Investment Business Analytics: Strategy Consulting (Mobile)

|  |
| --- |
| A Project Report Presented to Chandrasekar Vuppalapati |
| San Jose State University CMPE 277  Software Application Development – Data Analytics |

|  |
| --- |
| By |
| Lakshmi Venkateswaran  (October 2016) |
|  |

|  |
| --- |
| Copyright © 2016 |
|  |
| ALL RIGHTS RESERVED |

**Table of Contents**

Contents

[Chapter 1. Project Overview 4](#_Toc464938752)

[1.1 Introduction/Description 4](#_Toc464938753)

[Chapter 2. Requirements 5](#_Toc464938754)

[Functional Requirements 5](#_Toc464938755)

[Non-Functional Requirements 5](#_Toc464938756)

[Statistic Analytical Layer 6](#_Toc464938757)

[Middle Tier:  The Spring Framework with Maven build 11](#_Toc464938758)

[Entity Classes 13](#_Toc464938759)

[Repositories 14](#_Toc464938760)

[Controller 15](#_Toc464938761)

# Project Overview

## Introduction/Description

The currently trending topic in the past decade with the growth in the cyber arena is the analysis of big data. The availability of smart devices is catalyzing this to a different level too.

The various dimensions of data that are developed based on the figured patterns using numerous algorithmic techniques are key to this drive. The market trends, financial businesses, healthcare industry, people’s choice of products, their likes and dislikes these data can be aggregated by the data scientist to discover. Therefore, a fruitful outcome can be expected if the exploitation of this data is done with the right strategy, scientific tools to benefit the various day to day users of analytical and logical information.

More than ever before, today big organizations are looking for innovation as a key driver to sustain in the current milieu of scientific growth and development in the global arena. Dedicated well learned individuals are appointed to establish high-level innovation programs and ideas which in turn has fueled a growth of innumerous number of startups around the world.

And in the current economic times, investment is the one that allows this industry to sustain. The investors also in turn expect a commercially beneficial outcome of this innovation driven investment program.

It is learnt that the investment strategies are predominantly based on intuition, coffee shop discussions, trending media hypes or past experiences. The result of which is that the investors feel the need to be addressed with a logical and analytical approach. Now the challenge posed in this project is could we come up with a rigorous analytical solution that can be used to identify the major factors impacting the success or the failure of the startups and giving the information on the risk level of the investment.

The model that will be developed as a part of this project will be the tool allowing the investors to make a well-informed decision and rely less on their random intuitions. This project will need to focus on data groundwork and examination. Before any modeling algorithm could be suitable the given data needs to be prepared, cleansed and explored to make the data fit for analysis. The features are finally selected to prepare the model that will be the base engine to support the complete mobile application.

.

# Requirements

## Functional Requirements

**R1.1** The system shall allow the legit users to log in.

**R1.2** The system shall allow existing legit users to log in.

**R1.3** The system shall allow users to register.

**R1.4** The system shall list all the startups enrolled

**R1.5** The system shall display the risk measurement on the company.

**R1.6** The system shall provide complete details of each company.

## Non-Functional Requirements

**R2.1 Performance**

R2.1.1 The system should be able to accommodate considerable amount of incoming request.

R2.1.2 The system’s response time while loading the application should be less than 10 seconds.

R2.1.3 The response time in loading a screen should be less than 10 seconds.

**R2.2 Scalability**

R2.2.1 The turnaround time of the system on the database breakdown should not be more than 10 minutes.

R2.2.2. Considering the immense reliance on the database calls the system should be able to vertically and horizontally scale with the growth of number of users as well as the store of the incoming data.

**R2.3 Availability**

R2.3.1 The app, specifically the calls to the remote database, must be made available for use always.

**R2.4 Security**

R2.4.1 An authentication and authorization system is kept in place with user/password login system.

1. **High Level Architecture Design**

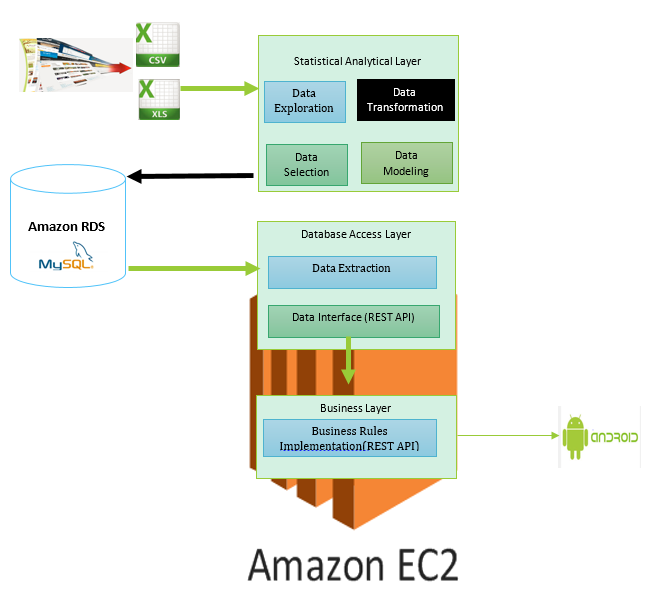


Figure 1: Architecture Diagram

## Statistic Analytical Layer

**Data Exploration:** This module includes whole exploration and analysis of each variable listed in the CSV files. It also makes the validation of all the variables which are then subjected to the next phase.

**Data Transformation:** The contribution to this layer is the data gathered from the data collection component. This component assemblies the data into list of CSV files. This information is fed into the data exploration layer.

**Data Selection**: This component forms the primary aspect of the layer composing the selection of the prime variables required for further analysis.

**Data Modelling:** This component forms the engines core that provides the prediction model on the basic system.

## 

**Database Access Layer**

**Data Extraction:** This is the part of the layer that extracts data from the data base in the raw format of storage for the next layer.

**Data Interface:** This is the layer that exposes the data in the API Restful format for business consumption.

**Business Layer**

**Business Rules Implementation (RestAPI)**: This layer is mainly used for the project to apply business rules demanded by the application of consumption.

**User Interface**

The user interface in this case is an Android application (Lollipop) for display.

