To fail or not to fail?

Startup treasures detection based on historical data

APPS UCU Linear Algebra Final Project, Spring 2019

Sofiya Hevorhyan
Iryna Popovych

INTRO

When choosing a topic for this project, we wanted to find something that would be interesting for us to do and valuable in terms of business perspective.

After changing different datasets (couple of times) that were not very informative, and inspired by our classmates' stories, we decided to choose startups as our topic.

BUSINESS OBJECTIVE

Investment strategies for startup companies are usually based just on intuition or past experience. The question we pose here is,

Can we perform some analysis that can be used to identify relevant factors and score prospective startups based on their potential to be successful?

DATA

We have a dataset from CrowdAnalytix that represents startups and covers information about various aspects of company, cofounders, investments, industry, activities of company, details about employees and technologies used.

Here is the link to train and test to datasets with information about 314 startups each having 50 characteristics.

You can take a look at data dictionary for descriptions of variables.

WORK DONE

Data Exploration

- Data retyping & splitting
- Graphics

Variable Selection

- Information Value
- Random Forest
- PCA

Specialized methods

- LDA
- K-Nearest Neighbor
- Support Vector Machine

Results Evaluation

- Regressions important features
- Confusion Matrices
- Comparison Table

1. Data Exploration

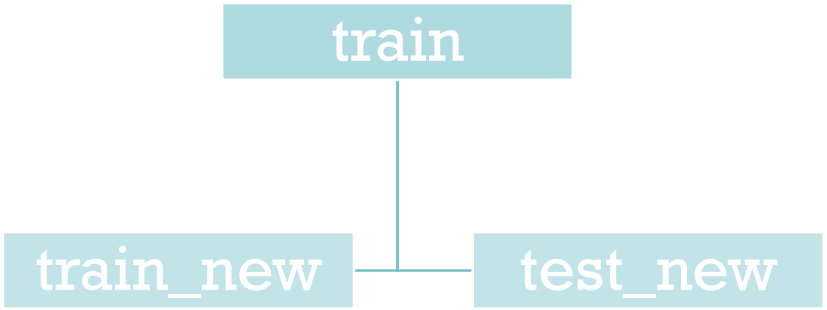
- Data retyping and splitting
- Graphics

- First we transformed characters to factors and round the numeric data types where possible.

For example, the column Founders_previous_company_employee_count transforms from
to

Large	3
Small	1
Medium	2
Large	3
…	…

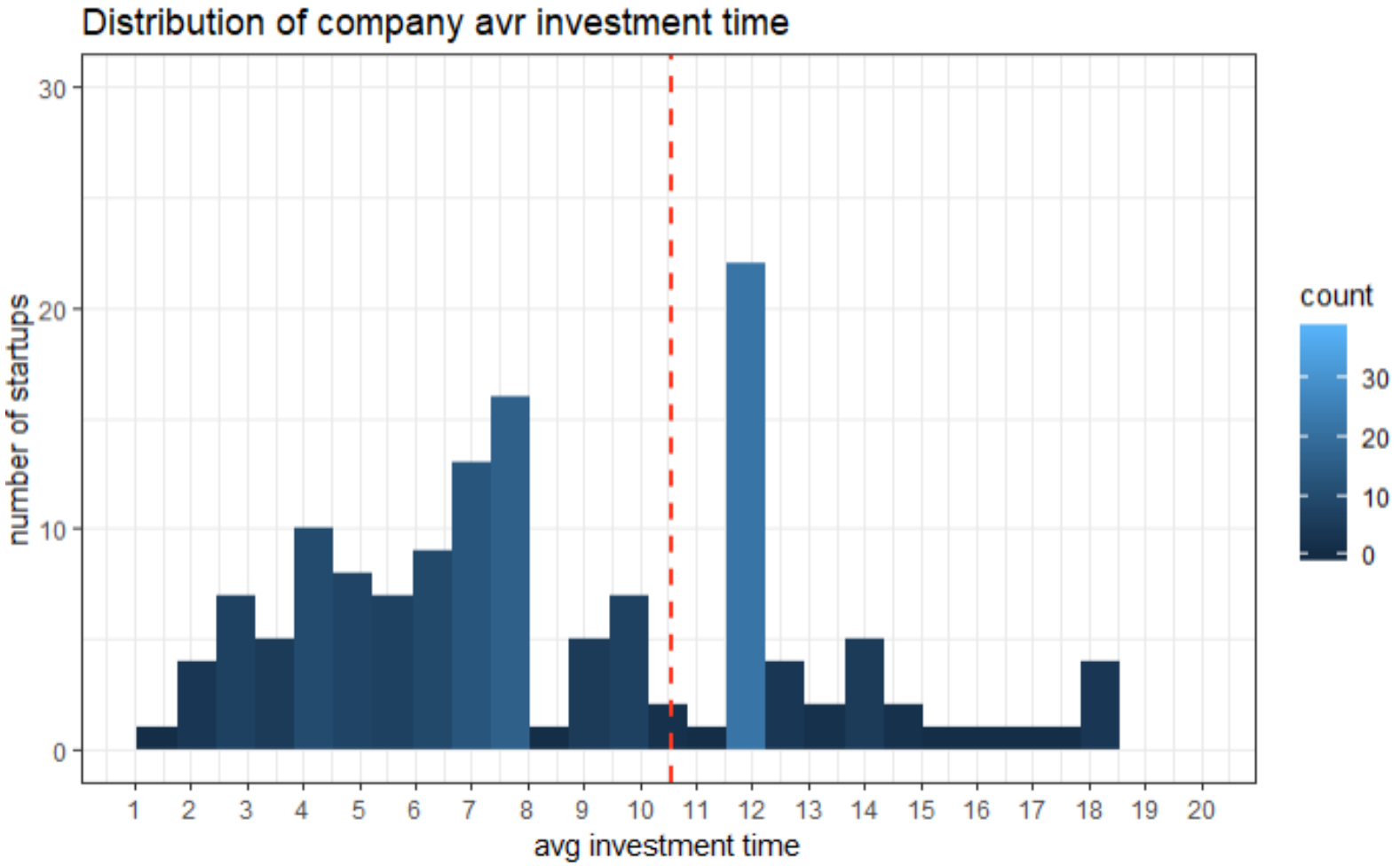
- We split data in test set in ratio 90/10 to train our models and use it for methods and then evaluate by new observations in test



Data retyping and splitting

Graphics

- We played with visuals to understand the data better. Here are some of them.

