	15
Date	Thu, 15 Jul 2021	Pseudo R-squ	0.1093
Time	14:32:24	Log-Likelihood	-56.778
converged	True	LL-Null	-63.748
Covariance Type	nonrobust	LLR p-value	0.5301

Feature	coef	std err	z	P>|z|	[0.025	0.975]
Industry	-0.0292	0.065	-0.451	0.652	-0.156	0.098
Gender	0.0913	0.556	0.164	0.870	-0.999	1.181
number	0.7877	0.483	1.632	0.103	-0.159	1.734
Sales	-0.0090	0.580	-0.016	0.988	-1.146	1.128
Ages	0.0171	0.025	0.680	0.497	-0.032	0.067
Beauty	0.0365	0.029	1.246	0.213	-0.021	0.094
Probability	-0.0820	0.058	-1.402	0.161	-0.197	0.033
speech	0.9393	0.690	1.361	0.173	-0.413	2.292
articulation	-0.1292	0.539	-0.240	0.810	-1.185	0.927
f0 mean	0.0224	0.085	0.263	0.792	-0.144	0.189
f0 std	-0.0425	0.047	-0.913	0.361	-0.134	0.049
f0 median	0.0606	0.045	1.359	0.174	-0.027	0.148
f0 min	-0.0125	0.047	-0.267	0.790	-0.105	0.080
f0 max	0.0035	0.007	0.494	0.621	-0.010	0.018
f0 qu25	-0.0382	0.037	-1.045	0.296	-0.110	0.033
f0 qu75	-0.0308	0.047	-0.650	0.516	-0.124	0.062

Analysis of voice characteristics from entrepreneurs' pitches

Machine Learning

- First, I found the best model's hyperparameters and accuracy using the GradientBoostingClassifier.
    - Best Model parameters:

subsample	0.9
n_estimators	568
min_samples_split	10
min_samples_leaf	20
max_features	4
max_depth	9
learning_rate	0.001

    - Best Model mean accuracy: 0.59

- I used 4 machine learning models to predict if a team will make or will not make a deal with all the features that I studied in the data analysis part of the project. The machine learning models are logistic regression, decision tree, random forest and SVM.
- For the evaluation part I used the 82 rows for training purposes and the other 10 rows for testing.
- I run this part for 10 times, I found the averages and got the following results

Model Name	Accuracy (overall correct predictions)	Auc
Logistic Regression	0.5	0.46
Decision Tree	0.58	0.61
Random Forest	0.57	0.57
SVM	0.40	0.42

Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

Data for value= 0:

Model Name	precision	recall	f1Score
Logistic Regression	0.43	0.75	0.55
Decision Trees	0.49	0.75	0.59
Random Forest	0.48	0.75	0.58
SVM	0.38	0.75	0.5

Data for value= 1:

Model Name	precision	recall	f1Score
Logistic Regression	0.67	0.33	0.44
Decision Trees	0.73	0.47	0.57
Random Forest	0.72	0.45	0.55
SVM	0.5	0.17	0.25

Data for Macro average:

Model Name	precision	recall	f1Score
Logistic Regression	0.55	0.54	0.49
Decision Trees	0.61	0.61	0.58
Random Forest	0.6	0.6	0.57
SVM	0.44	0.46	0.38

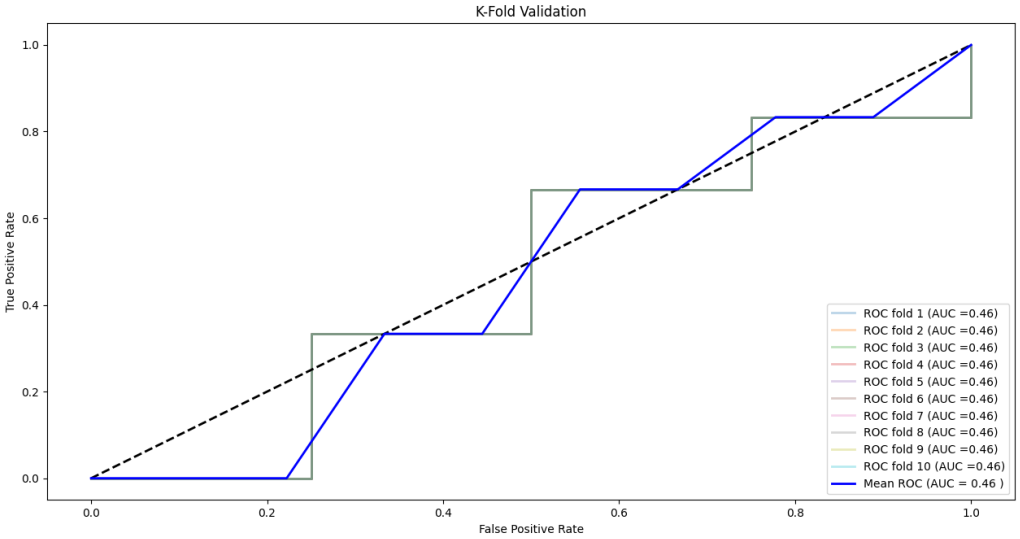
Data for weighted average:

Model Name	precision	recall	f1Score
Logistic Regression	0.57	0.5	0.48
Decision Trees	0.63	0.58	0.58
Random Forest	0.63	0.57	0.57
SVM	0.45	0.4	0.35

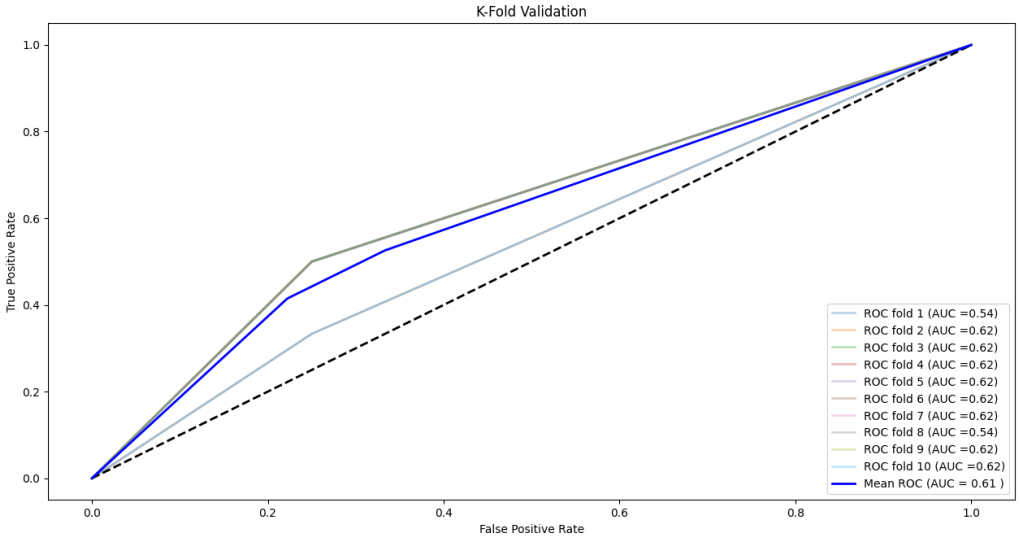
Analysis of voice characteristics from entrepreneurs' pitches

- I also, made the K-FoldValidation with the previous data.

Logistic Regression:



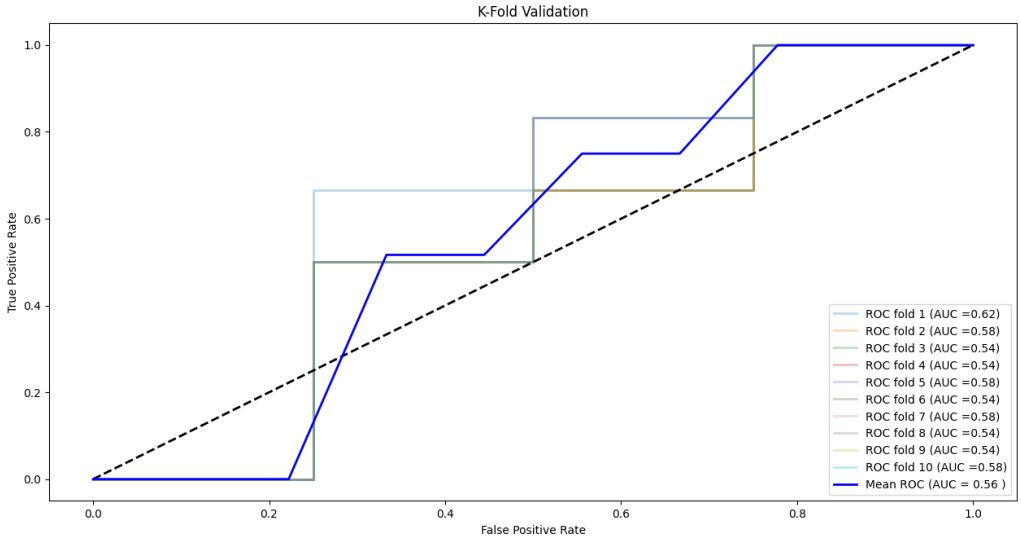
Decision Tree:



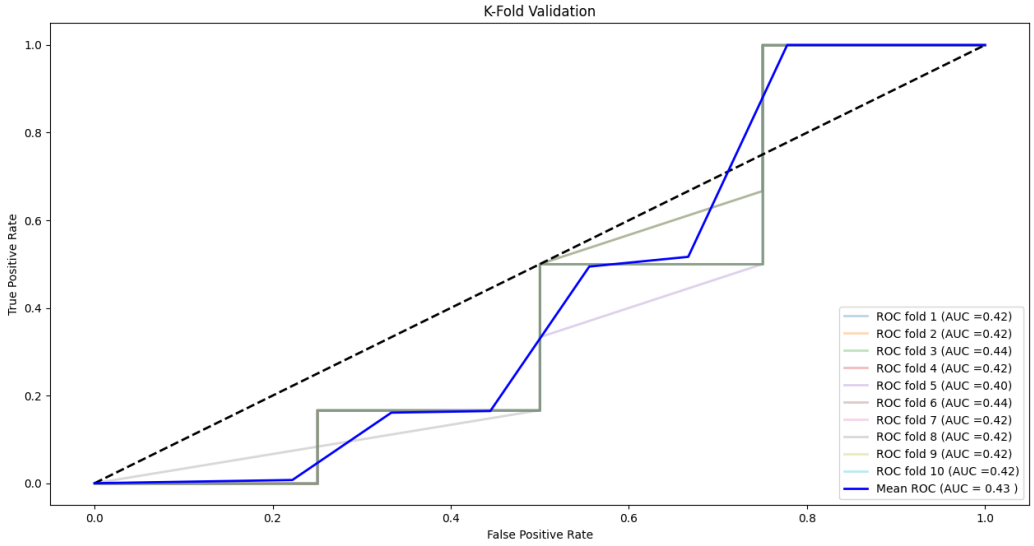
Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

Random Forest:



SVM:



Analysis of voice characteristics from entrepreneurs' pitches

- I also trained the models and made predictions to take the following data and plots

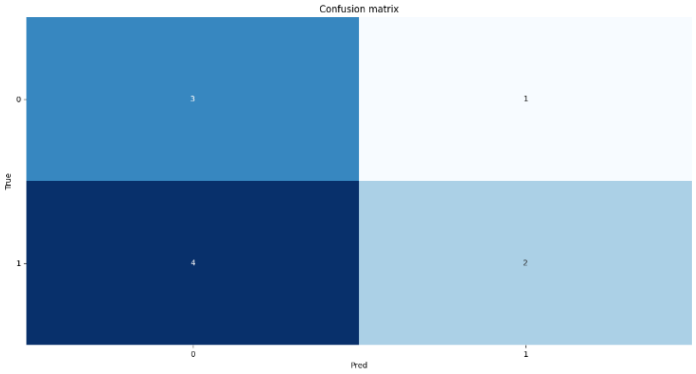
True	Logistic Regression		Decision Tree		Random Forest		SVM	
1	0	0.15	1	1	1	0.51	0	0.47
0	0	0.18	0	0	0	0.42	0	0.52
0	0	0.31	0	0	0	0.48	0	0.51
1	0	0.37	1	1	1	0.51	0	0.52
1	1	0.58	1	1	1	0.52	1	0.46
1	1	0.58	0	0	0	0.46	0	0.51
0	0	0.47	1	1	0	0.44	0	0.52
1	0	0.46	0	0	0	0.45	0	0.51
1	0	0.20	0	0	0	0.44	0	0.51
0	1	0.83	0	0	1	0.54	1	0.44

Student's Report. Used for educational Purposes.

