Capstone Project

Machine Learning Engineer Nanodegree

Keyvan Tajbakhsh
Aug 12th, 2019

Customer Segmentation Report for Arvato Financial Services

I. Definition

Project Overview

Nowaday, all business organizations are adopting datadriven strategies to generate more profits out of their business. Growing startups are investing a lot of funds in data structures to maximize profits of the business group by developing intelligent tools backed by machine learning and artificial intelligence.

Customer segmentation allows a business to precisely reach a consumer with specific needs and wants. In the long term, this benefits the company, because they are able to use their corporate resources more effectively and make better strategic marketing decisions.

In other words this is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

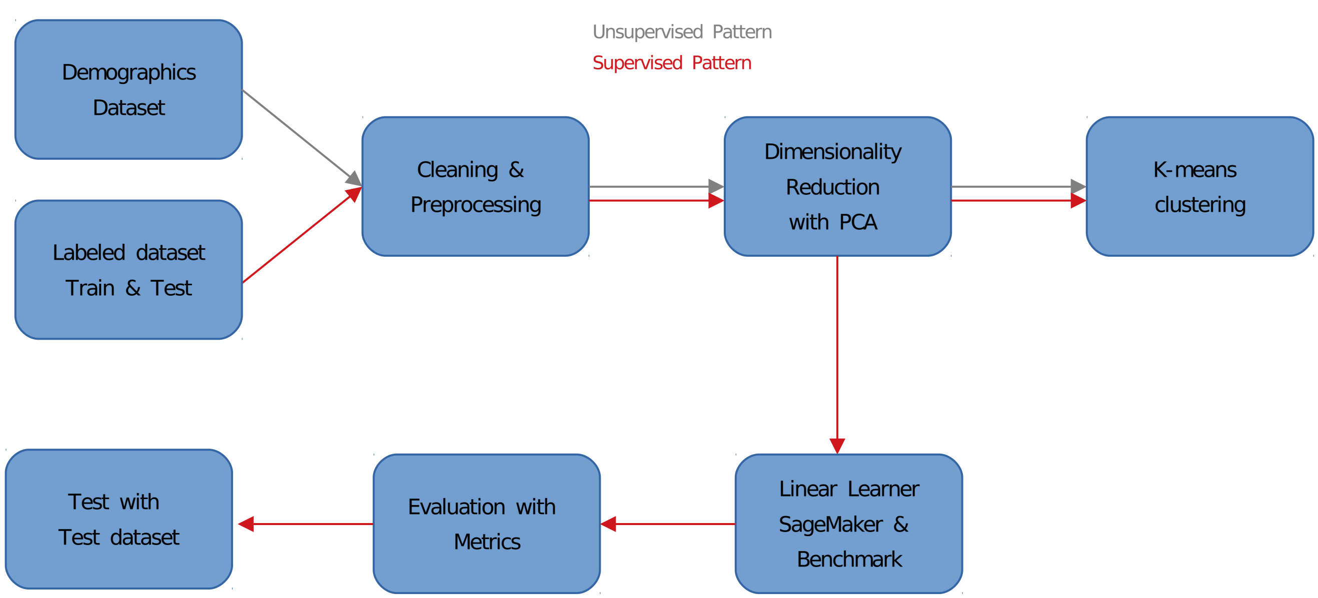
Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

This project relies on identifying key differentiators that divide customers into groups that can be targeted. Information such as a customers demographics (age, race, religion, gender, family size, ethnicity, income, education level), geography (where they live and work), psychographic (social class, lifestyle and personality characteristics) and behavioral (spending, consumption, usage and desired benefits) tendencies are taken into account when determining customer segmentation practices.

For this purpose we will use unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general geographical population of Germany. The datasets provided need to be treated and prepared before implementing machine learning algorithms.

Then our cluster analysis will be used to implement our supervised learning algorithm. In this context we will train and implement a supervised algorithm able to predict if a customer will respond positively to the mail-order campaign or not (binary classification problem). Then we will create a benchmark model to compare our final result and test the data.

Here below we have represented the workflow of this project and how we will proceed.



Problem Statement

The goal in this project is to create a model capable of predicting which individual is likely to be in mail-order list of the marketing campaign or not. To do this we break it into two parts as described below:

1. Customer Segmentation Report

In this section two datasets are provided for creating our unsupervised model. We have to note that all datasets provided will be treated (cleaning and preprocessing) in the same manner before implementing our models. We will discuss later about how to preprocess the data. For now we describe this section by dividing it in two parts; 1) Principal Components Analysis (PCA) for dimensionality reduction (simplification of data) and 2) K-means Clustering to create group of individuals and relate these groups to our mail-order marketing campaign.