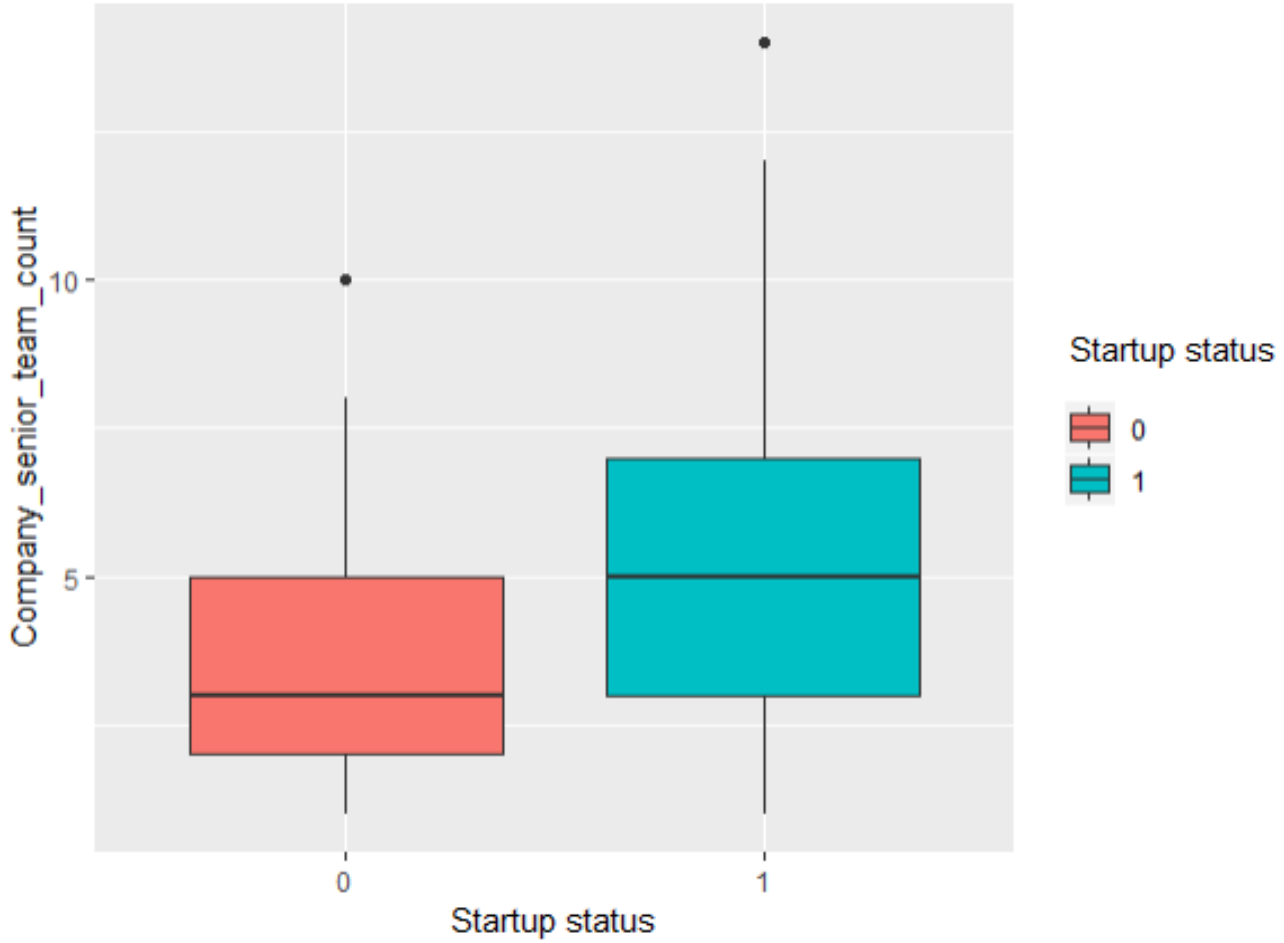


Data retyping and splitting

Graphics

- We played with visuals to understand the data better. Here are some of them.