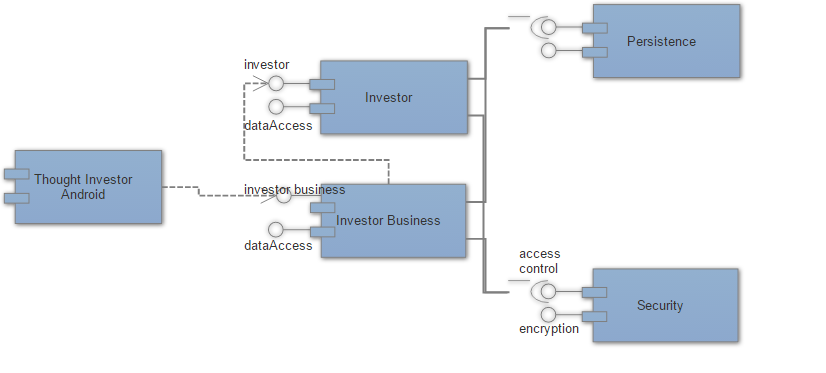


Figure 2: Component Level Design

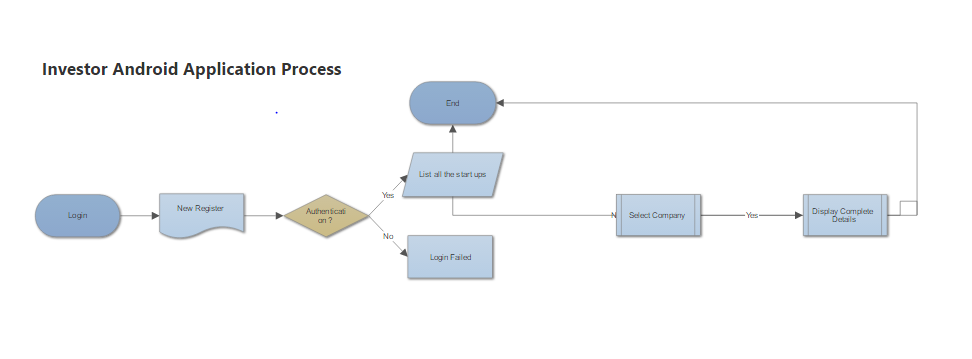
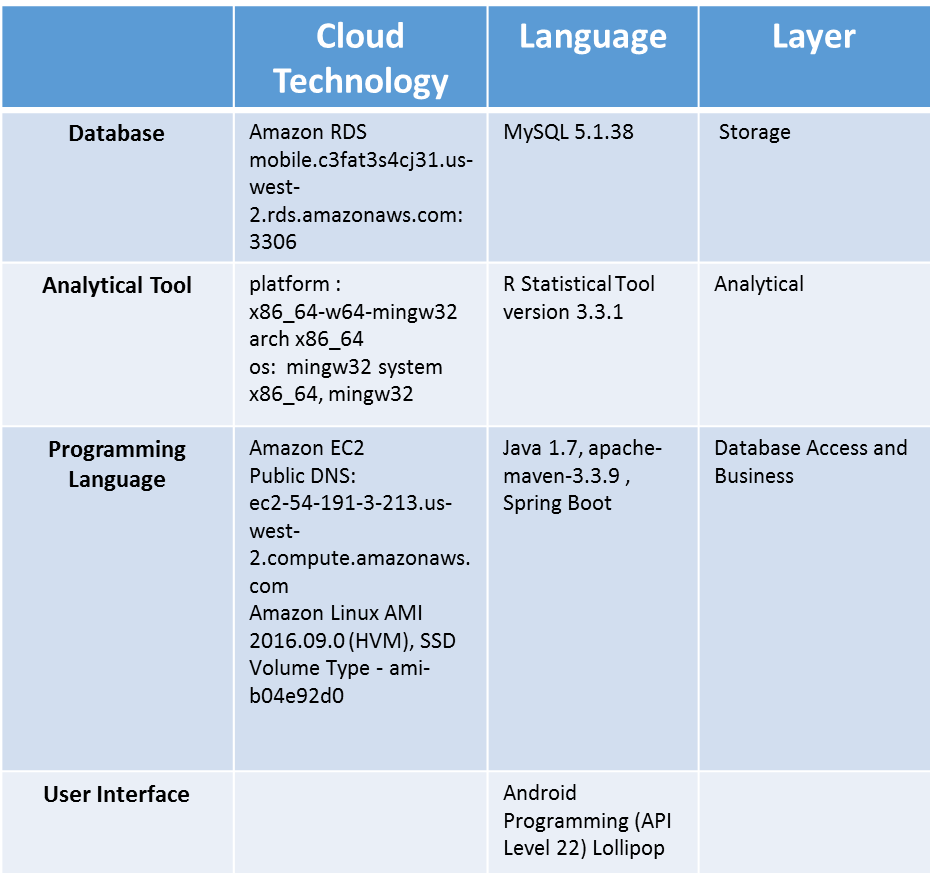


Figure 3: System Workflow Diagram

1. **Mobile & Cloud Technologies Used & Descriptions**

****

**Deployment Links:**

<http://ec2-54-191-3-213.us-west-2.compute.amazonaws.com:1111/investor/retrieve/companies>...

A current tendency in software is SaaS (Software-as-a-Service), is a software delivery model that allows the software service companies to deliver the software as on-demand service instead of connecting the request on every computer. This is an exclusive enterprise that helps the investor to use different software services that are hosted on the cloud

An Android application is used as a mobile platform given its wide usage across the world.

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale cloud computing easier for developers.

Amazon EC2’s simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon’s proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you use. Amazon EC2 provides developers the tools to build failure resilient applications and isolate themselves from common failure scenarios.

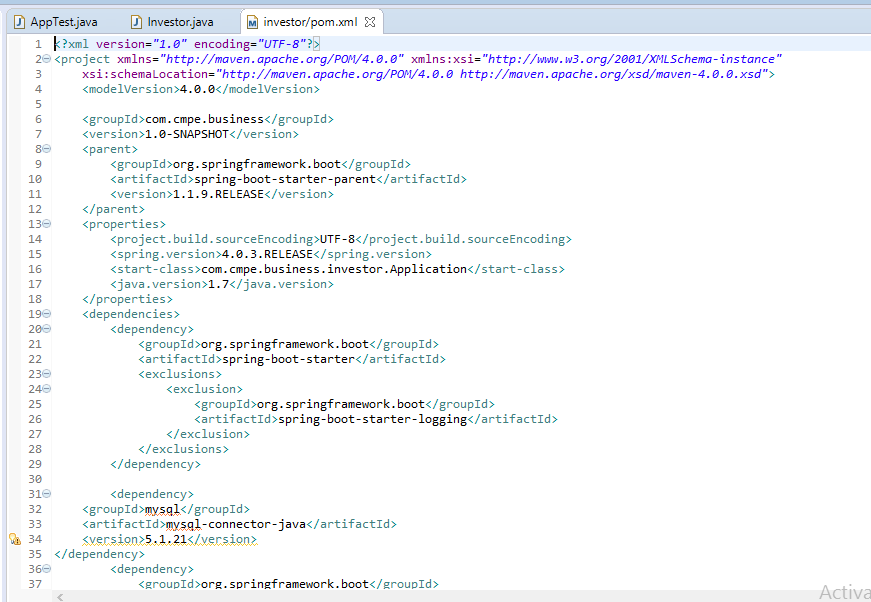
References : https://aws.amazon.com/ec2/