2. Supervised Learning Model

Once we have created our unsupervised model with K-means Clustering to groups of customers and identify in which cluster customers are more likely to be in mail-order campaign. Now it’s time to build a supervised prediction model. The goal in this section is to create a binary classifier model. This model will be able to describe parts of the general population that are more likely to be part of the mail-order company's main customer base (labeled 1), and which parts of the general population are less so (labeled 0). Then we will create another model as a benchmark to comparison to our binary classifier.

Metrics

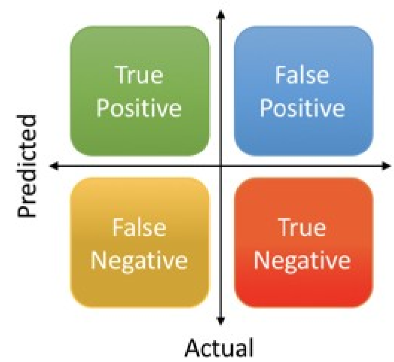
In this project, we'll want to evaluate the performance of our binary classifier and compare it to our benchmark; training it on some training data and testing it on test data that it did not see during the training process.

Once our model is trained, we can see how it performs when applied to the test data. To evaluate our predictor we'll calculate false negatives and positives as well as recall, precision, and accuracy.

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



True Positive = Correctly predicted part of mail list (1s are 1s)
True Negative = Correctly predicted not part of mail list (0s are 0s)
False Positive = Uncorrectly predicted part of mail list (0s are 1s)
False Negative = Uncorrectly predicted not part of mail list (1s are 0s)

II. Analysis

Data Exploration

Data features of datasets are provided in `DIAS Information Levels - Attributes 2017.csv` file and described as follows:

1. KBA05_DIESEL : share of cars with Diesel-engine in the microcell
2. KBA13_BJ_2009 : share of cars built in 2009 within the PLZ8
3. KBA05_ANTG4 : number of >10 family houses in the cell
4. KBA13_VW : share of VOLKSWAGEN within the PLZ8
5. KBA05_HERST4 : share of European manufacturer (e.g. Fiat, Peugeot, Rover,...)
6. KBA13_KW_110 : share of cars with an engine power between 91 and 110 KW - PLZ8
7. KBA13_SITZE_5 : number of cars with 5 seats in the PLZ8
8. KBA13_KW_30 : share of cars up to 30 KW engine power - PLZ8
9. D19_GESAMT_OFFLINE_DATUM : actuality of the last transaction with the complete file OFFLINE
10. KBA13_KMH_140_210 : share of cars with max speed between 140 and 210 km/h within the PLZ8
11. KBA13_KRSHERST_FORD_OPEL : share of FORD/Opel (referred to the county average) - PLZ8
12. KBA05_SEG2 : share of small and very small cars (Ford Fiesta, Ford Ka etc.) in the microcell
13. KBA05_ZUL2 : share of cars built between 1994 and 2000

14.ZABEOTYP : typification of energy consumers

15.ORTSGR_KLS9 : size of the community

16.PLZ8_ANTG1 : number of 1-2 family houses in the PLZ8

17.D19_VERSAND_ONLINE_QUOTE_12 : amount of online transactions within all transactions in the segment mail-order

18.KBA13_ANZAHL_PKW : number of cars in the PLZ8

19.KBA13_CCM_1500 : share of cars with 1400ccm to 1499ccm within the PLZ8

20.D19_BANKEN_ANZ_12 : transaction activity BANKS in the last 12 months

21.KBA05_MOTRAD : share of motorcycles per household

22.KBA13_KW_80 : share of cars with an engine power between 71 and 80 KW - PLZ8

23.KBA13_MOTOR : most common motor size within the PLZ8

24.GREEN_AVANTGARDE : Green avantgarde

25.KBA05_KRSKLEIN : share of small cars (referred to the county average)

26.KBA05_MAXBJ : most common age of the cars in the microcell

27.KBA13_KW_121 : share of cars with an engine power more than 120 KW - PLZ8

28.SEMIO_TRADV : affinity indicating in what way the person is traditional minded

29.KBA05_ANTG2 : number of 3-5 family houses in the cell

30.SEMIO_DOM : affinity indicating in what way the person is dominant minded

31.KBA13_HALTER_50 : share of car owners between 46 and 50 within the PLZ8

32.D19_VERSAND_ONLINE_DATUM : actuality of the last transaction for the segment mail-order ONLINE

33.FINANZ_SPARER : financial typology: money saver

34.KBA13_KMH_180 : share of cars with max speed between 110 km/h and 180km/h within the PLZ8

35.PLZ8_ANTG4 : number of >10 family houses in the PLZ8

36.KBA13_SEG_OBEREMITTELKLASSE : share of upper middle class cars and upper class cars (BMW5er, BMW7er etc.)