Company status by number of top management employees

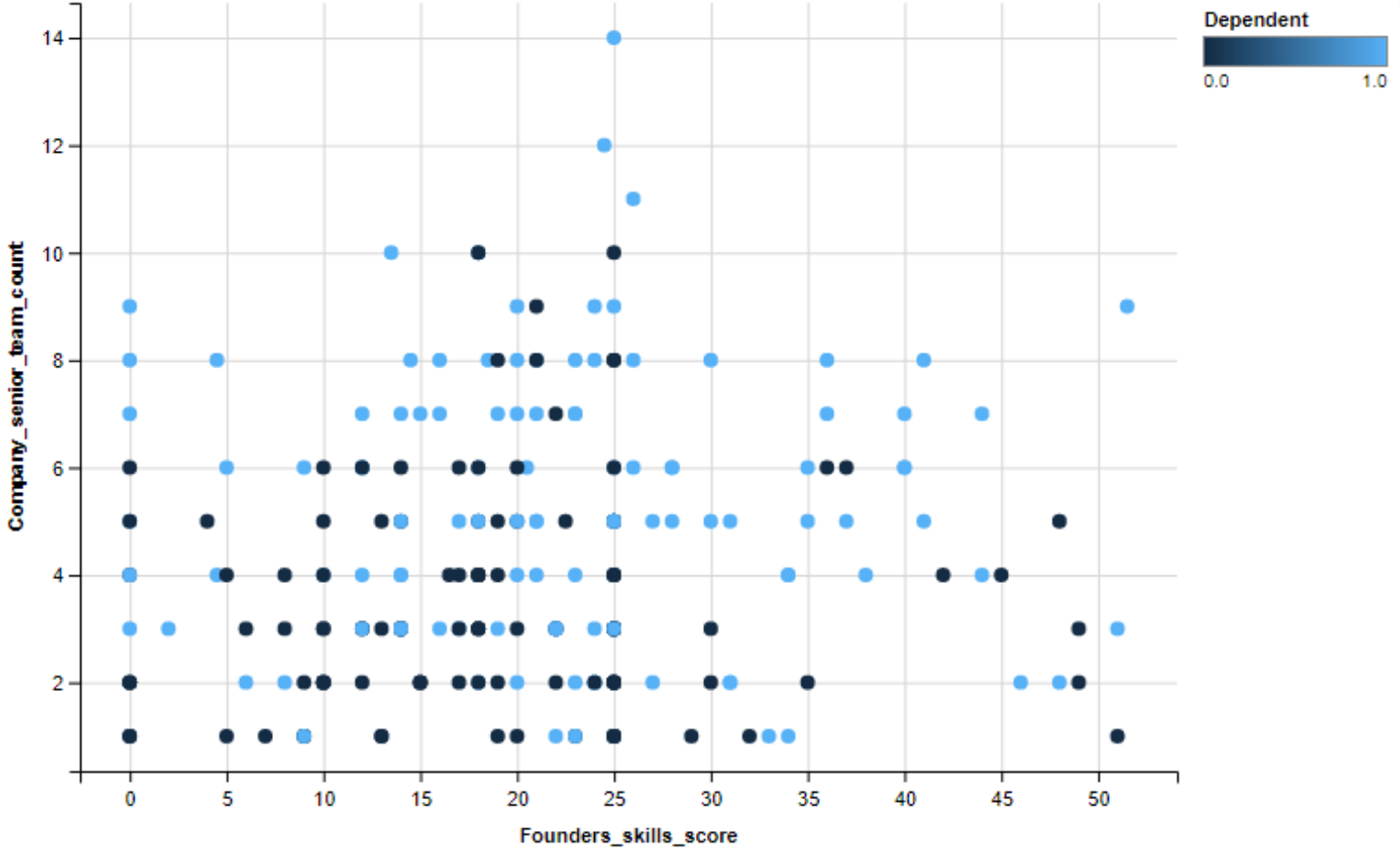


Data retyping and splitting

Graphics

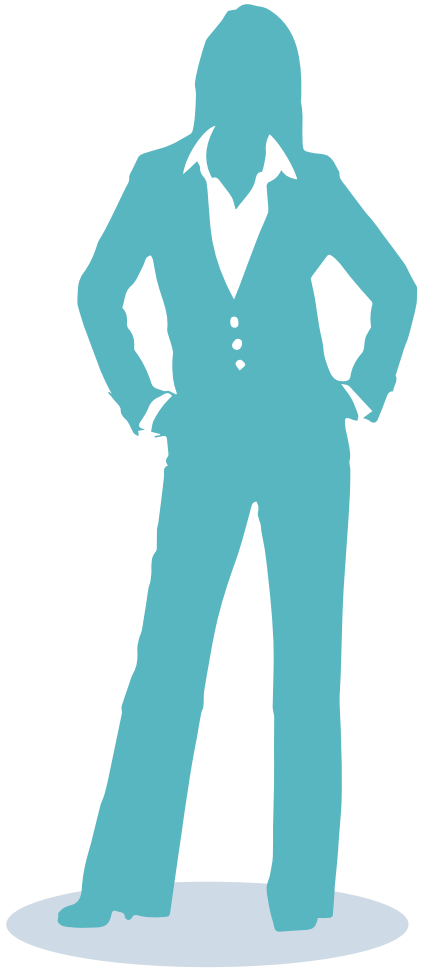
- We played with visuals to understand the data better. Here are some of them.

Company status by Founders skills score and senior team count



Data retyping and splitting

Graphics



2. Variable Selection

- Information Value
- Random Forest
- PCA

- We needed to select features for modeling, because we aren't able to use all of them.
- One of the approaches that we used for variable selection is Information Value.
  - Information Value (IV) for logistic regression is analogous to correlation for linear regression.
  - Information value tells us how well an independent variable is able to distinguish two categories of dependent variables.
  - We selected variables with IV of 0.1 to 0.5 for modeling which are considered to be strong predictors

```
# selecting variables with good information values
var1 <- IV[which(IV$InformationValue > 0.1),]
var1 <- c(var1, var1[which(var1$InformationValue < 0.5),])
x_train <- train_new[var1$Variable]
```