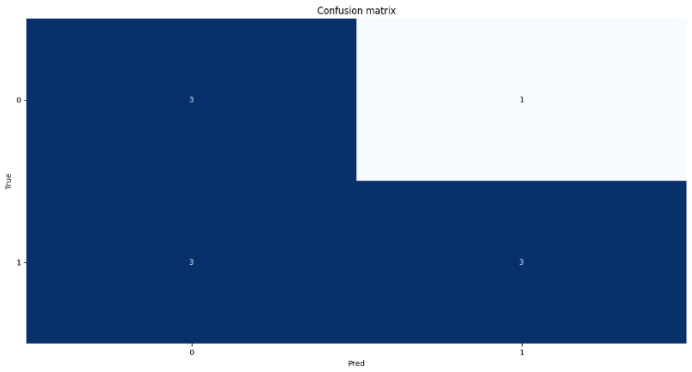
Analysis of voice characteristics from entrepreneurs' pitches

- Confusion Matrix:

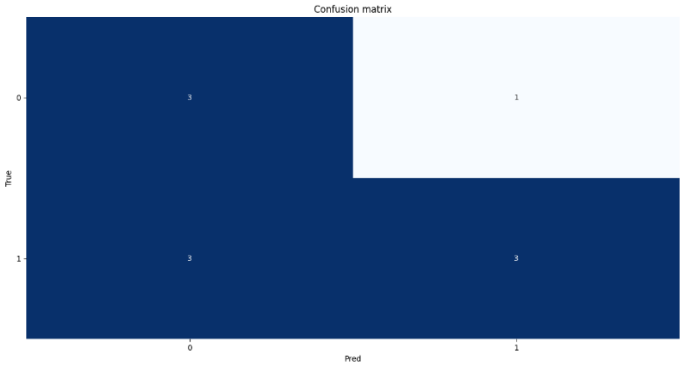
Logistic Regression



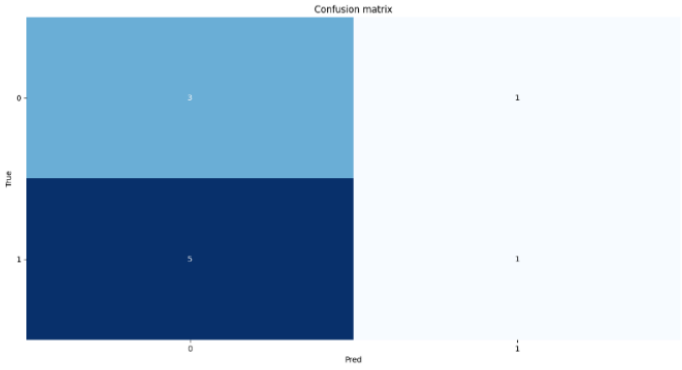
Decision Tree



Random Forest



SVM

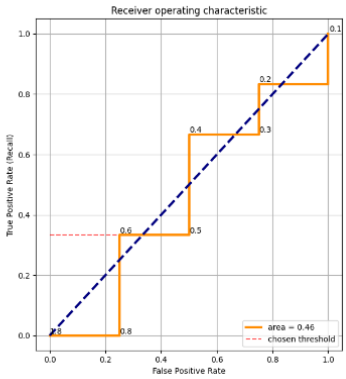


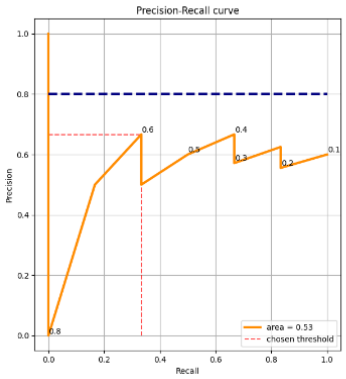
Student's Report. Used for educational Purposes.

