Shark Tank

Prediction Report

Contents

1. Introduction

1.1 Abstract

This report is an analysis of the shark tank's show. We first decided to analyze the important features of a business success on the show with random forest and Principal component analysis. The latter will provide us insight of criteria future candidates to the show might want to look at before their pitch. Then, we would conduct Classification techniques such as linear discriminant analysis and logistic regression to determine the likelihood of a business's success. In the third phase, we tried to validate our first model's result .

1.2 Our motivation

Being an entrepreneur is the hottest path to follow nowadays. This trend is fueled by the common desire of "becoming your own boss" shared by its adherents and the freedom that comes with it. The Shark tank's show allows SMEs to be guided and financed by "sharks" (angel investors). Whether you have an idea of creating a phone soap charger or a foldable magnetic lamp, the possibilities are infinite on the show. However, an outstanding observation was discovered when we looked at the dataset of the show. 42% of the business deals proposed to sharks which got rejected are still running in 2019. (Our dataset goes from 2009 to 2015). The sharks are missing out on lucrative deals, and we want to fix this problem and allow sharks to take enlightened decisions during the show.

1.3 Problem statement

We want to answer this problem: What is the survival rate of businesses participating in the show? Our goal is to predict if entrepreneurs' business are going to close their doors or will successfully thrive at least 4 years after their appearance in the show. To clarify a point, we refer at the survival rate as the likelihood to stay in business in the long run . Our model will be a useful tool for the sharks since they seem to be wrong 42% of the time.

2. Data Description

2.1 Data dictionary

Dependant variable

- still open - if the deal survived post getting a closing deal or even after not getting a deal

Deal Characteristics

- deal - if the entrepreneurs got a deal or no deal
- description - about what the business idea is
- Len_desc - length of description
- catChanges - business idea category
- Website - official website having details about the product
- exchangeForStake - amount exchanged for business proposal
- Valuation - what is the proposed value of entrepreneurs business
- Title - Name of the business
- in exchange for - amount asked for in exchange of company stakes
- implied valuation - what is the actual value of entrepreneurs business
- Dif_equity - difference in what was pitched and what was actual equity exchange for sealing the deal
- Dif_valuation - difference in what was pitched and what was actual value for sealing the deal

Entrepreneur Characteristics

- Entrepreneurs - name of entrepreneurs who appeared on the show
- Gender - gender of entrepreneurs
- multiple entrepreneurs - when entrepreneurs came in a team instead of coming as individuals

Episode Characteristics

- Episode - episode number
- season - season number
- Year - year in which a particular season and episode was aired

Shark Characteristics

- Shark1 - name of one of the first shark in the deciding panel

- Shark2 - name of one of the second shark in the deciding panel
- Shark3 - name of one of the third shark in the deciding panel
- Shark4 - name of one of the fourth shark in the deciding panel
- Shark5 - name of one of the fifth shark in the deciding panel

Demographics

- Location - where the business is/was originated
- State - where the business is/was originated
- Region - clubbing each of the state into 4 geographically separated regions

2.2 Data pre-processing

The original Shark Tank data set had a total of 19 features. To enhance the scope of our project and make the data set cleaner and richer we performed the below steps as part of data pre-processing:

First, we web scrapped and added the below columns, Second, using values in "entrepreneurs" column we built another column gender having values "M", "F" and "X" for male entrepreneurs, female entrepreneurs and mixed (team) entrepreneurs respectively Third, we transformed "location" column by considering only last characters representing state Then, we clustered column "category" by binning similar categories like "Women's Accessories", "Women's Apparel", "Women's Shoes" into one single category "Fashion" Finally, we category and state values for PCA to function

2.3 Data overview

Pictorial representation of data is one of the best ways to understand the data set. Utilising the same idea, we tried understanding the nitty gritty of the data by plotting a graph for each of the predictors involved. One might think that Shark Tank is just a show with just a handful of risk takers trying their luck. However, that is not the case, in fact it is one of the most popular shows with tons of aspiring entrepreneurs trying their hardest to their business ideas. Due to this reason there is always a huge turn around of budding entrepreneurs with an almost equal number of rejection and acceptance of deals. The same was depicted by graph 1.1 in the appendix. This arose the question: Are the sharks signing contracts randomly? Graph 1.2 shows how the popularity of the show impacted more business driven minds to present their ideas which resulted in more number of episodes over a period of time. Issues like gender equality and women's empowerment still remains a topic of concern. As seen in graph 1.3, Shark Tank had a starling high number of male entrepreneurs as compared to women or mixed team entrepreneurs Interestingly, Los Angeles left the Silicon Valley, San Francisco behind by having the most number of entrepreneurial minds coming on the show and pitching for their ideas. The same is exemplified by graph 1.4. Having one of the best platforms and the brightest bunch of Sharks supporting you, does not necessarily mean that your business would flourish altogether. Even after getting the money they asked for, some of the deals turned out to be unsuccessful. The same can be seen from graph 1.5 which shows the count of businesses still surviving or no.