Information Value

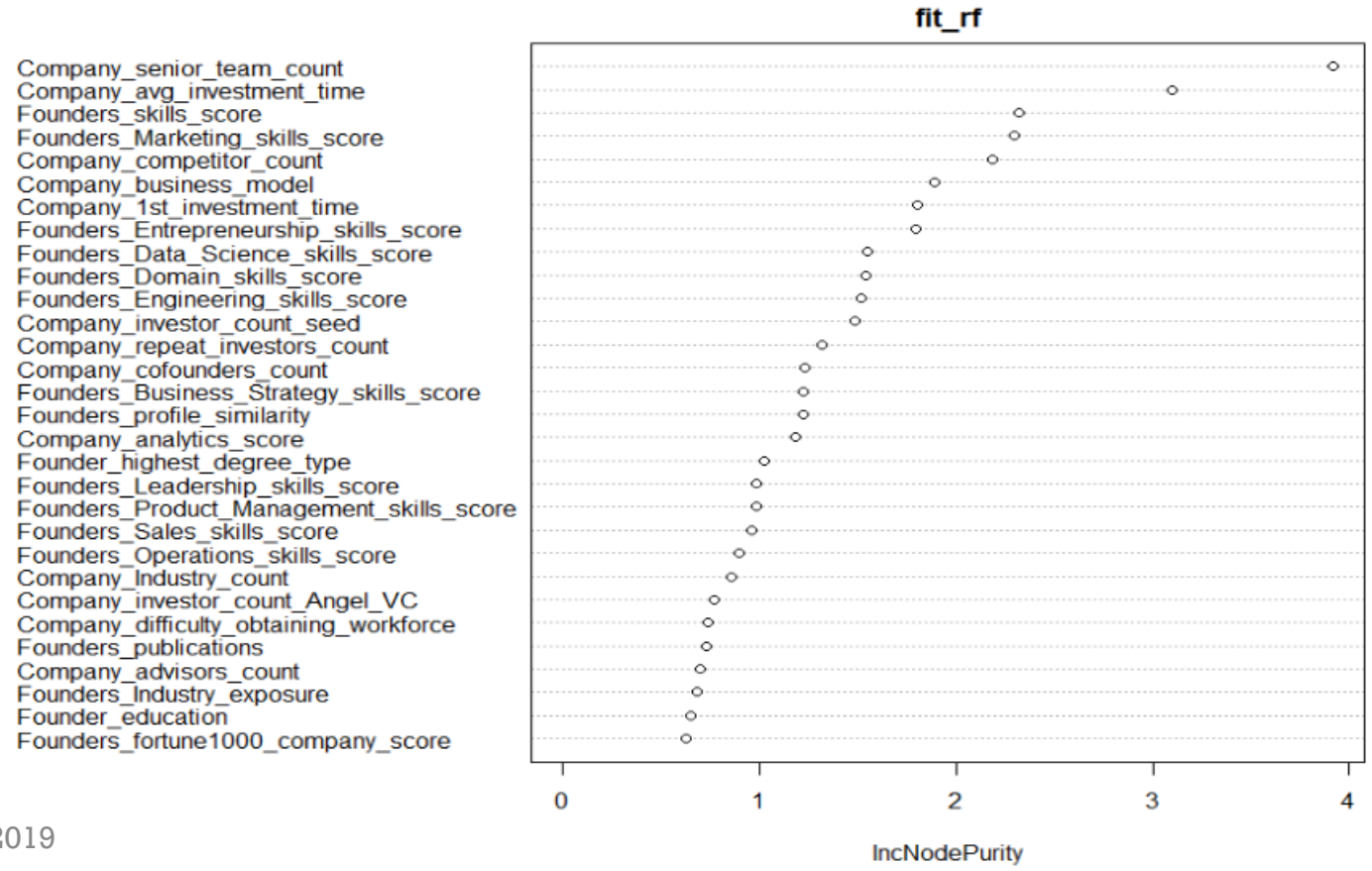
Random Forest

PCA

- Another way of determining importance is random forest that helps to select the most relevant features
- Perform random forest on some train data and the algorithm will come up with some set of rules that would be applied on your test data to actually predict your dependent variable

Advantages of random forest:

- They can deal with messy data. No problem with lots of predictors
- It automatically does a good job of finding interactions as well.



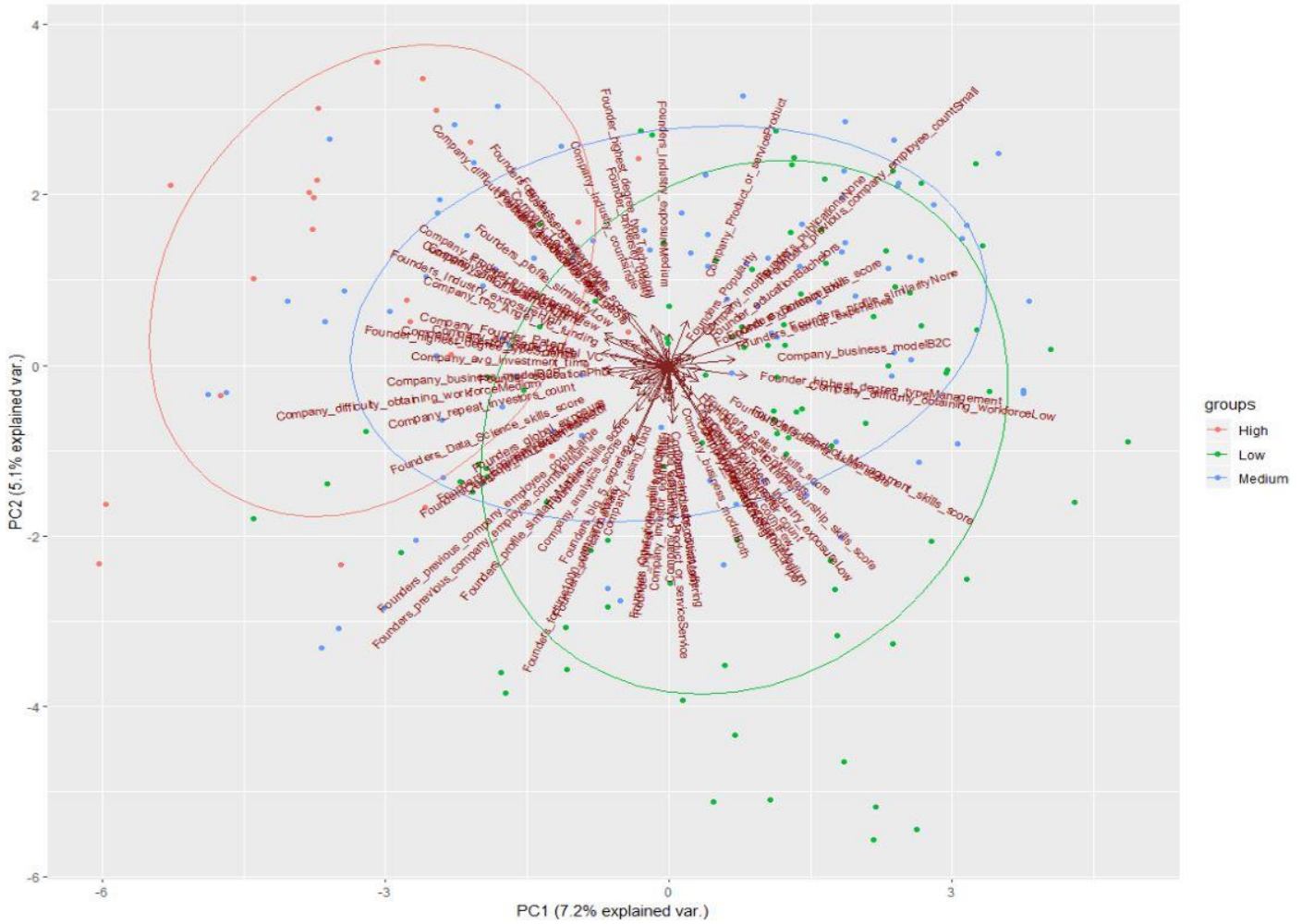
Information Value

Random Forest

PCA

In our case it was very useful to try PCA as a feature selection technique for modelling.
PCA projects the entire dataset onto a different feature subspace.

In the figure below you can see what we got after calculating the rotation matrix.
It is clear that first principal component explains 7.3% of variance
and second component explains 5.1% variance.