Analysis of voice characteristics from entrepreneurs' pitches

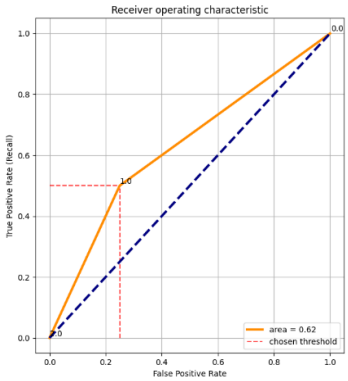
- SVC-ROC-Precision-Recall-Curves

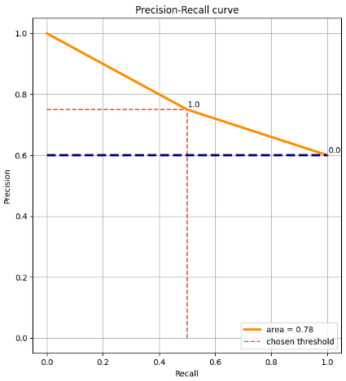
Logistic Regression



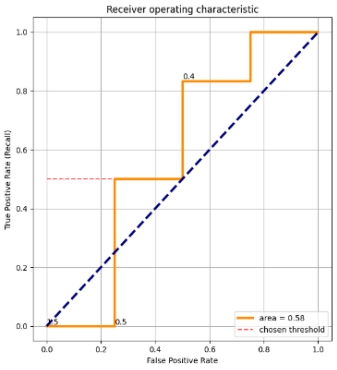


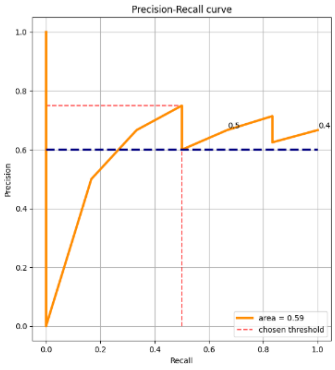
Decision Tree



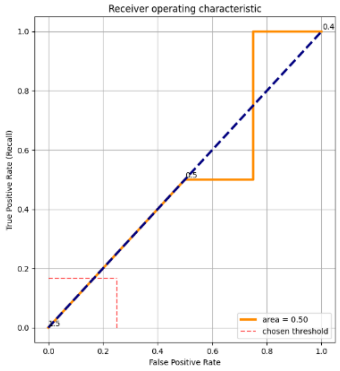


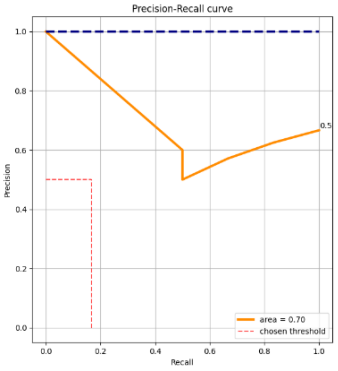
Random Forest





SVM



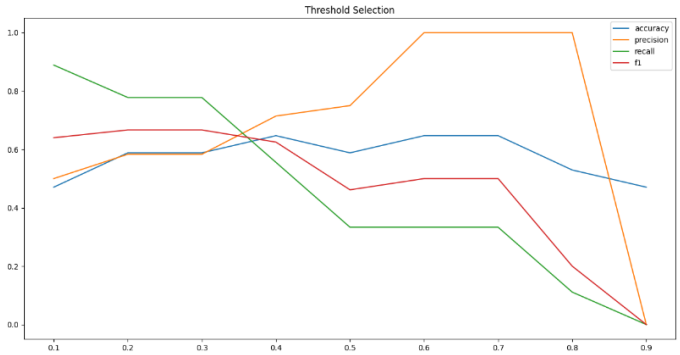


Student's Report. Used for educational Purposes.

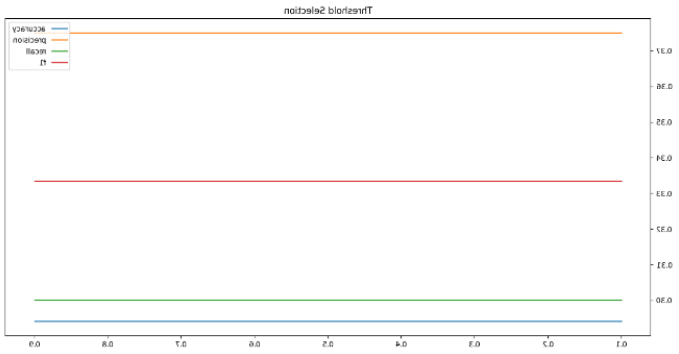
Analysis of voice characteristics from entrepreneurs' pitches