2.4 Features Selection

2.4.1 Intuition

We do not think that sharks(judges), the season and the episodes are indicators of the future success of a business. Plus, we can clearly say that valuation and difference in valuation are correlated, which

is expected because as mentioned above diff_valuation has been manually added which takes into account the difference in valuation asked for and the actual valuation. We are not going to consider diff_valuation for this reason. (See Appendices)

2.4.2 PCA

To enhance the readability for the user, we performed PCA with a bunch of options which helped us explore the possibility of correlation and reveal some unexpected correlated pairs. Firstly, it was performed on the data set having no dummified predictors i.e. on the original dataset with only numerical data. It was noticed that deal and difference in equity were highly correlated. Secondly, we performed PCA on the same data set but only concentrating on states by dummifying all of the state variable. Since the number of variables multiplied tremendously, the plot was visually hard to interpret. We decided to cluster states into regions to make the interpretation easier. It was noticed that episode and year and deal difference in equity were highly correlated. Lastly, PCA was performed with the original dataset with nothing being dummified except categories. We notice that episode season and deal and difference in equity were highly correlated. The same is justified by graph 2.1, graph 2.2 and graph 2.3 respectively in the appendix section.

2.4.3 Random Forest

Random forest is a must have when undergoing into a feature selection process. The results obtained for the best predictors are:

Predictors	Gini coefficient
website	0.118554168
year	0.107581859
dif_equity	0.078182056
len_desc	0.06221787
title	0.057363888
category	0.054830755
entrepreneurs	0.040821454
location	0.039956272
valuation	0.028807762

As illustrated by the random forest, having a website is a determining factor of long term success. Equity and valuation are also good indicators of a wealthy business. Choose the right business at the right place is essential, because the category and the location where you operate your business should not be ignored. The length of words used to explain the business model is important too. According to us, year should not be considered because random forest overestimate its importance (as year increases, we have more episode and then more businesses in the shark tank show). Entrepreneurs and title won't be selected as predictors in our model because there are too specific information which will over-fit our model. As we know that sharks are not able to spot the great businesses, we won't be considering the deal predictor. Based on our previous analysis in the data description, we believe we need to take into account the gender of the entrepreneur and the number of entrepreneurs managing each business. Finally, we are not going to consider the predictors which were correlated in our PCA analysis in order to have an accurate model.

3. Model Selection

3.1 Classification models

Our dependent variable is a binary variable, so we decided to explore which classification method suits the best our dataset.

3.1.1 Linear discriminant analysis

For our first classification model we built a LDA model.We considered for this model. valuation, len_desc and exchange.for as variables. We split the data set into a training set (80%) and a test set (20%). We normalize the data set and transform the training test sets accordingly. Then, we used the lda function to fit our model with the training data set transformed. Finally, we predicted our data from the test transformed set. The procedure we later followed for quadratic discriminant analysis (QDA) did not differ from the LDA except fitting the model with the qda function. Also, we notice that the error rate was lower with QDA(24.7%) , so we automatically stop considering the LDA model for our purpose (26.4%). But can we consider the QDA, based solely on its better performance? Linear discriminant analysis assumes all predictors are continuous. Knowing the high predominance of categorical variables in our data set, we chose to analyze another model.

3.1.2 Logistic Regression

As our first model was limited to continuous variables and given the fact we had numerous categorical variables; we built a logistic regression. Plus our predicted variable is dichotomous and applies more to Logistic regression compared to LDA which prefers nominal predicted variable. To build this model, we split the data set into a training set (70%) and a test set (30%). We do not normalize the data set because our focus was on categorical variables. Before running the logistic regression, we checked the weight of each class and identified our data was imbalanced. Indeed, we had more businesses still operating than closed businesses in our sample. To avoid overfitting, we decided to follow the oversampling balancing method and using the SMOTE implementation from the library DMwR in R. Our data set is limited, so the under-sampling balancing method was not an good option for us. After balancing the data, the probability that a business goes bankrupt after the show was

53% against 27% at the beginning. We run both a logistic regression with valuation and a multiple logistic regression.

3.2 Unsupervised learning algorithm

3.2.1 Clustering method

Our shark data set consists of a lot of labelled categorical data. So, as we tried to create beautiful clusters we realised that clustering categorical data was not a good idea as most of the clustering algorithms do not work well with such type of data. Even algorithms like k means which uses distance, would not work well on data set like ours which consists of majority of categorical predictors. This demonstrates the importance of understanding the data we are using to be able to pick the right algorithm to apply.

3.3 Limitation

The size of the data set was extremely limited and due to imbalanced data and difficulty to have access to more recent data.

4. Results

4.1 Our Model

Our final model is composed of 8 predictors: 2 continuous (the valuation of the business, the equity of the business) and 8 categorical (the entrepreneur's gender and region, the amount of words put in the description of the business, the category and website of the business and the number of entrepreneurs working in group or alone).