37.KBA05_SEG4 : share of middle class cars (Ford Mondeo etc.) in the microcell

38.KBA13_HALTER_65 : share of car owners between 61 and 65 within the PLZ8

39.ANZ_HAUSHALTE_AKTIV : number of households in the building

40.KBA13_KW_90 : share of cars with an engine power between 81 and 90 KW - PLZ8

41.SHOPPER_TYP : shopping typology

42.KBA05_KRSOBER : share of upper class cars (referred to the county average)

43.FINANZ_MINIMALIST : financial typology: low financial interest

44.GEBAEUDETYP : type of building (residential or commercial)

45.EWDICHTE : density of inhabitants per square kilometer

46.KBA13_KRSSEG_VAN : share of vans (referred to the county average) - PLZ8

47.KBA13_VORB_2 : share of cars with 2 preowner - PLZ8

48.LP_STATUS_GROB : social status rough

49.FINANZ_VORSORGER : financial typology: be prepared

50.PRAEGENDE_JUGENDJAHRE : dominating movement in the person's youth (avantgarde or mainstream)

51.KBA05_ALTER4 : share of cars owners elder than 61 years

52.KBA13_SITZE_4 : number of cars with less than 5 seats in the PLZ8

53.OST_WEST_KZ : flag indicating the former GDR/FRG

54.KBA05_AUTOQUOT : share of cars per household

55.KBA13_BJ_2004 : share of cars built before 2004 within the PLZ8

56.KBA13_KW_120 : share of cars with an engine power between 111 and 120 KW - PLZ8

57.KBA05_GBZ : number of buildings in the microcell

58.D19_TELKO_ONLINE_DATUM : actuality of the last transaction for the segment telecommunication ONLINE

59.KBA05_KRSAQUOT : share of cars per household (reffered to county average)

60.D19_BANKEN_DATUM : actuality of the last transaction for the segment banks TOTAL

61.KBA05_HERST5 : share of asian manufacturer (e.g. Toyota, Kia,...)

62.KBA05_MOD3 : share of Golf-class cars (in an AZ specific definition)

63.KBA05_SEG5 : share of upper middle class cars and upper class cars (BMW5er, BMW7er etc.)

64.KONSUMNAEHE : distance from a building to PoS (Point of Sale)

65.CAMEO_DEU_2015 : CAMEO classification 2015 - detailled classification

66.SEMIO_KRIT : affinity indicating in what way the person is critical minded

67.AGER_TYP : best-ager typology

68.KBA13_FIAT : share of FIAT within the PLZ8

69.HEALTH_TYP : health typology

70.ALTERSKATEGORIE_GROB : age classification through prename analysis

71.KBA13_HALTER_20 : share of car owners below 21 within the PLZ8

72.SEMIO_KULT : affinity indicating in what way the person is cultural minded

73.KBA13_NISSAN : share of NISSAN within the PLZ8

74.D19_BANKEN_OFFLINE_DATUM : actuality of the last transaction for the segment banks OFFLINE

75.KBA13_HALTER_60 : share of car owners between 56 and 60 within the PLZ8

76.FINANZ_UNAUFFAELLIGER : financial typology: unremarkable

77.KBA05_KRSHERST1 : share of Mercedes/BMW (reffered to the county average)

78.KBA05_MOD2 : share of middle class cars (in an AZ specific definition)

79.D19_VERSAND_ANZ_24 : transaction activity MAIL-ORDER in the last 24 months

80.KBA13_KW_0_60 : share of cars up to 60 KW engine power - PLZ8

81.KBA05_VORB0 : share of cars with no preowner

82.KBA13_BJ_2008 : share of cars built in 2008 within the PLZ8

83.KBA13_CCM_1200 : share of cars with 1000ccm to 1199ccm within the PLZ8

84.KBA13_KRSHERST_BMW_BENZ : share of BMW/Mercedes Benz (referred to the county average) - PLZ8