- SVC-ThresholdSelection

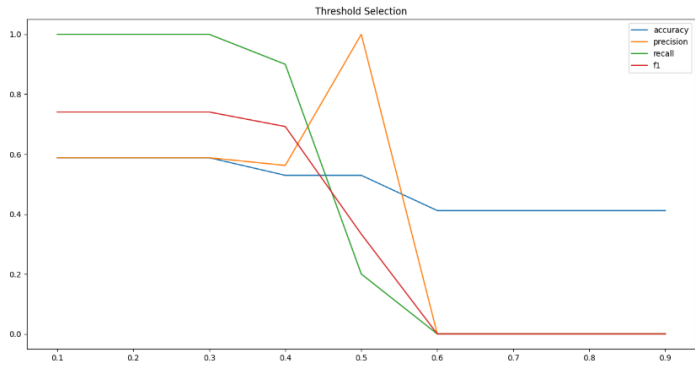
Logistic Regression



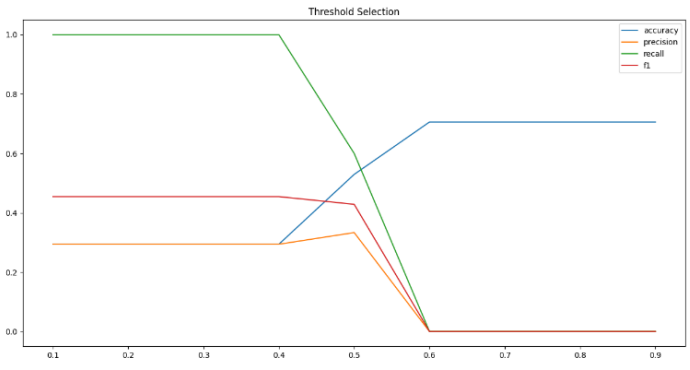
Decision Tree



Random Forest



SVM

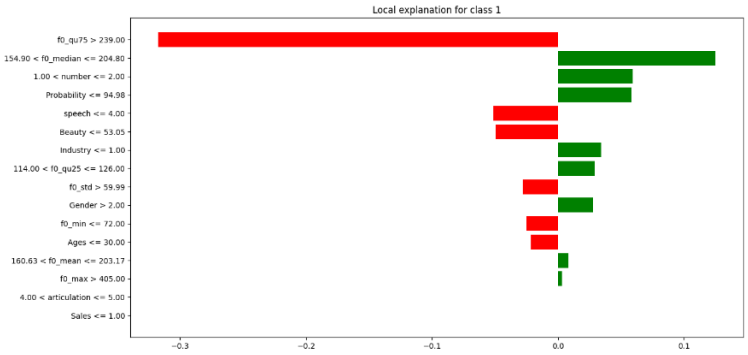


Student's Report. Used for educational Purposes.

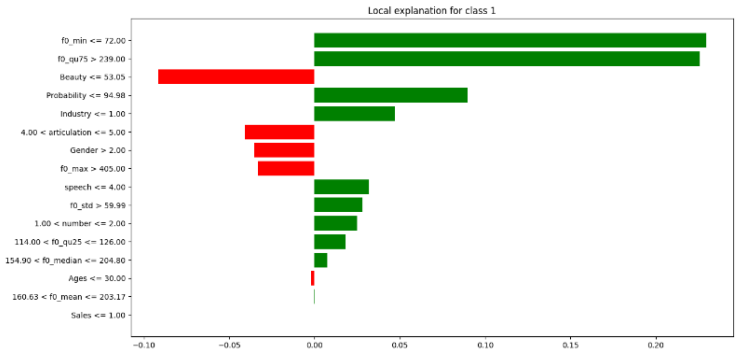
Analysis of voice characteristics from entrepreneurs' pitches