4.1.1 Categorical Variables

For the following part, we based our interpretation relative to the automotive industry,thus its acting as a reference. The first coefficient in front of each variables gives us the impact on the probability of predicting if a business is still open in the future. For example, if the category is baby, Toys and Games, we expect the probability to increase by 2.3% and 1.89% if category is Entertainment and events. If the category is fashion, we expect the probability to increase by 1.6% and 2.21% if category is food and beverage. For category health and beauty, we expect the probability to increase by 2% and 1.1% if category is decor and utilities. For category novelties, we expect the probability to increase by 1.8% and 4.7% if category is pet products. For professional services category, we expect the same probability to increase by 1% and 1.1% if category is software, electronics and education. For sports and fitness category, we expect that to increase by 1.3%.

If the entrepreneurs come in a team then it adds more value and its expected that the probability of the business to remain open will increase by 0.633% compared to not having a team. Having a website also adds weightage to this probability and gives a jump of 3.53% compared to a business without any website. Gender is an interesting characteristic to look at; we noticed that being a male entrepreneur decreases the probability of the long term viability of your business by 0.02% and by 0.058% if the team is gender-mixed compared to a female entrepreneur.

Lastly, a proposition coming from the Northeast region will increase the probability of a business being open by 1.4% compared to a proposition coming from the North region. However both propo-

sitions coming from South region and West region would decrease the probability by respectively 1.2% and 2.4% compared to a proposition coming from the North region. So the best promising deals could be found in the Northeast region relative to the North region.

4.1.2 Continuous Variables

The contributions of the continuous variables are less obvious from just looking at the coefficients. They reflect a more complex relationship between a business surviving later on and the respective independent variable. For example, the probability of survival rate decreases by 4.65% if the exchanged amount for stake increases by one unit. The length of description increases the survival rate of a company by 0.003% for each letter added in the business presentation. Our results reveal that valuation do not actually have any impact on the survival of a business in the future. This is a key explanation that could justifies why sharks were not good at finding everytime good deals, since they are heavily considering the valuation when giving an offer to entrepreneurs.

4.2 Cross-Validation

Since we are assessing a classification model, we should look at the accuracy. Our model miss-classified the data 23.37% of the time. We conduct a cross validation analysis with the help of the library Caret in R. We had initially two sets of data: training and test. Unfortunately, due too the difficulty to get recent data we could not add a validation test to this process. We did three type of cross validation: LOOCV, booot and repeatedcv; and surprisingly all of these cross validation gave us the same accuracy we had mentioned before(76.63%). This can be explained by the size of our data, the model learned quickly the training data set and did not require multiple iteration to do it. We believe this is the highest performance we can achieve with the tools we had.

4.3 Conclusion to sharks' attention

Having a great kick start is not the recipe of guaranteed future success. There are a lot of factors which impact the possibilities having a business survive in competitive market. Also, as the time changes, the KPIs also changes and thus it is very difficult to say that what are the most important predictors in determining the same.

After a deep analysis we could answer our problem raised at the beginning of this report. Sharks usually focus their decisions on financial indicators such as valuation and equity; however, sharks should focus on more qualitative attributes such as the diversity in the team. We have learned that having a strong multiple gender team mix do adds value and the capability to survive and shine throughout time.

Another important prospect would be that the north eastern market has quite strong entrepreneurs, when compared to the remaining three regions. Focusing on categories, we could say that Pet products are really worth the investment and would help our sharks fetch a lot of money as compared to other categories.

5. Appendices

Our best single-variable logistic regression

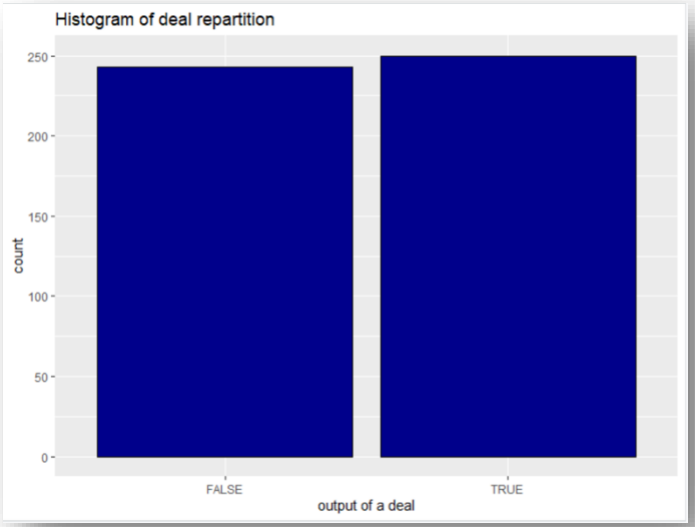
glm(survival valuation,family="binomial", data = train2)