85.D19_GESAMT_ANZ_24 : transaction activity TOTAL POOL in the last 24 months

86.KBA05_SEG8 : share of roadster and convertables in the microcell

87.D19_VERSAND_OFFLINE_DATUM : actuality of the last transaction for the segment mail-order OFFLINE

88.SEMIO_KAEM : affinity indicating in what way the person is of a fightfull attitude

89.W_KEIT_KIND_HH : likelihood of a child present in this household

90.KBA13_MAZDA : share of MAZDA within the PLZ8

91.KBA05_ANTG3 : number of 6-10 family houses in the cell

92.KBA05_MOTOR : most common engine size in the microcell

93.ANZ_PERSONEN : number of adult persons in the household

94.KBA13_OPEL : share of OPEL within the PLZ8

95.KBA13_KMH_251 : share of cars with a greater max speed than 250 km/h within the PLZ8

96.KBA13_CCM_2501 : share of cars with more than 2500ccm within the PLZ8

97.KBA13_VORB_1 : share of cars with 1 preowner - PLZ8

98.KBA13_MERCEDES : share of MERCEDES within the PLZ8

99.KBA13_VORB_3 : share of cars with 3 or more preowner - PLZ8

100.ONLINE_AFFINITAET : online affinity

101.PLZ8_ANTG3 : number of 6-10 family houses in the PLZ8

102.D19_TELKO_ANZ_12 : transaction activity TELCO in the last 12 months

103.KBA05_SEG3 : share of lowe midclass cars (Ford Focus etc.) in the microcell

104.KBA05_ZUL1 : share of cars built before 1994

105.KBA13_SEG_UTILITIES : share of MUVs/SUVs within the PLZ8

106.KBA05_HERSTTEMP : development of the most common car manufacturers in the neighbourhood

107.KBA05_MAXVORB : most common preowner structure in the microcell

108.KBA05_ANTG1 : number of 1-2 family houses in the cell

109.KBA05_MAXAH : most common age of car owners in the microcell

110.KBA13_KMH_250 : share of cars with max speed between 210 and 250 km/h within the PLZ8

111.KBA13_SEG_MITTELKLASSE : share of middle class cars (Ford Mondeo etc.) in the PLZ8

112.KBA13_SEG_MINIVANS : share of minivans within the PLZ8

113.RELAT_AB : share of unemployed in relation to the county the community belongs to

114.ANREDE_KZ : gender

115.GFK_URLAUBERTYP : vacation habits

116.KBA05_MOD1 : share of upper class cars (in an AZ specific definition)

117.KBA13_CCM_3001 : share of cars with more than 3000ccm within the PLZ8

118.KBA05_KRSVAN : share of vans (referred to the county average)

119.KBA13_CCM_3000 : share of cars with 2500ccm to 2999ccm within the PLZ8

120.KBA13_PEUGEOT : share of PEUGEOT within the PLZ8

121.KBA13_TOYOTA : share of TOYOTA within the PLZ8

122.KBA13_HALTER_35 : share of car owners between 31 and 35 within the PLZ8

123.KBA13_BJ_1999 : share of cars built between 1995 and 1999 within the PLZ8

124.KBA13_CCM_2500 : share of cars with 2000ccm to 2499ccm within the PLZ8

125.KBA05_MAXHERST : most common car manufacturer in the microcell

126.KBA13_RENAULT : share of RENAULT within the PLZ8

127.KBA13_HALTER_40 : share of car owners between 36 and 40 within the PLZ8

128.D19_VERSI_ANZ_24 : transaction activity INSURANCE in the last 24 months

129.D19_VERSAND_ANZ_12 : transaction activity MAIL-ORDER in the last 12 months

130.KBA13_HALTER_45 : share of car owners between 41 and 45 within the PLZ8

131.KBA13_SEG_KLEINWAGEN : share of small and very small cars (Ford Fiesta, Ford Ka etc.) in the PLZ8

132.D19_BANKEN_ANZ_24 : transaction activity BANKS in the last 24 months

133.KBA05_SEG10 : share of more specific cars (Vans, convertables, all-terrains, MUVs etc.)