**Design Patterns Used – Front End, Middle Tier, Data Store and Cloud Technologies**



**Middle Tier:  The Spring Framework with Maven build**

Maven configuration file, “pom.xml”



## Entity Classes

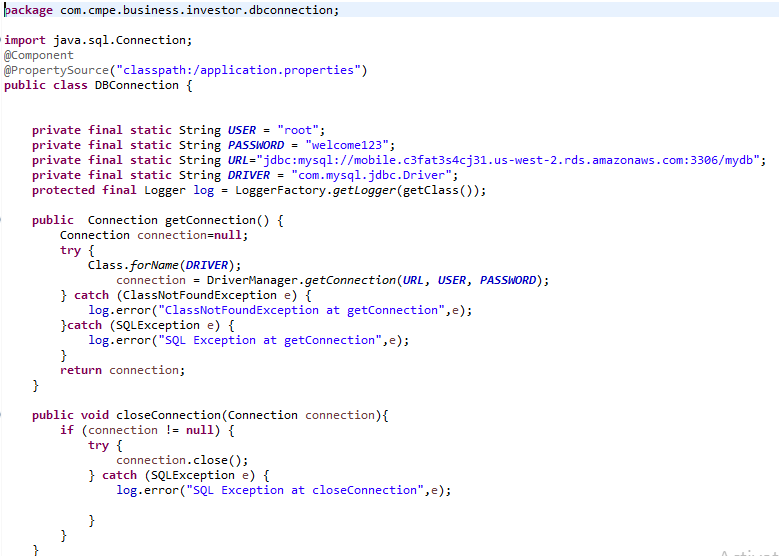
To be able to store information about investor Start-Ups an entity class called: “Investor” is developed.



### Repositories

With JPA Investor Repository that allows CRUD operation:

* “respository.listAll()”: returns all Companies List



## Controller

The controller will map request URIs to view templates and perform all necessary processing in between.



**Data Cleansing, Preparation, Exploration & Modeling Using R (Testing and Automation)**

The Analytic layer involves all the major phases of software development in the project

The following steps were followed in establishing the Analytical Layer and Testing

* Import the data in R
* Checking for anomalies
* Individual Variable Exploration
* Creating new Feature
* Treating the data missing for consistency
* Hypothesis testing and feature selection

table(char\_df$Country.of.company,useNA = "always")

Argentina Austria Azerbaijan

2 2 2

Belgium Bulgaria Canada

5 3 3

Czech Republic Denmark Estonia

1 3 1

Finland France Germany

2 8 6

India Israel Italy

10 4 1

Russian Federation Singapore Spain

1 1 5

Sweden Switzerland United Kingdom

1 2 33

United States

305

char\_df$Continent.of.company[char\_df$Continent.of.company=='Asia' ||

char\_df$Continent.of.company == 'South America'] <-'Others'table(char\_df$Continent.of.company,useNA = "always")

Asia Europe South America USA

15 76 2 308

char\_df$Continent.of.company[char\_df$Continent.of.company=='Asia'] <-'Others'

table(char\_df$Continent.of.company,useNA = "always")

Europe Others South America USA

76 15 2 308

char\_df$Continent.of.company[char\_df$Continent.of.company=='South America']

<-'Others'

table(char\_df$Continent.of.company,useNA = "always")

tab<-table(char\_df$Dependent.Company.Status,char\_df$Continent.of.company)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 32.644, df = 2, p-value = 8.156e-08

Feature Creation

char\_df$Company.industry.count<-length(strsplit(char\_df$Industry.of.company, "|",fixed=T))

for (i in (1:length(char\_df$Industry.of.company)))

{

if(is.na(char\_df$Industry.of.company[i])==T){

char\_df$Company.industry.count[i]<- NA}

else{

lst<-strsplit(char\_df$Industry.of.company[i], "|", fixed=T)

clength<-length(lst[[1]])

if(clength ==1){

char\_df$Company.industry.count[i]<-‘single’}

if(clength 1 & clength < 3)){

char\_df$Company.industry.count[i]<-‘few’}  
 if(clength 3){

char\_df$Company.industry.count[i]<-‘Many’ }

} }

char\_df$Company.industry.count

[1] NA "472" "472" "few" "472" "few"

[7] "single" "few" "472" "472" "few" NA

[13] "472" "few" "472" "few" "single" "single"

[19] NA "472" "single" NA "single" NA

[25] "few" "single" "few" NA "single" "single"

[31] "472" "single" "single" "single" NA "few"

[37] "472" "472" "472" "Many" "472" "few"

[43] NA NA NA NA "few" NA

[49] "single" NA "few" "472" "single" NA

[55] NA "single" "472" "Many" NA "few"

[61] "single" "few" "few" "472" NA NA

[67] "Many" "472" "472" "single" "472" "472"

[73] "few" "single" "few" "472" "few" "Many"

[79] "few" "single" "few" "single" "few" "472"

[85] "few" "Many" "Many" "Many" "single" "472"

[91] "single" "single" "few" "single" "single" "single"

[97] NA "single" NA "single" NA "single"

[103] "single" "single" "few" "Many" "472" "472"

[109] "single" "single" "single" NA NA "few"

[115] "few" NA "few" "single" "single" "472"

[121] "Many" "Many" "Many" "few" "Many" NA

[127] "single" NA "single" "single" "472" "single"

[133] "few" "single" "472" NA "single" "single"

[139] NA "single" NA "single" "472" "single"

[145] "single" "few" "472" NA NA "single"

[151] "single" "472" NA NA NA NA

[157] "single" NA NA NA "few" "few"

[163] "few" NA "472" "Many" "single" "Many"

[169] NA NA "472" "single" "single" NA

[175] "472" NA "472" NA "472" "few"

[181] "472" "472" NA "few" "few" "single"

[187] "few" "472" NA "single" NA "472"

[193] NA "single" "few" "single" "Many" "few"

[199] "single" "single" NA "few" NA "few"

[205] "single" "few" "single" "single" "Many" "few"

[211] "few" NA "single" "single" "472" "472"

[217] NA NA "single" "Many" NA "single"

[223] "single" NA "single" "few" "single" "single"

[229] NA "few" "few" "few" NA NA

[235] "472" NA NA "few" "Many" "single"

[241] NA "few" "few" "single" "few" "single"

[247] "single" "few" "single" "single" "few" "single"

[253] "Many" "472" "single" "few" "472" "few"

[259] "single" "few" "single" "single" "single" "few"