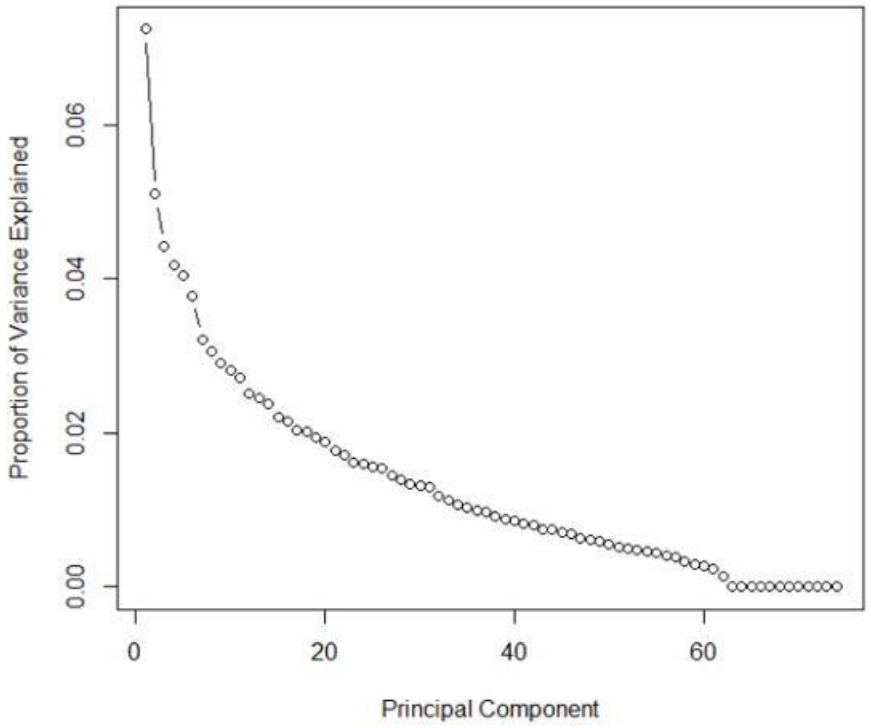


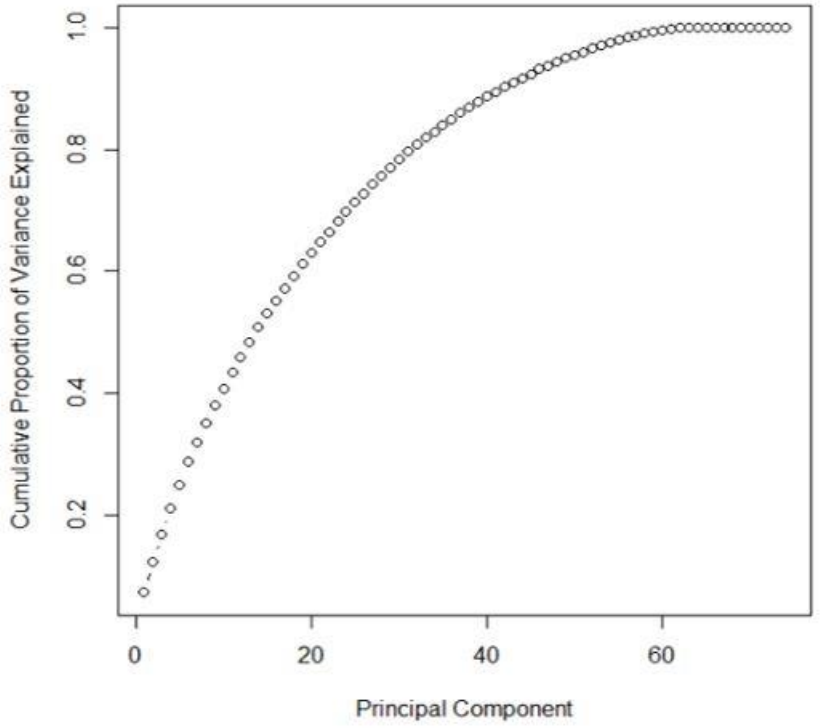
Information Value

Random Forest

PCA

After we know the principal components, we need to decide how many of them to choose for further modelling. We used 60 components of 75 available, because more than 98%percent of data were described by them. You can clearly see this from a scree plot (line plot of the eigenvalues of principal components) and cumulative scree plot.