134.KBA13_HERST_FORD_OPEL : share of Ford & Opel/Vauxhall within the PLZ8

135.KKK : purchasing power

136.KBA05_KW1 : share of cars with less than 59 KW engine power

137.KBA05_MAXSEG : most common car segment in the microcell

138.SEMIO_VERT : affinity indicating in what way the person is dreamily

139.KBA05_MOD4 : share of small cars (in an AZ specific definition)

140.D19_VERSAND_DATUM : actuality of the last transaction for the segment mail-order TOTAL

141.BALLRAUM : distance to next urban centre

142.KBA13_BMW : share of BMW within the PLZ8

143.KBA13_SEG_GELAENDEWAGEN : share of allterrain within the PLZ8

144.LP_LEBENSPHASE_GROB : lifestage rough

145.KBA13_BJ_2006 : share of cars built between 2005 and 2006 within the PLZ8

146.KBA05_ZUL4 : share of cars built from 2003 on

147.SEMIO_PFLICHT : affinity indicating in what way the person is dutyfull traditional minded

148.KBA05_SEG7 : share of all-terrain vehicles and MUVs in the microcell

149.MIN_GEBAEUDEJAHR : year the building was first mentioned in our database

150.KBA05_ALTER1 : share of car owners less than 31 years old

151.LP_LEBENSPHASE_FEIN : lifestage fine

152.KBA05_BAUMAX : most common building-type within the cell

153.D19_VERSI_ANZ_12 : transaction activity INSURANCE in the last 12 months

154.KBA13_KW_60 : share of cars with an engine power between 51 and 60 KW - PLZ8

155.ANZ_HH_TITEL : number of academic title holder in building

156.KBA13_SEG_GROSSRAUMVANS : share of big sized vans within the PLZ8

157.KBA05_CCM3 : share of cars with 1800ccm to 2499 ccm

158.KBA13_ALTERHALTER_45 : share of car owners between 31 and 45 within the PLZ8

159.KBA13_HALTER_66 : share of car owners over 66 within the PLZ8

160.MOBI_REGIO : moving patterns

161.CAMEO_DEUG_2015 : CAMEO classification 2015 - Uppergroup

162.KBA13_ALTERHALTER_61 : share of car owners elder than 61 within the PLZ8

163.ANZ_TITEL : number of professional title holder in household

164.SEMIO_REL : affinity indicating in what way the person is religious

165.KBA13_CCM_1800 : share of cars with 1600ccm to 1799ccm within the PLZ8

166.KBA13_HALTER_25 : share of car owners between 21 and 25 within the PLZ8

167.KBA13_HERST_EUROPA : share of European cars within the PLZ8

168.D19_TELKO_OFFLINE_DATUM : actuality of the last transaction for the segment telecommunication OFFLINE

169.KBA13_CCM_0_1400 : share of cars with less than 1400ccm within the PLZ8

170.D19_GESAMT_ANZ_12 : transaction activity TOTAL POOL in the last 12 months

171.KBA13_AUDI : share of AUDI within the PLZ8

172.KBA13_KRSZUL_NEU : share of newbuilt cars (referred to the county average) - PLZ8

173.GEBAEUDETYP_RASTER : industrial areas

174.FINANZ_ANLEGER : financial typology: investor

175.KBA13_ALTERHALTER_60 : share of car owners between 46 and 60 within the PLZ8

176.KBA13_FAB_ASIEN : share of other Asian Manufacturers within the PLZ8

177.FINANZTYP : best descirbing financial type for the person

178.KBA05_HERST3 : share of Ford/Opel

179.KBA13_CCM_1600 : share of cars with 1500ccm to 1599ccm within the PLZ8

180.FINANZ_HAUSBAUER : financial typology: main focus is the own house

181.KBA13_CCM_1400 : share of cars with 1200ccm to 1399ccm within the PLZ8

182.KBA13_KW_61_120 : share of cars with an engine power between 61 and 120 KW - PLZ8

183.KBA13_SEG_VAN : share of vans within the PLZ8

184.D19_GESAMT_ONLINE_DATUM : actuality of the last transaction with the complete file ONLINE

185.D19_TELKO_DATUM : actuality of the last transaction for the segment telecommunication TOTAL

186.KBA13_KW_70 : share of cars with an engine power between 61 and 70 KW - PLZ8

187.SEMIO_MAT : affinity indicating in what way the person is material minded

188.KBA05_MOD8 : share of vans (in an AZ specific definition)

189.KBA05_CCM1 : share of cars with less than 1399ccm

190.D19_BANKEN_ONLINE_DATUM : actuality of the last transaction for the segment banks ONLINE

191.PLZ8_GBZ : number of buildings within the PLZ8

192.KBA05_KRSHERST3 : share of Ford/Opel (reffered to the county average)

193.KBA05_VORB1 : share of cars with one or two preowner

194.KBA05_VORB2 : share of cars with more than two preowner

195.KBA05_KRSHERST2 : share of Volkswagen (reffered to the county average)