[265] "single" "single" NA "single" NA "few"

[271] NA "single" "few" "472" NA NA

[277] NA "single" "few" NA "few" "single"

[283] "single" "472" "few" "single" NA NA

[289] "single" NA NA "472" "472" "single"

[295] NA NA "472" "few" "Many" "few"

[301] "Many" "Many" "472" "Many" "472" "Many"

[307] "Many" "472" "472" "Many" "472" "Many"

[313] "single" "few" "single" "single" "472" "single"

[319] "few" "single" "few" "single" "Many" "few"

[325] "few" "few" "few" "Many" "single" "few"

[331] "single" "single" "few" "Many" "472" "single"

[337] "single" "few" "few" "472" "few" "few"

[343] "few" "few" "472" NA NA "few"

[349] NA "few" "single" "single" "single" "single"

[355] "few" "few" "few" "Many" "single" "472"

[361] "single" "few" "Many" "few" "Many" "few"

[367] "Many" "few" "few" "472" "few" "few"

[373] NA NA "single" NA "few" "few"

[379] NA NA "few" "Many" "472" "Many"

[385] NA "single" NA "few" "single" "single"

[391] NA "few" "single" "single" NA "single"

[397] "472" "single" "472" "few" NA NA

[403] "few" NA NA NA NA NA

[409] NA "few" "Many" NA NA NA

[415] "472" NA "single" "few" "few" "single"

[421] NA "few" NA NA "few" "few"

[427] "single" NA NA "single" NA "single"

[433] "single" NA "single" "single" "single" NA

[439] "472" NA NA NA "single" "few"

[445] NA NA NA NA "single" NA

[451] "single" NA NA "single" NA NA

[457] "single" NA "single" "few" "few" NA

[463] NA NA "single" "few" NA "single"

[469] NA "few" "few" "few"

table(char\_df$Company.industry.count,useNA = "always")

472 few Many single <NA

66 112 37 133 124

char\_df$Company.industry.count[char\_df$Company.industry.count=='472']<-'few'

table(char\_df$Company.industry.count,useNA = "always")

few Many single <NA

178 37 133 124

tab<-table(char\_df$Dependent.Company.Status,char\_df$Company.industry.count)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 8.9577, df = 2, p-value = 1.231e-14

Clean <- function(x) {

u <- unique(x)

u[which.min(tabulate(match(x, u)))]

}

char\_df$Company.industry.count[is.na(char\_df$Company.industry.count)]<-Clean(char\_df$Company.industry.count)

table(char\_df$Company.industry.count,useNA = "always")

char\_df$Company.mobile.app<- (strsplit(char\_df$Industry.of.company, "|",fixed=T))

for (i in (1:length(char\_df$Industry.of.company)))

{

if(is.na(char\_df$Industry.of.company [i])==T){

char\_df$Company.mobile.app[i]<- NA}

if(isTRUE(str\_detect(strsplit(char\_df$Industry.of.company, "|",fixed=T),"Mobile"))==T)

{ char\_df$Company.mobile.app[i]<-yes}

if(isTRUE(str\_detect(strsplit(char\_df$Industry.of.company, "|",fixed=T),"Mobile"))==F){ char\_df$Company.mobile.app[i]<-no}

}

for (i in (1:length(char\_df$Industry.of.company)))

{

if(is.na(char\_df$Industry.of.company [i])==T){

char\_df$Company.mobile.app[i]<- NA}

if(isTRUE(str\_detect(strsplit(char\_df$Industry.of.company[i], "|",fixed=T),"Mobile"))==T)

{ char\_df$Company.mobile.app[i]<-'yes'}

else { char\_df$Company.mobile.app[i]<-'no'}

}

char\_df$Company.mobile.app

[1] "no" "no" "no" "yes" "no" "no" "no" "no" "yes"

[10] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[19] "no" "yes" "no" "no" "no" "no" "yes" "no" "no"

[28] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[37] "no" "no" "no" "yes" "no" "no" "no" "no" "no"

[46] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[55] "no" "no" "no" "no" "no" "no" "no" "yes" "no"

[64] "yes" "no" "no" "yes" "yes" "yes" "no" "no" "yes"

[73] "no" "no" "no" "yes" "yes" "no" "no" "no" "no"

[82] "no" "no" "no" "no" "yes" "yes" "no" "no" "no"

[91] "no" "no" "yes" "no" "no" "no" "no" "no" "no"

[100] "no" "no" "no" "no" "no" "no" "yes" "no" "no"

[109] "no" "no" "no" "no" "no" "yes" "yes" "no" "no"

[118] "no" "no" "yes" "no" "no" "no" "no" "no" "no"

[127] "no" "no" "no" "no" "yes" "no" "no" "no" "no"

[136] "no" "no" "no" "no" "no" "no" "yes" "yes" "yes"

[145] "no" "yes" "yes" "no" "no" "no" "no" "no" "no"

[154] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[163] "no" "no" "yes" "no" "no" "yes" "no" "no" "no"

[172] "no" "no" "no" "yes" "no" "no" "no" "no" "no"

[181] "no" "no" "no" "no" "yes" "no" "no" "no" "no"

[190] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[199] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[208] "no" "no" "no" "no" "no" "no" "no" "yes" "no"

[217] "no" "no" "no" "yes" "no" "no" "no" "no" "no"

[226] "no" "yes" "no" "no" "no" "no" "no" "no" "no"

[235] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[244] "no" "yes" "no" "no" "no" "no" "no" "no" "no"

[253] "yes" "no" "no" "no" "no" "no" "no" "no" "no"

[262] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[271] "no" "yes" "no" "no" "no" "no" "no" "no" "no"

[280] "no" "yes" "no" "no" "no" "no" "no" "no" "no"

[289] "no" "no" "no" "yes" "yes" "no" "no" "no" "no"

[298] "no" "no" "no" "no" "no" "no" "yes" "yes" "yes"

[307] "no" "no" "no" "no" "no" "yes" "no" "no" "no"

[316] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[325] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[334] "no" "no" "no" "no" "no" "yes" "no" "no" "no"

[343] "no" "no" "yes" "no" "no" "no" "no" "no" "no"

[352] "no" "no" "no" "yes" "no" "no" "no" "no" "no"

[361] "no" "no" "yes" "no" "yes" "no" "no" "no" "no"

[370] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[379] "no" "no" "no" "yes" "yes" "no" "no" "no" "no"

[388] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[397] "yes" "no" "no" "no" "no" "no" "no" "no" "no"

[406] "no" "no" "no" "no" "yes" "no" "no" "no" "no"

[415] "no" "no" "no" "no" "no" "yes" "no" "no" "no"