Our best multiple logistic regression

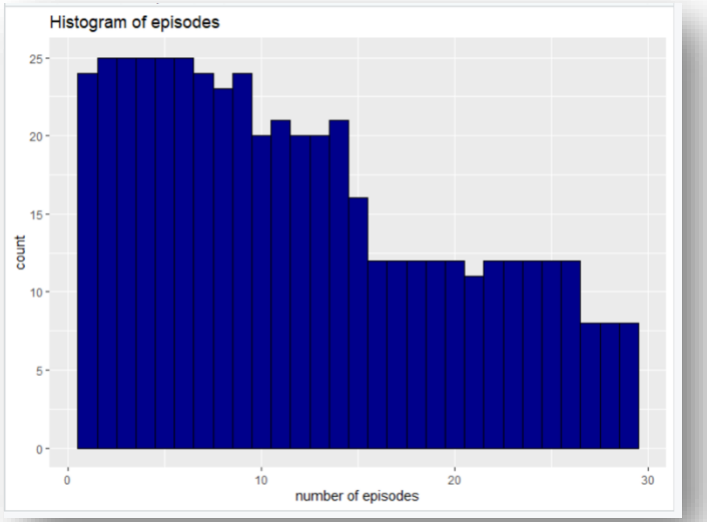
glm(survival valuation+exchangeForStake+gender+len$_desc + catChanges + Multiple.Entrepreneurs + website + region, family = "binomial", data = train2)$