196.KBA05_KRSZUL : share of newbuilt cars (referred to the county average)

197.KBA13_CCM_1000 : share of cars with less than 1000ccm within the PLZ8

198.KBA13_HERST_ASIEN : share of Asian Manufacturers within the PLZ8

199.KBA13_HERST_BMW_BENZ : share of BMW & Mercedes Benz within the PLZ8

200.KBA13_KMH_140 : share of cars with max speed between 110 km/h and 140km/h within the PLZ8

201.KBA13_AUTOQUOTE : share of cars per household within the PLZ8

202.KBA13_FAB_SONSTIGE : share of other Manufacturers within the PLZ8

203.KBA13_KRSHERST_AUDI_VW : share of Volkswagen (referred to the county average) - PLZ8

204.KBA13_VORB_0 : share of cars with no preowner - PLZ8

205.KBA13_KW_40 : share of cars with an engine power between 31 and 40 KW - PLZ8

206.KBA05_MODTEMP : development of the most common car segment in the neighbourhood

207.KBA05_SEG6 : share of upper class cars (BMW 7er etc.) in the microcell

208.KBA13_SEG_SPORTWAGEN : share of sportscars within the PLZ8

209.CJT_GESAMTTYP : customer journey typology

210.KBA13_KMH_0_140 : share of cars with max speed 140 km/h within the PLZ8

211.D19_BANKEN_ONLINE_QUOTE_12 : amount of online transactions within all transactions in the segment bank

212.KBA05_CCM4 : share of cars with more than 2499ccm

213.KBA05_SEG9 : share of vans in the microcell

214.VERS_TYP : insurance typology

215.KBA05_KW2 : share of cars with an engine power between 60 and 119 KW

216.TITEL_KZ : flag whether this person holds an academic title

217.KBA05_HERST1 : share of top German manufacturer (Mercedes, BMW)

218.KBA13_SEG_KLEINST : share of very small cars (Ford Ka etc.) in the PLZ8

219.KBA13_SEG_SONSTIGE : share of other cars within the PLZ8

220.KBA13_CCM_2000 : share of cars with 1800ccm to 1999ccm within the PLZ8

221. D19_GESAMT_ONLINE_QUOTE_12 : amount of online transactions within all transactions in the complete file

222. KBA05_CCM2 : share of cars with 1400ccm to 1799 ccm

223. KBA05_ZUL3 : share of cars built between 2001 and 2002

224. LP_FAMILIE_FEIN : familytyp fine

225. KBA13_SITZE_6 : number of cars with more than 5 seats in the PLZ8

226. SEMIO_RAT : affinity indicating in what way the person is of a rational mind

227. KBA05_HERST2 : share of Volkswagen-Cars (including Audi)

228. KBA13_HALTER_30 : share of car owners between 26 and 30 within the PLZ8

229. D19_GESAMT_DATUM : actuality of the last transaction with the complete file TOTAL

230. INNENSTADT : distance to the city centre

231. KBA13_KW_50 : share of cars with an engine power between 41 and 50 KW - PLZ8

232. KBA13_ALTERHALTER_30 : share of car owners below 31 within the PLZ8

233. KBA13_SEG_OBERKLASSE : share of upper class cars (BMW 7er etc.) in the PLZ8

234. LP_FAMILIE_GROB : familytyp rough

235. NATIONALITAET_KZ : nationaltity (scored by prename analysis)

236. SEMIO_ERL : affinity indicating in what way the person is eventful orientated

237. ALTER_HH : main age within the household

238. KBA05_ALTER3 : share of car owners inbetween 45 and 60 years of age

239. KBA05_KW3 : share of cars with an engine power of more than 119 KW

240. WOHNLAGE : residential-area

241. HH_EINKOMMEN_SCORE : estimated household net income

242.KBA13_KRSSEG_KLEIN : share of small cars (referred to the county average) - PLZ8

243.KBA13_SEG_KOMPAKTKLASSE : share of lowe midclass cars (Ford Focus etc.) in the PLZ8