[424] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[433] "no" "no" "no" "no" "no" "no" "yes" "no" "no"

[442] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[451] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[460] "no" "no" "no" "no" "no" "no" "no" "no" "no"

[469] "no" "no" "no" "no"

table(char\_df$Company.mobile.app)

no yes

418 54

tab<- table(char\_df$Dependent.Company.Status,char\_df$Company.mobile.app)

chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab

X-squared = 0.24778, df = 1, p-value = 0.6186

|  |
| --- |
| char\_df$Experience.in.Fortune.100.organizations[cnt\_df$Experience.in.Fortune.100.organizations==0]<-'no'  char\_df$Experience.in.Fortune.100.organizations[cnt\_df$Experience.in.Fortune.100.organizations==1]<-'yes'  tab<-(char\_df$Dependent.Company.Status,char\_df$Experience.in.Fortune.100.organizations)  Error: unexpected ',' in "tab<-(char\_df$Dependent.Company.Status,"  tab<-table(char\_df$Dependent.Company.Status,char\_df$Experience.in.Fortune.100.organizations)  chisq.test(tab)  Pearson's Chi-squared test with Yates' continuity correction  data: tab  X-squared = 7.4879, df = 1, p-value = 0.006211 |
|  |
| |  | | --- | | table(char\_df$Experience.in.Fortune.100.organizations)  no yes  285 105  Founders\_top\_company\_experience  char\_df$Linear.or.Non.linear.business.model[char\_df$Linear.or.Non.linear.business.model=='Linear']<-'B2B'  char\_df$Linear.or.Non.linear.business.model[char\_df$Linear.or.Non.linear.business.model=='Non-Linear']<-'B2C' | |  | |

tab<-table(char\_df$Dependent.Company.Status,char\_df$Average.Years.of.experience.for.founder.and.co.founder)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 2.4599, df = 2, p-value = 0.2923

tab<-table(char\_df$Dependent.Company.Status,char\_df$Exposure.across.the.globe)

chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab

X-squared = 21.107, df = 1, p-value = 4.344e-06

tab<-table(char\_df$Dependent.Company.Status,char\_df$Highest.education)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 4.5513, df = 2, p-value = 0.1027

tab<-table(char\_df$Dependent.Company.Status,char\_df$Number.of..of.Research.publications)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 17.646, df = 2, p-value = 0.0001473

char\_df$Crowdsourcing.based.business[is.na(char\_df$Crowdsourcing.based.business)]<-Mode(char\_df$Crowdsourcing.based.business)

table(char\_df$Crowdsourcing.based.business)

No Yes

442 30

tab<-table(char\_df$Dependent.Company.Status,char\_df$Crowdsourcing.based.business)

chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab

X-squared = 1.2752, df = 1, p-value = 0.2588

table(char\_df$Crowdfunding.based.business)

No Yes

445 22

tab<-table(char\_df$Dependent.Company.Status,char\_df$Crowdfunding.based.business)

chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab

X-squared = 5.5751, df = 1, p-value = 0.01822

|  |
| --- |
| table(char\_df$Product.or.service.company.)  Both Product Service  24 207 231  chisq.test(tab)  Pearson's Chi-squared test  data: tab  X-squared = 1.7579, df = 2, p-value = 0.4152 |
|  |
| |  | | --- | |  | |

table(char\_df$Dificulty.of.Obtaining.Work.force)

High Low Medium

58 178 150

tab<-table(char\_df$Dependent.Company.Status,char\_df$Dificulty.of.Obtaining.Work.force)

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 39.657, df = 2, p-value = 2.447e-09

tab<-table(char\_df$Dependent.Company.Status,char\_df$Proprietary.or.patent.position..competitive.position.)

chisq.test(tab)

Pearson's Chi-squared test with Yates' continuity correction

data: tab

X-squared = 3.025, df = 1, p-value = 0.08199

tab<-table(char\_df$Dependent.Company.Status,char\_df$Number.of..of.Research.publications

+

+ )

chisq.test(tab)

Pearson's Chi-squared test

data: tab

X-squared = 17.646, df = 2, p-value = 0.0001473

summary(cnt\_df$Number.of.Investors.in.Seed)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.000 0.000 0.000 1.546 2.000 24.000 49

quantile(cnt\_df$Number.of.Investors.in.Seed, probs = seq(0, 1, by= 0.05),na.rm=T)

0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55%

0 0 0 0 0 0 0 0 0 0 0 1

60% 65% 70% 75% 80% 85% 90% 95% 100%

1 1 1 2 2 3 5 7 24

quantile(cnt\_df$Number.of.Investors.in.Seed, probs = seq(.95, 1, by= 0.01),na.rm=T)

95% 96% 97% 98% 99% 100%

7.00 9.00 10.00 11.00 12.78 24.00

cnt\_df$Number.of.Investors.in.Seed[cnt\_df$Number.of.Investors.in.Seed7]<-7

quantile(cnt\_df$Number.of.Investors.in.Angel.and.or.VC

+ , probs = seq(0, 1, by= 0.05),na.rm=T)

0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55%

0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0

60% 65% 70% 75% 80% 85% 90% 95% 100%

0.0 0.0 0.0 0.0 1.0 1.0 2.0 3.9 9.0

quantile(cnt\_df$Number.of.Investors.in.Angel.and.or.VC

+ , probs = seq(.95, 1, by= 0.01),na.rm=T)

95% 96% 97% 98% 99% 100%

3.90 4.00 5.00 6.00 6.78 9.00

cnt\_df$Number.of.Investors.in.Angel.and.or.VC[cnt\_df$Number.of.Investors.in.Angel.and.or.VC2]<-1

cnt\_df$Number.of.Investors.in.Angel.and.or.VC[is.na(cnt\_df$Number.of.Investors.in.Angel.and.or.VC)]<-median(cnt\_df$Number.of.Investors.in.Angel.and.or.VC,na.rm=T)

summary(cnt\_df$Number.of.of.advisors)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.000 0.000 0.000 1.017 1.000 13.000

quantile(cnt\_df$Number.of.of.advisors, probs = seq(0, 1, by= 0.05),na.rm=T)

0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55%

0 0 0 0 0 0 0 0 0 0 0 0

60% 65% 70% 75% 80% 85% 90% 95% 100%

0 1 1 1 2 2 3 6 13

quantile(cnt\_df$Number.of.of.advisors, probs = seq(0.9, 1, by= 0.05),na.rm=T)