- SVC-Local Explanation for prediction 2

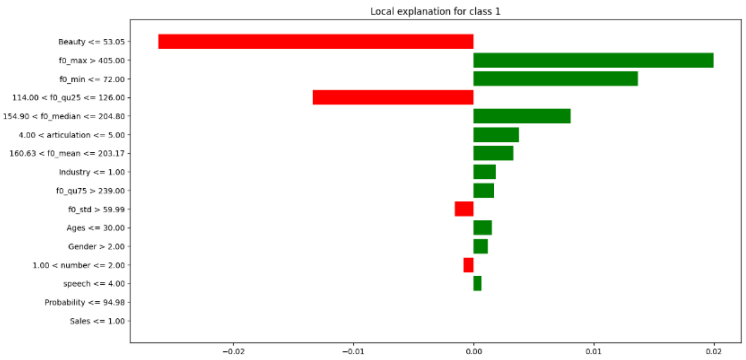
Logistic Regression



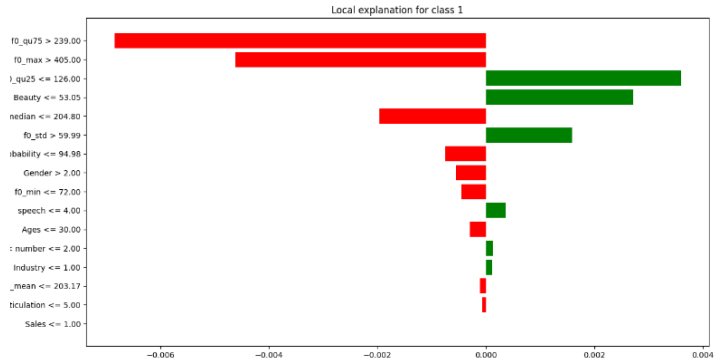
Decision Tree



Random Forest



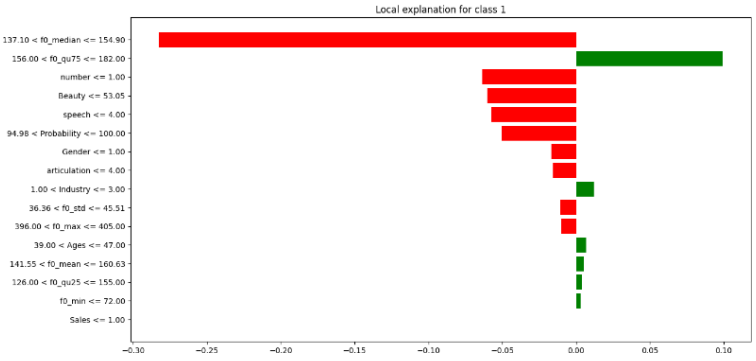
SVM



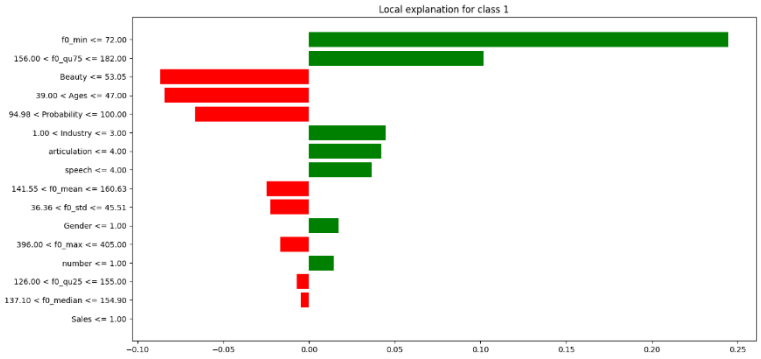
Analysis of voice characteristics from entrepreneurs' pitches