244.KBA05_ANHANG : share of trailers in the microcell

245.KBA05_FRAU : share of female car owners

246.D19_KONSUMTYP : consumption type

247.KBA13_VORB_1_2 : share of cars with 1 or 2 preowner - PLZ8

248.WOHNDAUER_2008 : length of residence

249.KBA05_SEG1 : share of very small cars (Ford Ka etc.) in the microcell

250.REGIOTYP : neighbourhood

251.KBA13_SEG_MINIWAGEN : share of minicars within the PLZ8

252.PLZ8_BAUMAX : most common building-type within the PLZ8

253.RETOURTYP_BK_S : return type

254.KBA13_FORD : share of FORD within the PLZ8

255.KBA13_HALTER_55 : share of car owners between 51 and 55 within the PLZ8

256.KBA13_HERST_AUDI_VW : share of Volkswagen & Audi within the PLZ8

257.KBA13_KMH_110 : share of cars with max speed 110 km/h within the PLZ8

258.PLZ8_ANTG2 : number of 3-5 family houses in the PLZ8

259.KBA05_ALTER2 : share of car owners inbetween 31 and 45 years of age

260.D19_TELKO_ANZ_24 : transaction activity TELCO in the last 24 months

261.KBA13_KMH_211 : share of cars with a greater max speed than 210 km/h within the PLZ8

262.KBA13_KRSAQUOT : share of cars per household (referred to the county average) - PLZ8

263.LP_STATUS_FEIN : social status fine

264.KBA13_BJ_2000 : share of cars built between 2000 and 2003 within the PLZ8

265.SEMIO_LUST : affinity indicating in what way the person is sensual minded

266.KBA13_SEG_WOHNMOBILE : share of roadmobiles within the PLZ8

267.SEMIO_FAM : affinity indicating in what way the person is familiar minded

268.PLZ8_HHZ : number of households within the PLZ8

269.KBA13_KRSSEG_OBER : share of upper class cars (referred to the county average) - PLZ8