Appendices

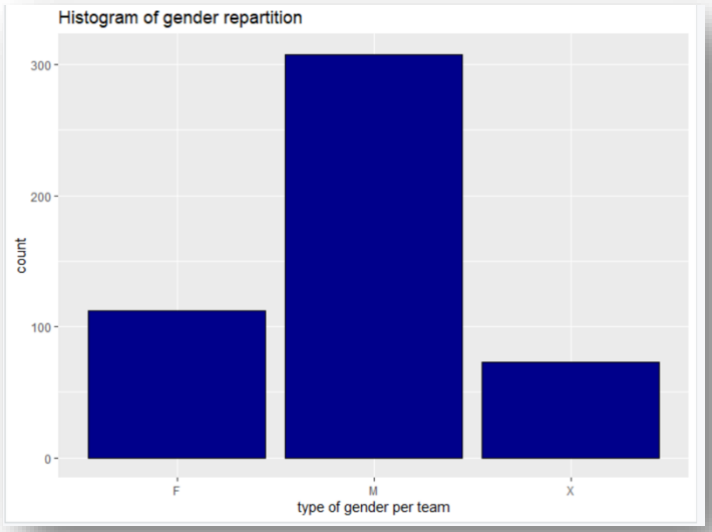
Data overview



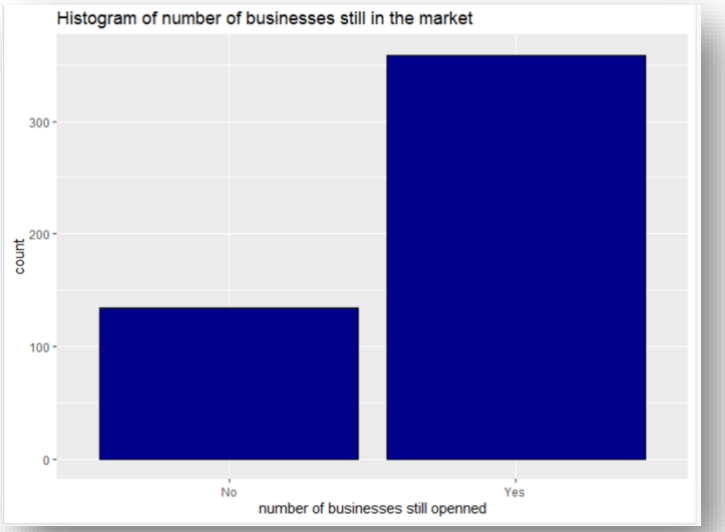
Graph 1.1 : Histogram of deal



Graph 1.2 : Histogram of episodes