- SVC-Local Explanation for prediction 2

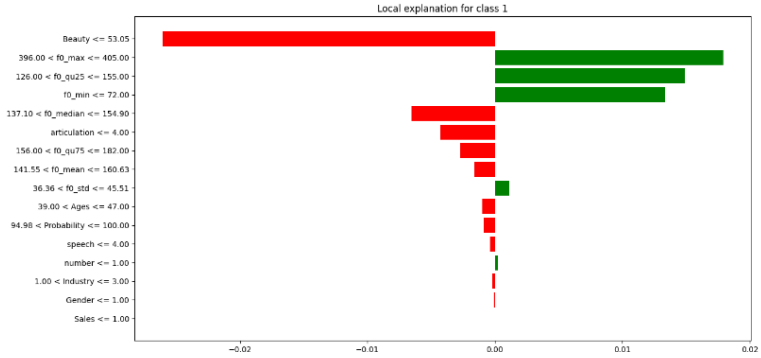
Logistic Regression



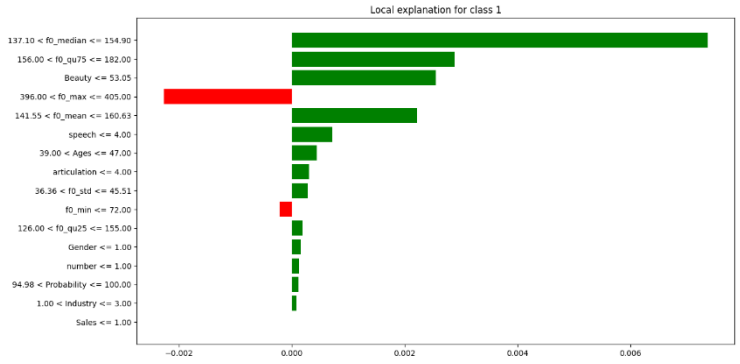
Decision Tree



Random Forest



SVM

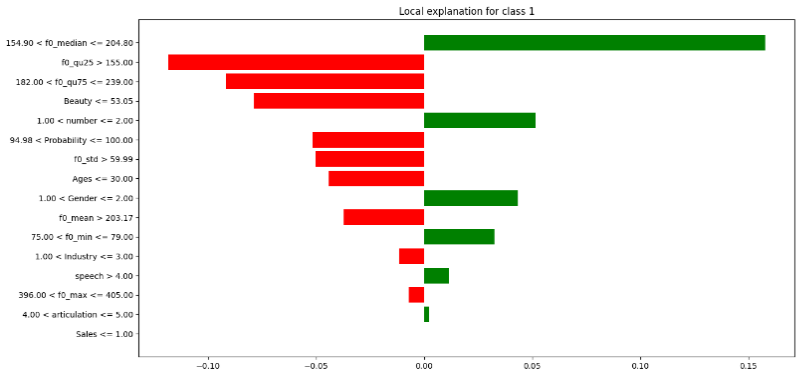


Student's Report. Used for educational Purposes.

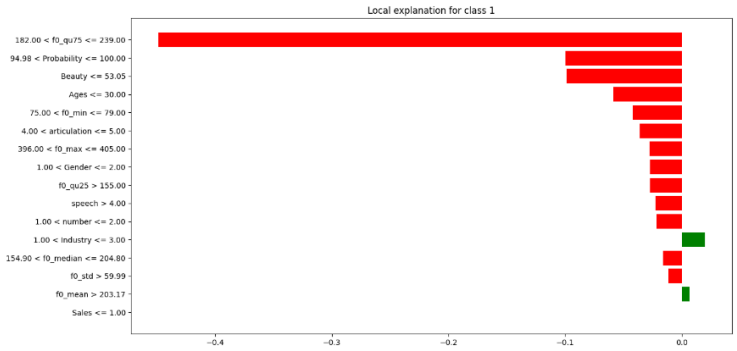
Analysis of voice characteristics from entrepreneurs' pitches

SVC-Local Explanation for prediction 4

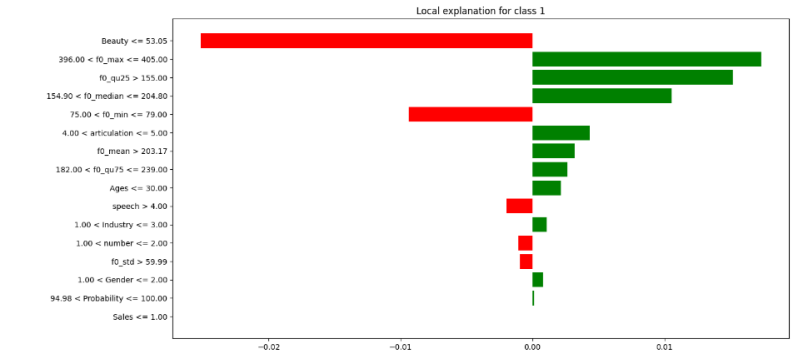
Logistic Regression



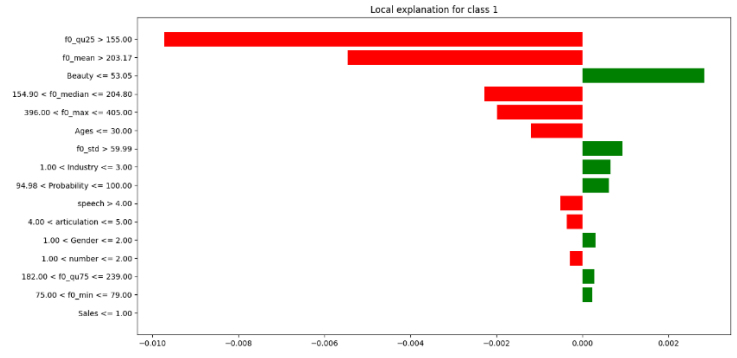
Decision Tree



Random Forest



SVM



Conclusion

The best models , that predict if a team will make or will not make a deal with the best accuracy,are the decision tree and the random forest. Unfortunately, I cant be certain about that as I used very little amount of data. But, the analysis I made and the results I got are really promising.