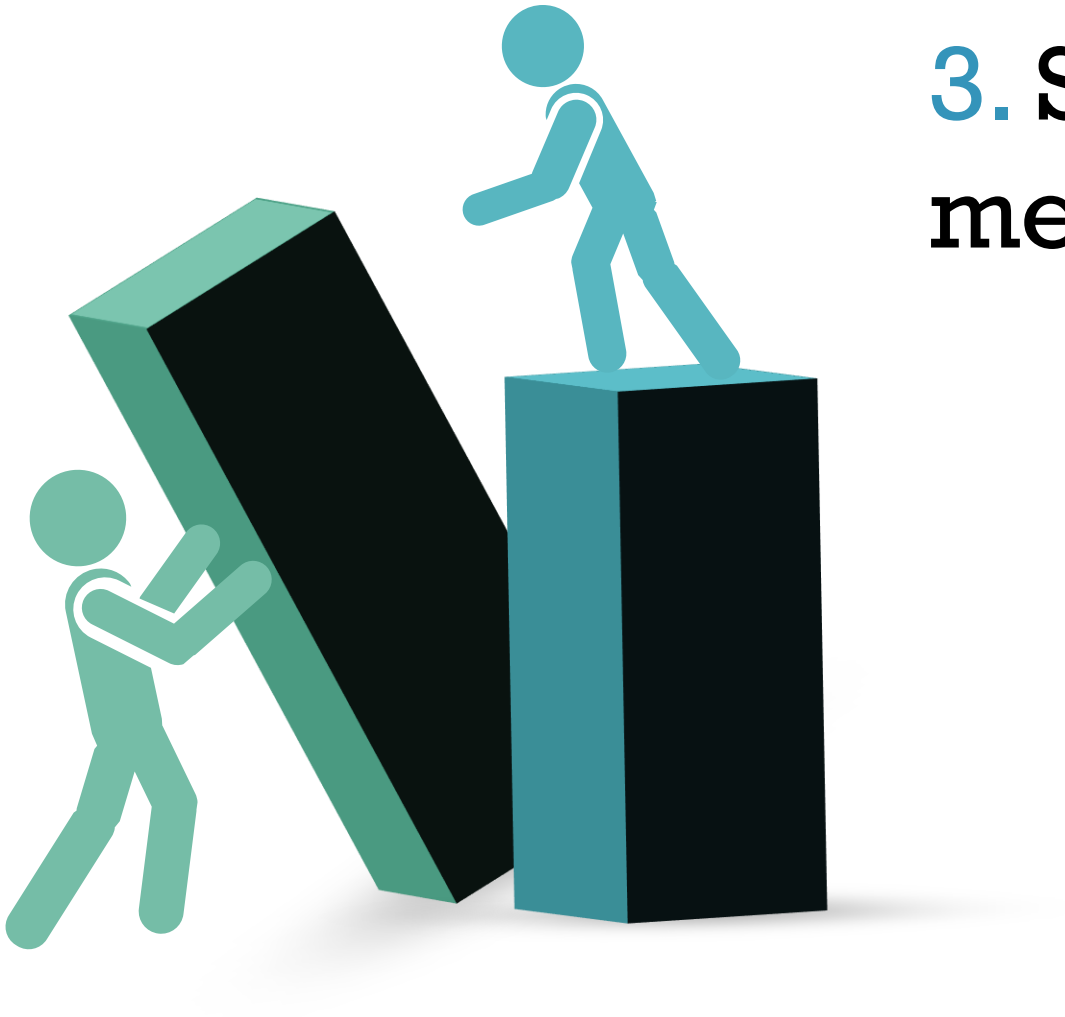




Information Value

Random Forest

PCA



3. Specialized methods

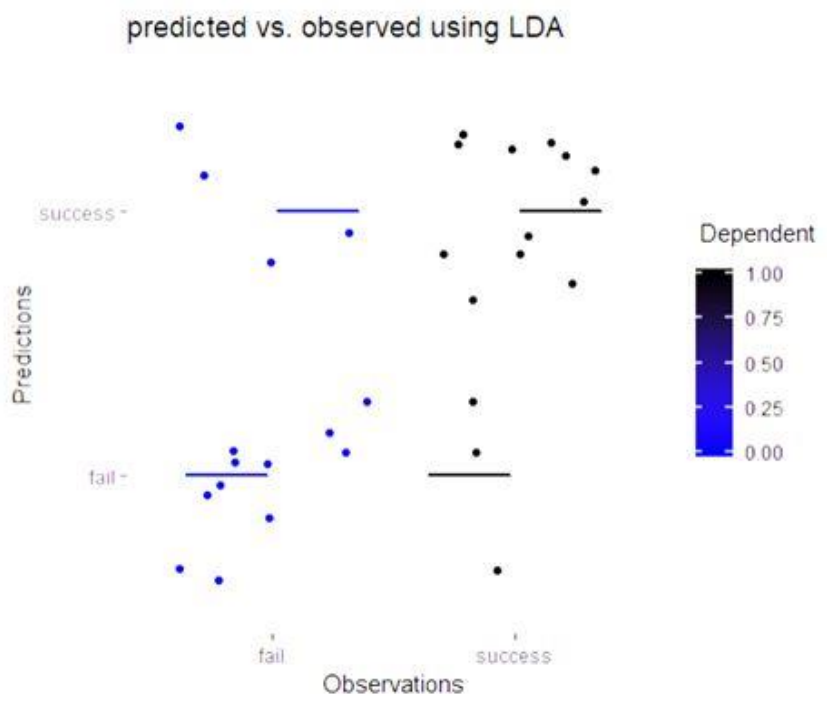
- LDA
- KNN
- SVM

- LDA which stands for Linear Discriminant Analysis is next most common method for feature selection after PCA

- Taking training set, LDA tries to project it onto the same line to make distances between different classes as far as possible while distances between the same category the smallest

```
          Reference
Prediction fail success
   fail      11       3
   success    4      12

               Accuracy : 0.7667
                 95% CI : (0.5772, 0.9007)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.002611
```



LDA

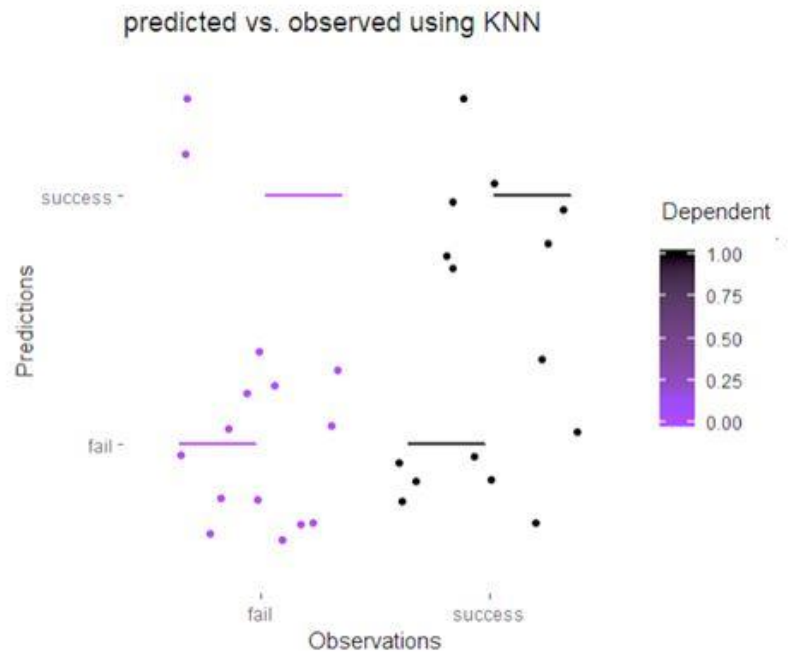
KNN

SVM

- KNN or K-Nearest Neighbor is the technique to classify data point to the given category
- For the new observation x, prediction is made by searching on the whole data set for the k-nearest neighbors (most similar cases) and summarizing the output variable for these neighbors
    - Determine the Manhattan or Euclidean distance
    - Determine the number of k neighbors for your data

```
          Reference
Prediction fail success
   fail      13       8
   success    2       7

               Accuracy : 0.6667
                 95% CI : (0.4719, 0.8271)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.04937
```