90% 95% 100%

3 6 13

cnt\_df$Number.of.of.advisors[Number.of.of.advisors3]<-3

Error in cnt\_df$Number.of.of.advisors[Number.of.of.advisors 3] <- 3 :

object 'Number.of.of.advisors' not found

cnt\_df$Number.of.of.advisors[cnt\_df$Number.of.of.advisors3]<-3

|  |
| --- |
| summary(cnt\_df$Team.size.Senior.leadership)  Min. 1st Qu. Median Mean 3rd Qu. Max.  1.000 2.000 3.000 3.731 5.000 24.000    quantile(cnt\_df$Team.size.Senior.leadership  + , probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20% 25% 30% 35% 40% 45% 50% 55%  1 1 1 2 2 2 2 2 3 3 3 3  60% 65% 70% 75% 80% 85% 90% 95% 100%  4 4 4 5 5 6 7 8 24  quantile(cnt\_df$Team.size.Senior.leadership  + , probs = seq(0.95, 1, by= 0.05),na.rm=T)  95% 100%  8 24  cnt\_df$Team.size.Senior.leadership[cnt\_df$Team.size.Senior.leadership8]<-8  quantile(cnt\_df$Team.size.Senior.leadership  + , probs = seq(0.95, 1, by= 0.05),na.rm=T)  95% 100%  8 8 |
|  |
| |  | | --- | | # Recoding variable levels and converting to factor variable  char\_df$Degree.from.a.Tier.1.or.Tier.2.university.[char\_df$Degree.from.a.Tier.1.or.Tier.2.university.=='Tier\_1']<-1  # Recoding variable levels and converting to factor variable  char\_df$Degree.from.a.Tier.1.or.Tier.2.university.[char\_df$Degree.from.a.Tier.1.or.Tier.2.university.=='Tier\_2']<-2  # Recoding variable levels and converting to factor variable  char\_df$Degree.from.a.Tier.1.or.Tier.2.university.[char\_df$Degree.from.a.Tier.1.or.Tier.2.university.=='Both']<-4  # Recoding variable levels and converting to factor variable  char\_df$Degree.from.a.Tier.1.or.Tier.2.university.[char\_df$Degree.from.a.Tier.1.or.Tier.2.university.=='None']<-0  table(char\_df$Degree.from.a.Tier.1.or.Tier.2.university.)  0 1 2 4  144 139 58 43  summary(cnt\_df$Skills.score)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.00 14.00 21.00 21.69 25.00 200.00 81    table(char\_df$Skills.score)  < table of extent 0  quantile(cnt\_df$Skills.score, probs = seq(0.95, 1, by= 0.05),na.rm=T)  95% 100%  45 200  quantile(cnt\_df$Skills.score, probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20% 25% 30% 35% 40% 45%  0.0 0.0 7.0 10.0 12.0 14.0 14.5 17.0 18.0 19.0  50% 55% 60% 65% 70% 75% 80% 85% 90% 95%  21.0 22.0 24.0 25.0 25.0 25.0 26.0 31.0 37.0 45.0  100%  200.0  quantile(cnt\_df$Skills.score, probs = seq(0.8, 1, by= 0.05),na.rm=T)  80% 85% 90% 95% 100%  26 31 37 45 200  quantile(cnt\_df$Skills.score, probs = seq(0.85, 1, by= 0.05),na.rm=T)  85% 90% 95% 100%  31 37 45 200  cnt\_df$Skills.score[cnt\_df$Skills.score26]<-31  t.test(Skills.score~Dependent.Company.Status, data=cnt\_df)  t.test(Skills.score~Dependent.Company.Status, data=cnt\_df)  Welch Two Sample t-test  data: Skills.score by Dependent.Company.Status  t = -1.3804, df = 136.54, p-value = 0.1697  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -3.6528554 0.6495551  sample estimates:  mean in group Failed mean in group Success  18.33333 19.83498 | |

t.test(Number.of.Investors.in.Seed~Dependent.Company.Status, data=cnt\_df)

Welch Two Sample t-test

data: Number.of.Investors.in.Seed by Dependent.Company.Status

t = -3.7638, df = 280.1, p-value = 0.0002039

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-1.1040033 -0.3457646

sample estimates:

mean in group Failed mean in group Success

0.7933884 1.5182724

t.test(Number.of.Investors.in.Angel.and.or.VC~Dependent.Company.Status, data=cnt\_df)

Welch Two Sample t-test

data: Number.of.Investors.in.Angel.and.or.VC by Dependent.Company.Status

t = -1.8294, df = 375.83, p-value = 0.06813

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-0.183787519 0.006628679

sample estimates:

mean in group Failed mean in group Success

0.1976048 0.2861842

|  |
| --- |
| summary(cnt\_df$Percent\_skill\_Entrepreneurship)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 5.882 7.538 11.110 100.000 61    quantile(cnt\_df$Percent\_skill\_Entrepreneurship  + , probs = seq(0.8, 1, by= 0.05),na.rm=T)  80% 85% 90% 95% 100%  11.76471 14.72222 16.66667 20.00000 100.00000  cnt\_df$Percent\_skill\_Entrepreneurship[cnt\_df$Percent\_skill\_Entrepreneurship20]<-21  quantile(cnt\_df$Percent\_skill\_Entrepreneurship  + , probs = seq(0.95, 1, by= 0.05),na.rm=T)  95% 100%  20 21  cnt\_df$Percent\_skill\_Entrepreneurship[cnt\_df$Percent\_skill\_Entrepreneurship20]<-20 |
| summary(cnt\_df$Percent\_skill\_Marketing)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 5.556 11.000 14.290 76.470 61    quantile(cnt\_df$Percent\_skill\_Marketing, probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20% 25% 30% 35% 40%  0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000  45% 50% 55% 60% 65% 70% 75% 80% 85%  2.777778 5.555556 5.800654 7.142857 9.232955 11.342593 14.285714 19.780220 24.128540  90% 95% 100%  33.333333 50.000000 76.470588    quantile(cnt\_df$Percent\_skill\_Marketing, probs = seq(0.8, 1, by= 0.04),na.rm=T)  80% 84% 88% 92% 96% 100%  19.78022 23.52941 30.00000 38.95833 54.50980 76.47059  cnt\_df$Percent\_skill\_Marketing[cnt\_df$Percent\_skill\_Marketing30]<-30 |
| |  | | --- | | cnt\_df$Percent\_skill\_Operations[cnt\_df$Percent\_skill\_Operations20]<-20    summary(cnt\_df$Percent\_skill\_Operations)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 0.000 2.385 3.452 50.000 61    quantile(cnt\_df$Percent\_skill\_Operations, probs = seq(0.8, 1, by= 0.05),na.rm=T)  80% 85% 90% 95% 100%  5.555556 5.752996 7.142857 11.764706 50.000000  cnt\_df$Percent\_skill\_Operations[cnt\_df$Percent\_skill\_Operations20]<-21  quantile(cnt\_df$Percent\_skill\_Operations, probs = seq(0.95, 1, by= 0.05),na.rm=T)  95% 100%  11.76471 50.00000  cnt\_df$Percent\_skill\_Operations[cnt\_df$Percent\_skill\_Operations20]<-20    summary(cnt\_df$Percent\_skill\_Engineering)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 9.804 18.630 28.660 100.000 61  quantile(cnt\_df$Percent\_skill\_Engineering, probs = seq(0.8, 1, by= 0.05),na.rm=T)  80% 85% 90% 95% 100%  36.11111 43.11146 50.00000 69.95798 100.00000 | |