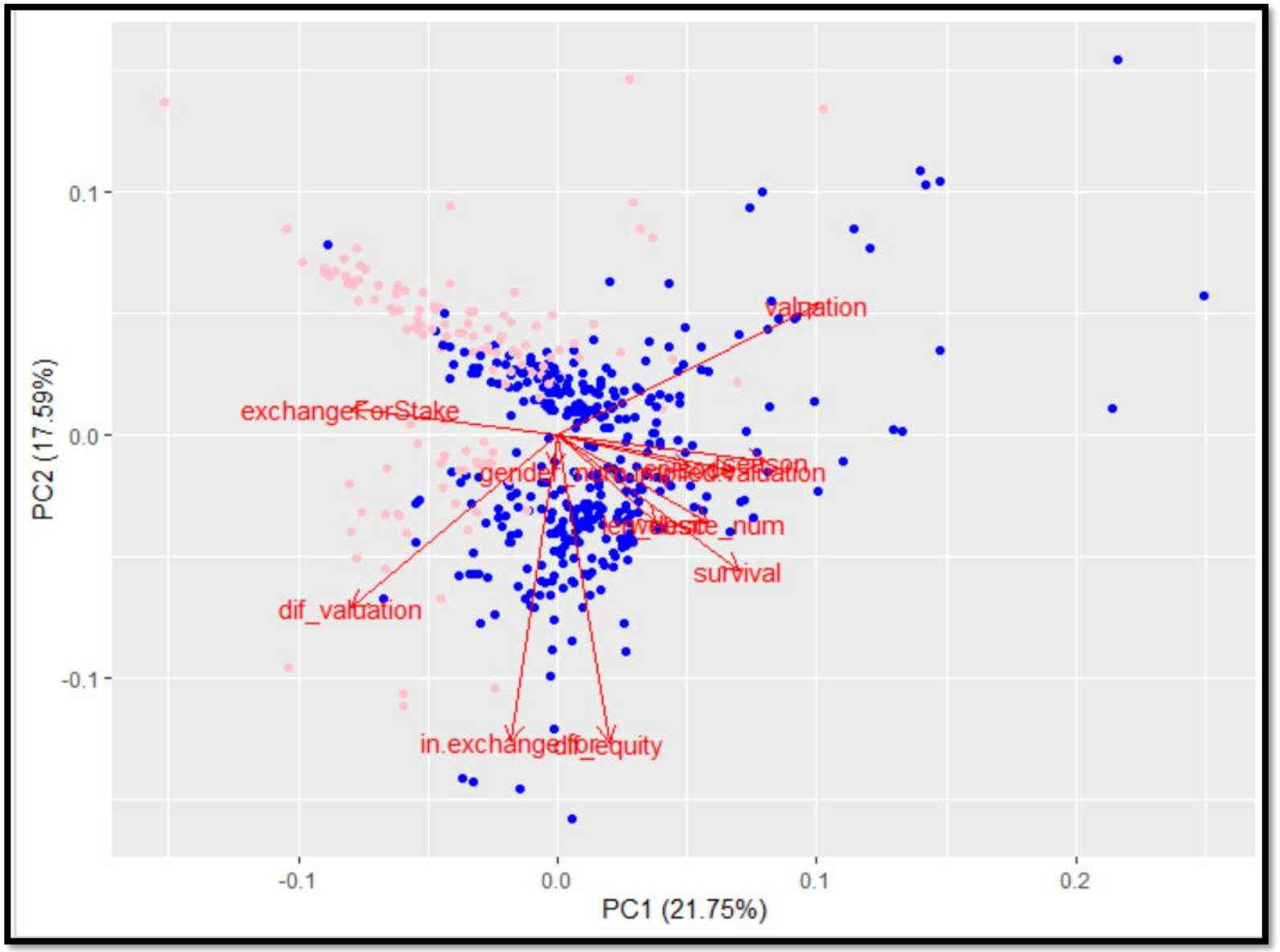


Graph 1.3: Histogram of gender repartition



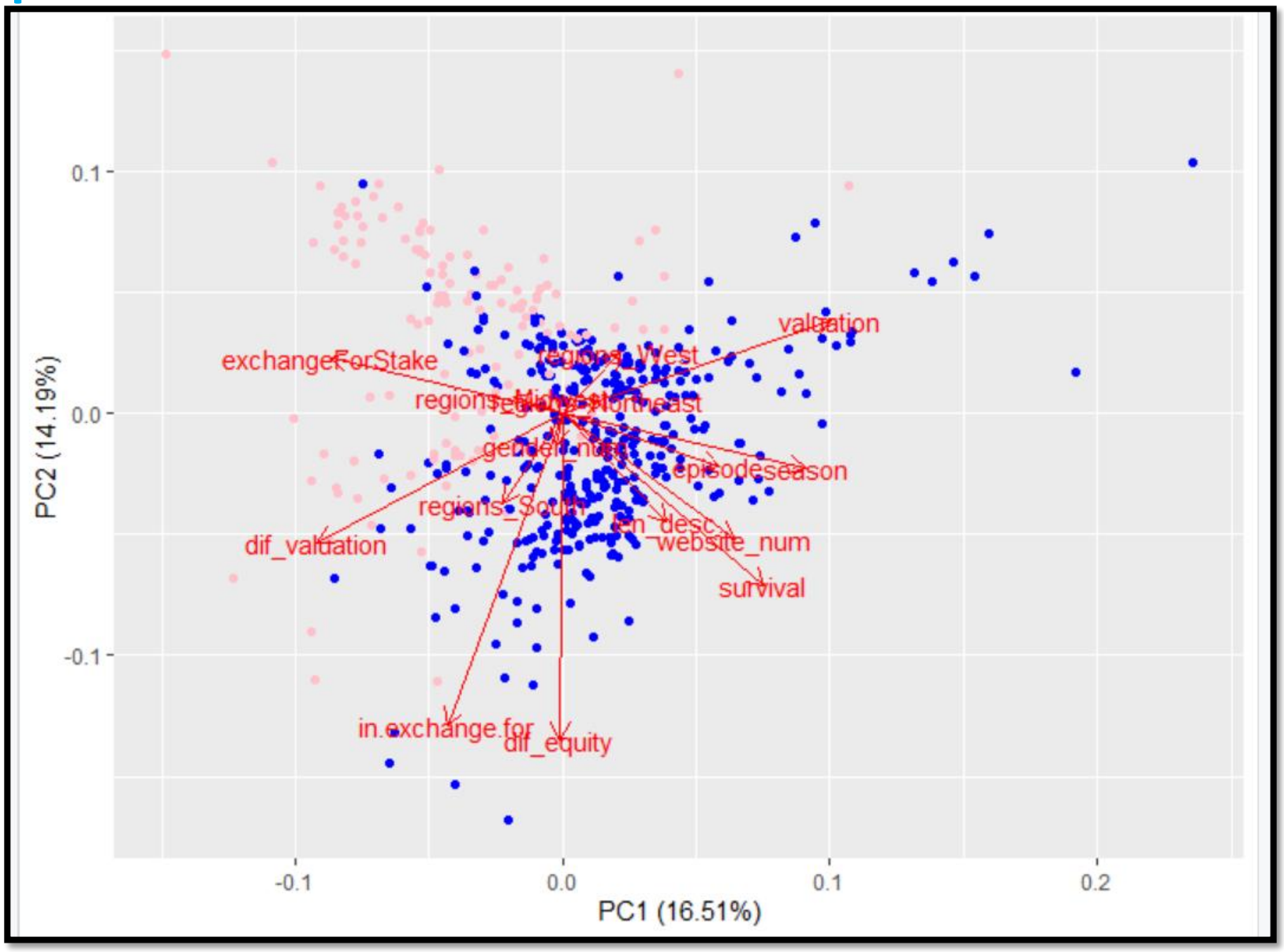
Graph 1.4: Histogram of nb. of business still in the market