LDA

KNN

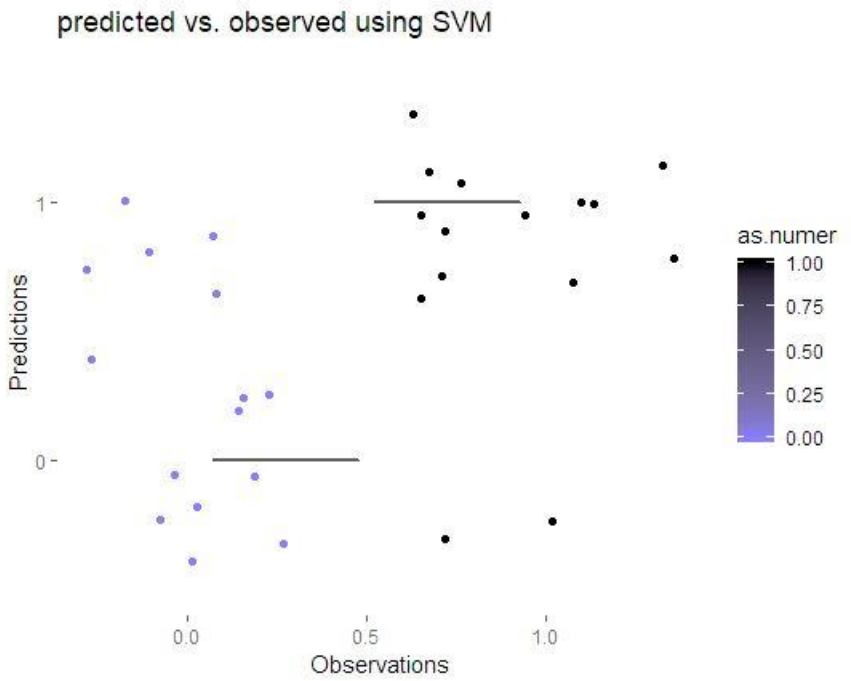
SVM

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N in our case is the number of features) that distinctly classifies the data points. SVM tries to maximize the margin between the closest support vectors by minimizing hinge loss.

Here we present the results of the PCA model based on Principal components derived from PCA. Accuracy of prediction is 76.67%, based on the test data from 30 observations.

```
          Reference
Prediction  0  1
         0 10  5
         1  2 13

               Accuracy : 0.7667
                 95% CI : (0.5772, 0.9007)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.04352
```



LDA

KNN

SVM

4. Results

- Important Features
- Comparison Table
- CrowdAnalytix

- We used different approaches to select predictors for various models for our project.
    - variable importance
    - randomForest + caret, varImp()
- After that, we've built different models with these groups of features, including one mixed group, and compared results.

```
Call:
glm(formula = Dependent ~ Company_senior_team_count +
Company_business_modelB2C +
    Company_competitor_count + Company_analytics_score, family =
binomial(link = logit),
    data = train.iv)
```

```
Call:
glm(formula = Dependent ~ Company_senior_team_count + Founders_Popularity +
    Founders_publicationsFew + Founders_Data_Science_skills_score +
    Company_crowdfunding + Founders_Engineering_skills_score +
    Company_competitor_count + Company_1st_investment_time, family =
binomial(link = logit),
    data = train.varimp)
```

Important Features

Comparison Table

CrowdAnalytix

Prediction accuracy for different methods used

		logit from varImp (randomForest)	logit from IV	KNN	LDA	IV+SVM	PCA+SVM
Prediction accuracy	Test sample 30 obs.	73.33%	76.67%	66.67%	76.67%	73.33%	76.67%
	Test sample 80 obs.	**70.46%**	56.82%	**78.28%**	**70.70%**	65.54%	68.93%

Important Features
Comparison Table
CrowdAnalytix

Online Data Challenge

Top 10 on Leaderboard

Position	Name	Score	Last Submission
1	Pranjal Rawat	0.84343	16/12/2018 22:3:56
2	Marco Scattolin	0.81818	20/1/2017 17:6:3
3	Abhishek Sharma	0.79293	2/4/2018 16:21:46
4	maxio	0.78914	12/4/2016 21:34:1
5	Fernando Luziani Gomes	0.78788	2/6/2016 13:37:33
6	Kranthi	0.78535	9/3/2016 7:51:39
7	Arjun K Vijayakumar	0.78535	6/2/2016 14:54:4
8	crazyBA	0.78283	15/5/2019 12:13:16
9	frankrizzo	0.77904	10/4/2017 17:56:29
10	Mark Landry	0.77020	10/12/2016 12:4:54

Important Features

Comparison Table

CrowdAnalytix

CONCLUSIONS

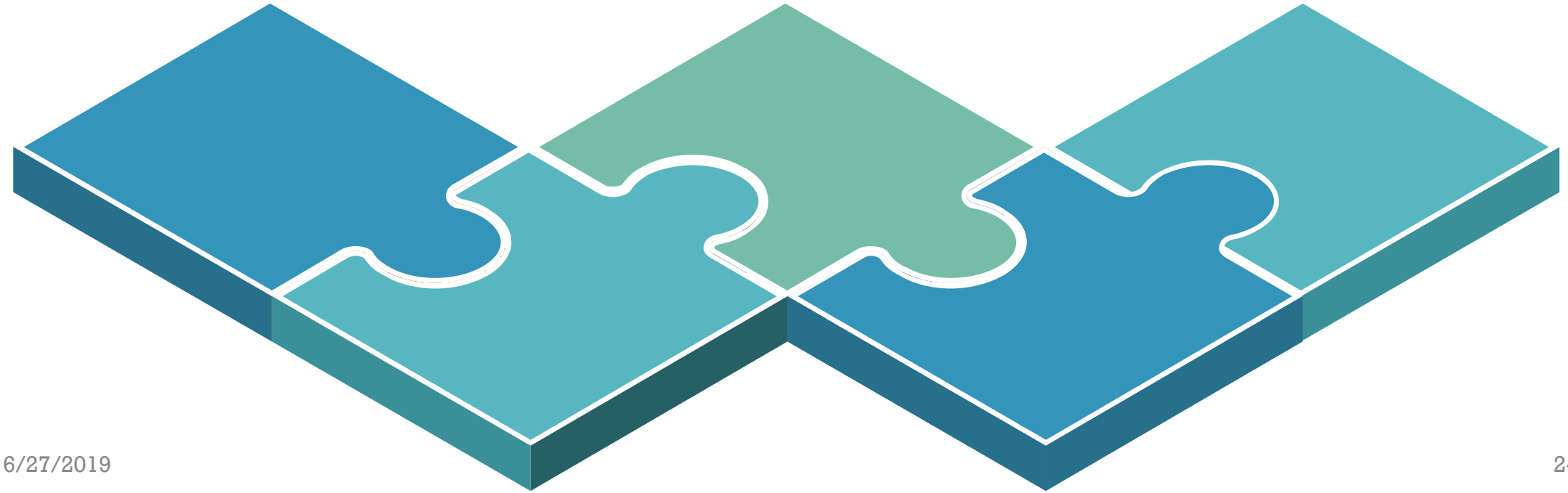
A Data exploration and preparation is a very significant part that influence your results

B Now understand different methods that can be used for prediction such as SVM, PCA, LDA and KNN

C We learned some new techniques for working with wide datasets.