|  |
| --- |
| summary(cnt\_df$Percent\_skill\_Leadership)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 0.000 2.870 5.556 40.000 61    quantile(cnt\_df$Percent\_skill\_Leadership, probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20% 25% 30% 35% 40%  0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000 0.000000  45% 50% 55% 60% 65% 70% 75% 80% 85%  0.000000 0.000000 0.000000 0.000000 1.960784 3.125000 5.555556 5.882353 6.250000  90% 95% 100%  9.375000 11.220044 40.000000    quantile(cnt\_df$Percent\_skill\_Leadership, probs = seq(0.8, 1, by= 0.04),na.rm=T)  80% 84% 88% 92% 96% 100%  5.882353 6.250000 8.205128 11.111111 14.285714 40.000000      cnt\_df$Percent\_skill\_Leadership[cnt\_df$Percent\_skill\_Leadership14]<-14 |
|  |
| |  | | --- | |  | |

|  |
| --- |
| quantile(cnt\_df$Percent\_skill\_Business.Strategy, probs = seq(0.8, 1, by= 0.04),na.rm=T)  80% 84% 88% 92% 96% 100%  20.00000 22.22222 24.73856 27.77778 29.76471 50.00000      summary(cnt\_df$Percent\_skill\_Product.Management)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 0.000 3.430 5.556 25.000 61    quantile(cnt\_df$Percent\_skill\_Product.Management,probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20%  0.000000 0.000000 0.000000 0.000000 0.000000  25% 30% 35% 40% 45%  0.000000 0.000000 0.000000 0.000000 0.000000  50% 55% 60% 65% 70%  0.000000 2.777778 2.941176 4.166667 5.555556  75% 80% 85% 90% 95%  5.555556 6.250000 7.233045 9.090909 12.500000  100%  25.000000    quantile(cnt\_df$Percent\_skill\_Product.Management  + ,probs = seq(0.8, 1, by= 0.04),na.rm=T)  80% 84% 88% 92% 96%  6.250000 7.142857 8.333333 10.176471 15.598291  100%  25.000000      cnt\_df$Percent\_skill\_Product.Management[cnt\_df$Percent\_skill\_Product.Management10]<-10 |
|  |
| |  |  |  |  |  | | --- | --- | --- | --- | --- | | |  | | --- | | summary(cnt\_df$Percent\_skill\_Sales)  Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  0.000 0.000 0.000 3.357 5.556 33.330 61    quantile(cnt\_df$Percent\_skill\_Sales,probs = seq(0, 1, by= 0.05),na.rm=T)  0% 5% 10% 15% 20%  0.000000 0.000000 0.000000 0.000000 0.000000  25% 30% 35% 40% 45%  0.000000 0.000000 0.000000 0.000000 0.000000  50% 55% 60% 65% 70%  0.000000 0.000000 2.380952 3.333333 5.263158  75% 80% 85% 90% 95%  5.555556 5.882353 6.683007 9.920635 13.392857  100%  33.333333    quantile(cnt\_df$Percent\_skill\_Sales  + ,probs = seq(0.8, 1, by= 0.04),na.rm=T)  80% 84% 88% 92% 96%  5.882353 6.416667 8.333333 11.111111 17.254902  100%  33.333333        cnt\_df$Percent\_skill\_Sales[cnt\_df$Percent\_skill\_sales17]<-17 | |  | | |  | | --- | |  |   **Model Creation and Data Store Automation** | | |

model<-glm(formula = Dependent ~ Company\_competitor\_count + Company\_business\_model

+ Company\_1st\_investment\_time +Company\_avg\_investment\_time

+ Company\_crowdsourcing + Company\_crowdfunding, , family = binomial(link = logit), data = train\_final)

summary(model)

exp(cbind(OR = coef(model), confint(model)))

library(ResourceSelection)

hoslem.test(train$Dependent,model$fitted.values, g=10)

# Prediction on test set

pred\_prob<-predict (model, newdata=test\_new, type="response")

# model accuracy measures

library (ROCR)

pred <- prediction (pred\_prob, test\_new$Dependent)

# Area under the curve

Performance (pred, 'auc')

# creating ROC curve

roc <- performance (pred,"tpr","fpr")

plot (roc)

prediction<- predict(model, newdata=test, type="response")

resultFrame<- cbind(test\_new$Company\_id,prediction)

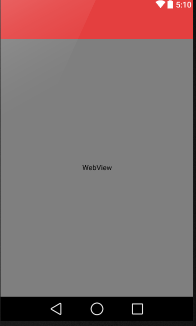
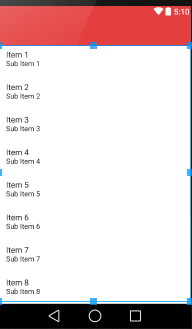
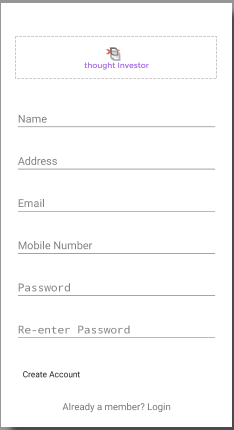
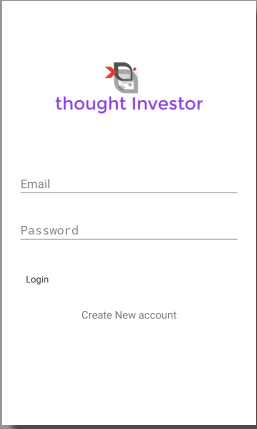
resultFrame<-cbind(test,test\_new)

library(RMySQL)

mydb = dbConnect(MySQL(), user='user', password='welcome123', dbname='mydb', host='mobile.c3fat3s4cj31.us-west-2.rds.amazonaws.com')

dbWriteTable(mydb, name='Prediction', value=resultFrame)

**Mobile UI Design Principles – Storyboard, Wireframes**



Existing

New

On Click of Item

Register

The above diagram gives a complete picture of the story of this Android Application .