Appendices



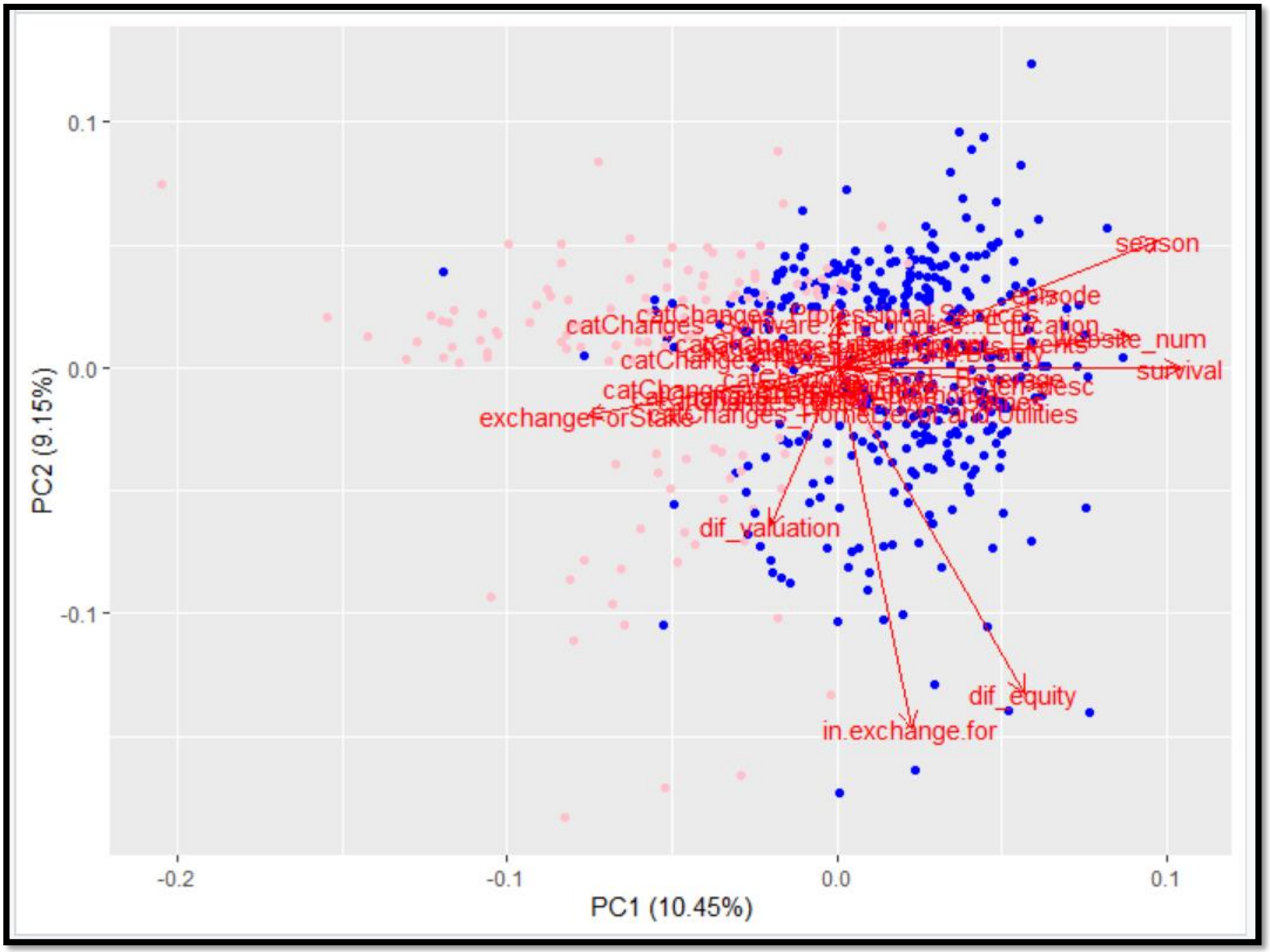
Graph 2.1 : PCA with numerical variable

Appendices



Graph 2.2 : PCA with numerical variables and the variable region dummified

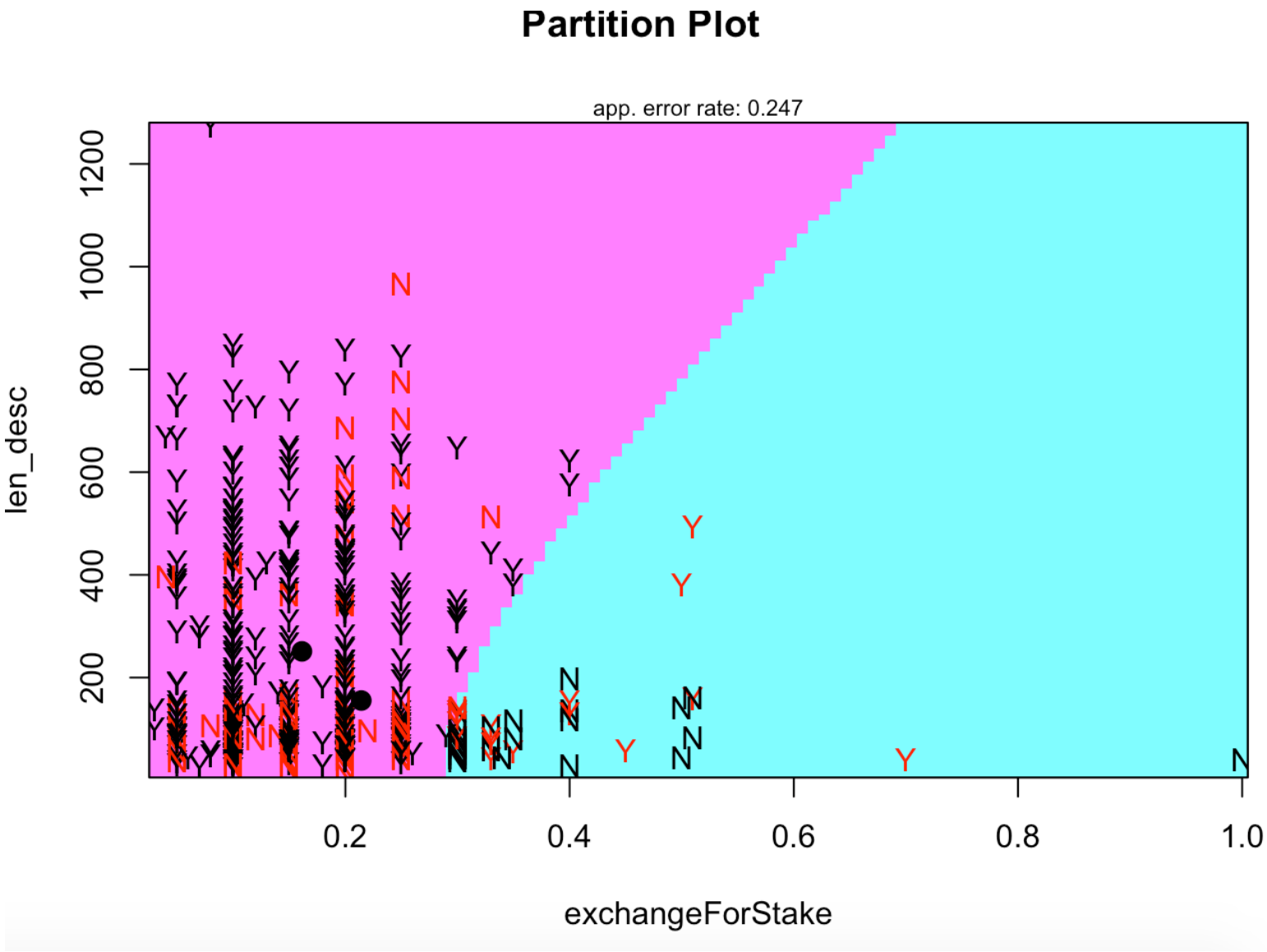
Appendices



Graph 2.3 : PCA with numerical variables and the variable category dummified

Appendices

Linear Discriminant Analysis - Quadratic



Appendices

Random forest results

predictor	Gini coefficient
shark2	0
catChanges_Pet Products	0
catChanges_Entertainment & Events	0.000431568
regions_West	0.000693068
shark3	0.000765053
shark4	0.000799598
catChanges_Software, Electronics & Education	0.000939046
deal	0.001053944
catChanges_Novelties	0.001295889
shark5	0.001401312
regions_Midwest	0.002315026
regions_South	0.002965456
catChanges_Food & Beverage	0.003953174
catChanges_Fashion	0.004865526
catChanges_Baby, Toys & Games	0.005249147
regions_Northeast	0.005727566
catChanges_Professional Services	0.005846455
catChanges_Sports & Fitness	0.006049204
catChanges_Health and Beauty	0.006070471
gender_num	0.006132109
in exchange for	0.006208224
gender	0.006309586
Multiple Entreprenuers	0.007438205
catChanges_Automotive	0.007936709
season	0.007939513