D Feature selection was one of the most important part of our work. We used different methods, and tried to compare them to choose best.

E We have used this methods mostly separately but lots of them can be performed in tandem (LDA and PCA) that can be done in the future.



AUTHORS

Iryna Popovych

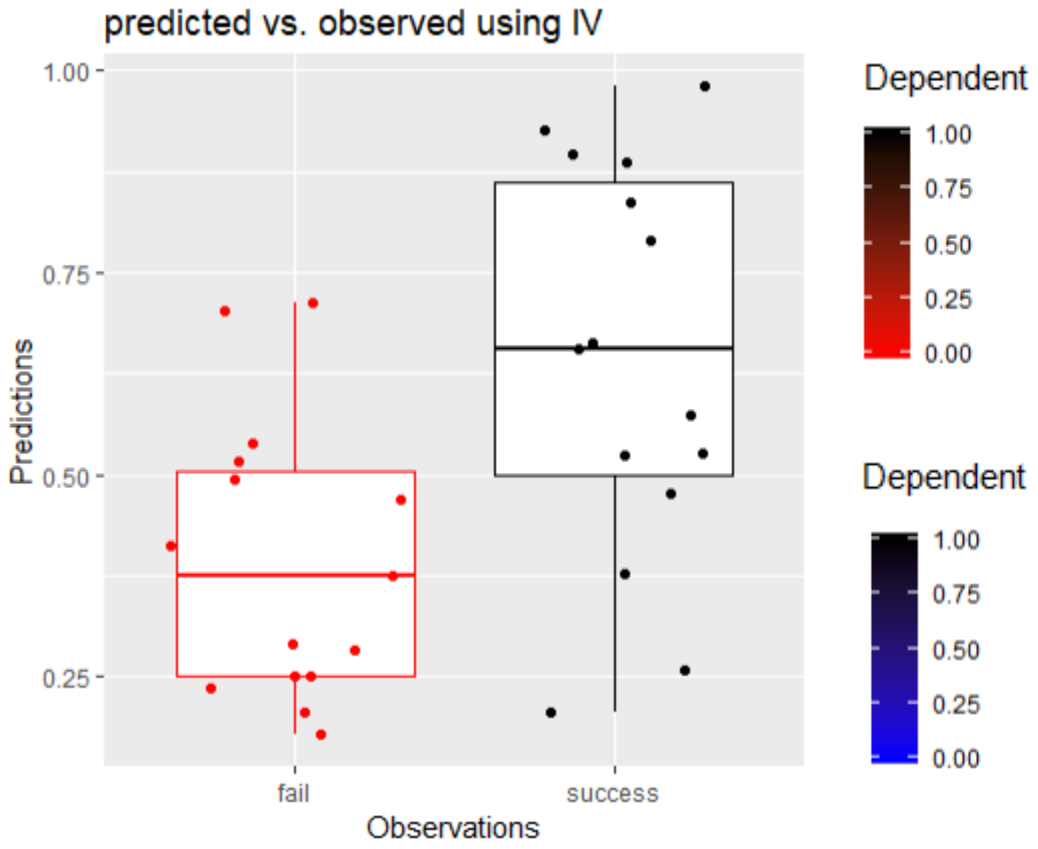
Sofiya Hevorhyan

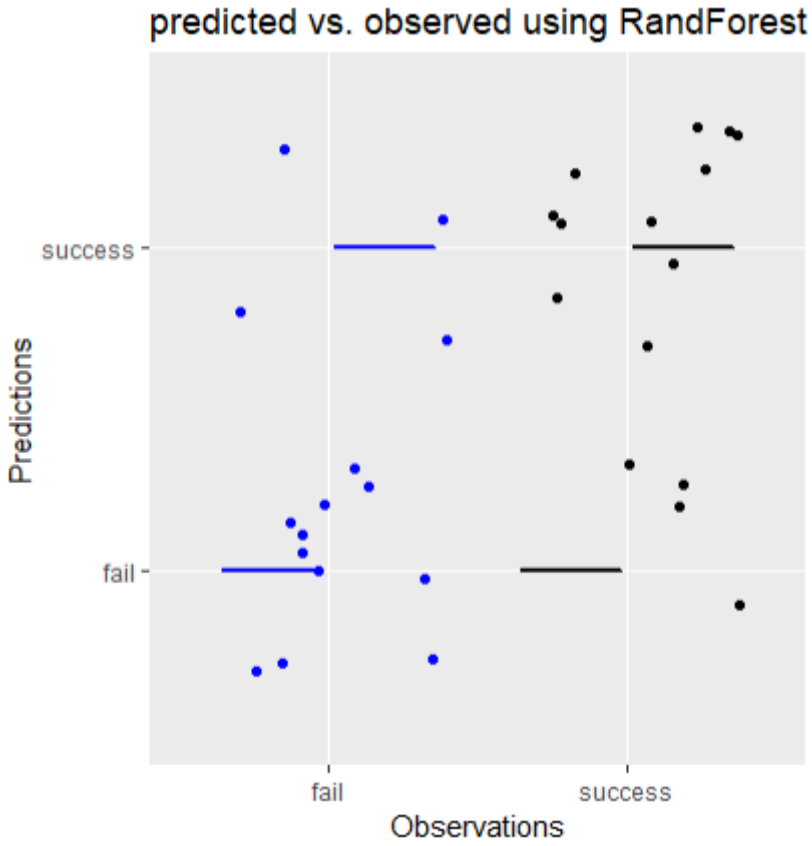
LINKS

/DataAnalysisR

link dataset

THANK YOU





- Regression
- Confusion Matrix
- CrowdAnalytix