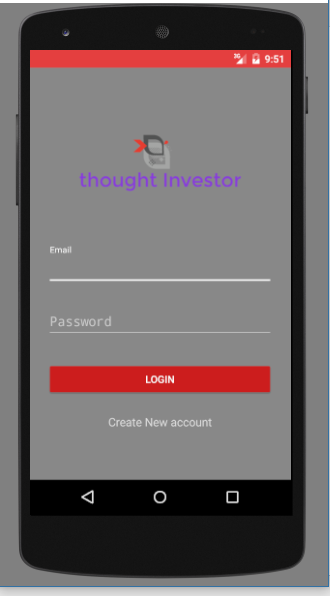
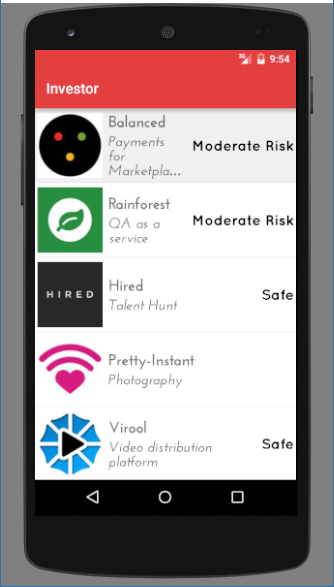
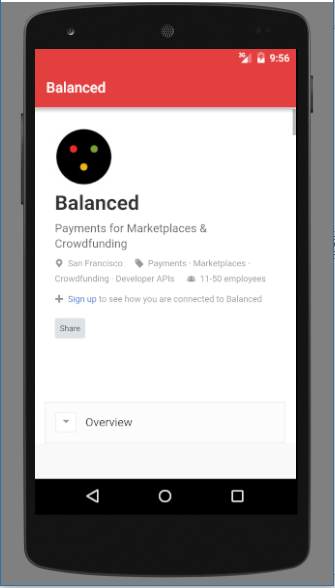
The user first views a login screen that further redirects to the list of items page in case the user is an existing one.

If the user is a new one and wants to register into the application that option is available on clicking the Create New Account button as shown above.

Then further s/he is redirected to the list of items frame.

To provide the user with the complete detail of the Items a web view is provided displaying the holistic information about the company selected.

The below diagram show the execution state of the story.

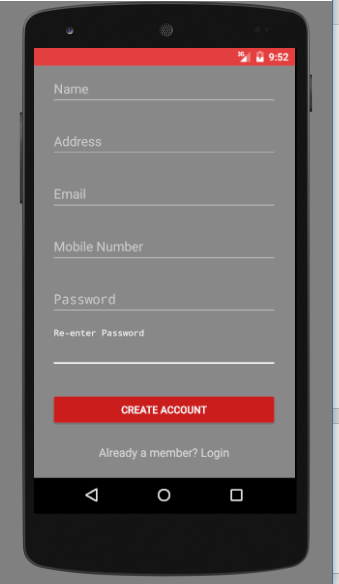


On Item Clicked

Existing

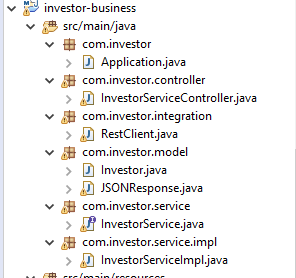
Register

New Login

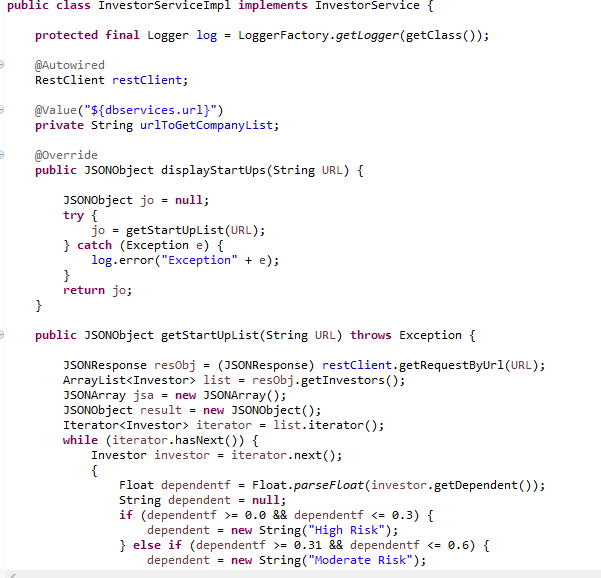


* 1. **Client Side Design**

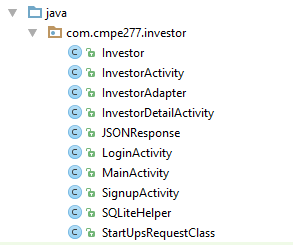
The client side design is important receptor of the restful services.

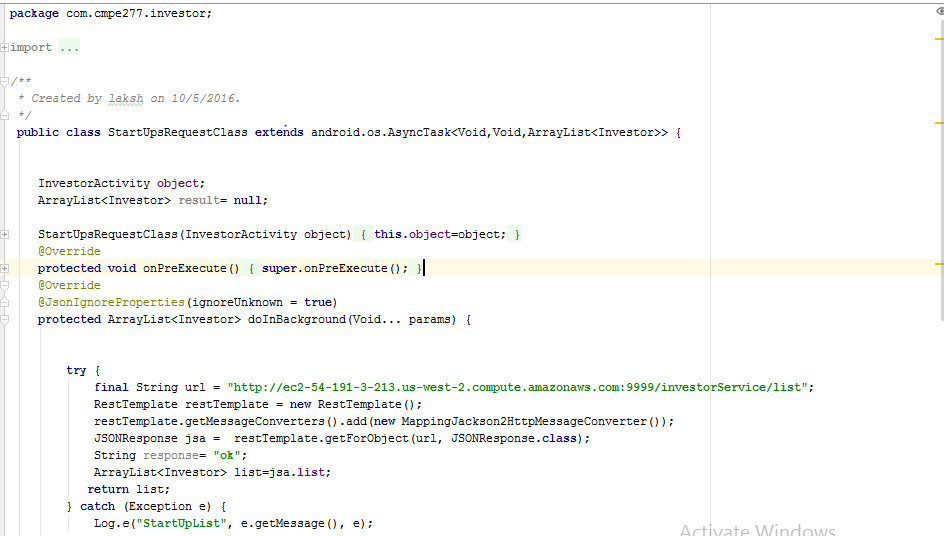


The business rules are applied in this phase and the Restful Client over http consumption for the middle tier API occurs at this location.



In the Android Application the Client Design is as follows





**Data Store**

This primarily involves a RDS MySQL system stores in Amazon RDS cloud.

For the Client side authentication storage is done in the SQLite .