predictor	Gini coefficient
catChanges_HomeDecor and Utilities	0.009711787
shark1	0.010271883
dif_valuation	0.020627739
valuation	0.028807762
implied valuation	0.029521964
state	0.029829056
episode	0.036617975
exchangeForStake	0.03916103
location	0.039956272
entrepreneurs	0.040821454
category	0.054830755
description	0.055760118
title	0.057363888
len_desc	0.06221787
dif_equity	0.078182056
website_num	0.087796513
year	0.107581859
website	0.118554168

Appendices

Multiple Logistic regression

Multiple Logistic Regression Results

	Dependent variable:
	survival
valuation	-0.00000**
	(0.00000)
exchangeForStake	-4.659***
	(1.017)
genderM	-0.020
	(0.184)
genderX	-0.058
	(0.287)
len_desc	0.003***
	(0.0005)
catChangesBaby, Toys	2.305***
	(0.431)
catChangesEntertainment	1.895***
	(0.437)
catChangesFashion	1.624***
	(0.406)
catChangesFood	2.215***
	(0.418)
catChangesHealth and Beauty	2.057***
	(0.496)
catChangesHomeDecor and Utilities	1.189***
	(0.413)

Multiple Logistic Regression Results

	Dependent variable:
	survival
catChangesNovelties	1.851***
	(0.471)
catChangesPet Products	4.770***
	(1.117)
catChangesProfessional Services	1.061**
	(0.418)
catChangesSoftware, Electronics	1.175*
	(0.460)
catChangesSports	1.361*
	(0.600)
Multiple.EntrepreneursYes	0.634***
	(0.183)
websiteYes	3.536***
	(0.461)
regionregions_Northeast	1.460***
	(0.302)
regionregions_South	-0.119
	(0.247)
regionregions_West	-0.247
	(0.226)
Constant	-4.885***
	(0.673)
Observations	1,220
Log Likelihood	-612.048
Akaike Inf. Crit.	1,268.096
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendices

Logistic regression with valuation

Logistic Regression Results

	Dependent variable:
	survival
valuation	-0.000
	(0.00000)
Constant	-0.129*
	(0.071)
Observations	1,220
Log Likelihood	-841.831
Akaike Inf. Crit.	1,687.661
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendices

Random forest code

Code

```
# -*- coding: utf-8 -*-
"""
Created on Tue Dec  3 16:05:25 2019

@author: bhavv
"""

import pandas as pd
shark = pd.read_csv(r"C:\Users\bhavv\OneDrive\Desktop\MMA2\Multivariate Statcl Analysis- Juan Serpa\FINAL PROJECT-SHARK TANK\SharkTank_Dataset_new-catChanges.csv"
shark_df=pd.get_dummies(shark, columns = ["catChanges"])
shark_df.to_csv(r"C:\Users\bhavv\OneDrive\Desktop\MMA2\Multivariate Statcl Analysis- Juan Serpa\FINAL PROJECT-SHARK TANK\SharkTank_Dataset_new-catChanges-df.csv")

shark_statedummy =pd.get_dummies(shark, columns = ["regions"])

shark_statedummy.to_csv(r"C:\Users\bhavv\OneDrive\Desktop\MMA2\Multivariate Statcl Analysis- Juan Serpa\FINAL PROJECT-SHARK TANK\SharkTank_Dataset_new-catChanges-
statedummy.csv")

#########RANDOM FOREST FOR FEATURE SELECTION#########

shark_dummy = pd.get_dummies(shark, columns = ["catChanges","regions"])

from sklearn.preprocessing import LabelEncoder

for column in shark_dummy.columns:
    if shark_dummy[column].dtype == type(object):
        le = LabelEncoder()
        shark_dummy[column] = le.fit_transform(shark_dummy[column].astype(str))

shark_dummy_nw = shark_dummy.loc[:,shark_dummy.columns != "still open"]
#split data into X and y
X = shark_dummy_nw
y = shark_dummy["still open"]


from sklearn.ensemble import RandomForestRegressor
randomforest = RandomForestRegressor(random_state=0)

model = randomforest.fit(X,y)

from sklearn.feature_selection import SelectFromModel
sfm = SelectFromModel(model,threshold=0.05)
sfm.fit(X,y)
for feature_list_index in sfm.get_support(indices=True):
    print(X.columns[feature_list_index])


result_RF = pd.DataFrame(list(zip(X.columns,model.feature_importances_)), columns = ['predictor','Gini coefficient'])
```