270.KBA13_HERST_SONST : share of other cars within the PLZ8

271.GEBURTSJAHR : year of birth

272.SEMIO_SOZ : affinity indicating in what way the person is social minded

273.CJT_TYP_3: not described

274.VHN: not described

275.D19_GARTEN: not described

276.D19_TECHNIK: not described

277.CJT_TYP_5: not described

278.D19_VERSICHERUNGEN: not described

279.D19_BEKLEIDUNG_REST: not described

280.MOBI_RASTER: not described

281.D19_GARTEN_RZ: not described

282.D19_KINDERARTIKEL: not described

283.D19_REISEN_RZ: not described

284.D19_BANKEN_LOKAL: not described

285. UMFELD_JUNG: not described

286. D19_BUCH_CD: not described

287. KONSUMZELLE: not described

288. D19_SAMMELARTIKEL_RZ: not described

289. D19_RATGEBER: not described

290. KBA13_ANTG3: not described

291. VK_ZG11: not described

292. KBA13_BAUMAX: not described

293. D19_HANDWERK: not described

294. VK_DHT4A: not described

295. CUSTOMER_GROUP: not described

296. D19_KK_KUNDENTYP: not described

297. AKT_DAT_KL: not described

298. BIP_FLAG: not described

299. ANZ_KINDER: not described

300. CJT_TYP_1: not described

301. SOHO_FLAG: not described

302. RT_SCHNAEPPCHEN: not described

303. D19_TECHNIK_RZ: not described

304. KBA13_ANTG4: not described

305. D19_SCHUHE_RZ: not described

306. KBA13_CCM_1400_2500: not described

307. D19_BILDUNG: not described

308. D19_REISEN: not described

309. D19_BIO_OEKO: not described

310. HH_DELTA_FLAG: not described

311. CJT_KATALOGNUTZER: not described

312. ALTER_KIND2: not described

313. D19_RATGEBER_RZ: not described

314. D19_LETZTER_KAUF_BRANCHE: not described

315. D19_KONSUMTYP_MAX: not described

316. D19_FREIZEIT: not described

317. CAMEO_DEUINTL_2015: not described

318. KOMBIALTER: not described

319. D19_TELKO_REST: not described

320. D19_WEIN_FEINKOST: not described

321. D19_SOZIALES: not described

322. D19_BANKEN_REST: not described

323. VK_DISTANZ: not described

324. D19_KINDERARTIKEL_RZ: not described

325. D19_VOLLSORTIMENT_RZ: not described

326. D19_TIERARTIKEL: not described

327. CJT_TYP_2: not described

328. D19_KOSMETIK_RZ: not described

329. D19_FREIZEIT_RZ: not described

330. CAMEO_INTL_2015: not described

331.DSL_FLAG: not described

332.D19_LEBENSMITTEL_RZ: not described

333.D19_BANKEN_DIREKT: not described

334.D19_ENERGIE_RZ: not described

335.ALTER_KIND4: not described

336.KBA13_GBZ: not described

337.UNGLEICHENN_FLAG: not described

338.CJT_TYP_6: not described

339.D19_BANKEN_GROSS_RZ: not described

340.D19_VERSAND_REST: not described

341.D19_HAUS_DEKO_RZ: not described

342.D19_VERSI_OFFLINE_DATUM: not described

343.D19_KOSMETIK: not described

344.D19_LEBENSMITTEL: not described

345.D19_VERSI_ONLINE_QUOTE_12: not described

346.KBA13_KMH_210: not described

347.SOHO_KZ: not described

348.D19_TELKO_REST_RZ: not described

349.GEMEINDETYP: not described

350.D19_DIGIT_SERV: not described

351.D19_TELKO_ONLINE_QUOTE_12: not described

352.D19_LOTTO: not described

353.RT_UEBERGROESSE: not described

354. D19_VERSICHERUNGEN_RZ: not described

355. D19_HANDWERK_RZ: not described

356. D19_BANKEN_REST_RZ: not described

357. EXTSEL992: not described

358. D19_BEKLEIDUNG_GEH_RZ: not described

359. RT_KEIN_ANREIZ: not described

360. VHA: not described

361. KBA13_CCM_1401_2500: not described

362. KK_KUNDENTYP: not described

363. KBA13_ANTG2: not described

364. D19_BILDUNG_RZ: not described

365. D19_BEKLEIDUNG_GEH: not described

366. D19_SCHUHE: not described

367. D19_BUCH_RZ: not described

368. STRUKTURTYP: not described

369. ALTER_KIND1: not described

370. D19_VOLLSORTIMENT: not described

371. ALTER_KIND3: not described

372. UMFELD_ALT: not described

373. D19_LOTTO_RZ: not described

374. VERDICHTUNGSRAUM: not described

375. WACHSTUMSGEBIET_NB: not described

376. FIRMENDICHTE: not described

377. KBA13_ANTG1: not described

378. D19_NAHRUNGSERGAENZUNG: not described

379. D19_HAUS_DEKO: not described

380. HAUSHALTSSTRUKTUR: not described

381. D19_NAHRUNGSERGAENZUNG_RZ: not described

382. D19_VERSI_ONLINE_DATUM: not described

383. EINGEZOGENAM_HH_JAHR: not described

384. ONLINE_PURCHASE: not described

385. PRODUCT_GROUP: not described

386. ANZ_STATISTISCHE_HAUSHALTE: not described

387. D19_TIERARTIKEL_RZ: not described

388. EINGEFUEGT_AM: not described

389. D19_BEKLEIDUNG_REST_RZ: not described

390. D19_TELKO_MOBILE: not described

391. D19_SONSTIGE: not described

392. CJT_TYP_4: not described

393. D19_DIGIT_SERV_RZ: not described

394. D19_BANKEN_LOKAL_RZ: not described

395. ALTERSKATEGORIE_FEIN: not described

396. D19_DROGERIEARTIKEL: not described

397. KBA13_HHZ: not described

398. D19_WEIN_FEINKOST_RZ: not described

399. GEOSCORE_KLS7: not described

400. D19_BANKEN_DIREKT_RZ: not described

401. D19_BANKEN_GROSS: not described

402. D19_SONSTIGE_RZ: not described

403. D19_TELKO_MOBILE_RZ: not described

404. D19_SAMMELARTIKEL: not described

405. ARBEIT: not described

406. D19_ENERGIE: not described

407. D19_VERSAND_REST_RZ: not described

408. D19_DROGERIEARTIKEL_RZ: not described

409. D19_VERSI_DATUM: not described

410. D19_BIO_OEKO_RZ: not described

The dataset provided is splitted into two part:

1 - The first part ***Customer Segmentation Report*** is composed of two csv file (unsupervised learning):

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns):
-

Sample

```
      LNR  AGER_TYP  AKT_DAT_KL  ALTER_HH  ALTER_KIND1  ALTER_KIND2  \
0  910215        -1         NaN       NaN          NaN          NaN
1  910220        -1         9.0       0.0          NaN          NaN
2  910225        -1         9.0      17.0          NaN          NaN
3  910226         2         1.0      13.0          NaN          NaN
4  910241        -1         1.0      20.0          NaN          NaN
```

Descriptive Stats

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	\...
count	8.912210e+05	891221.000000	817722.000000	817722.000000	
mean	6.372630e+05	-0.358435	4.421928	10.864126	
std	2.572735e+05	1.198724	3.638805	7.639683	
min	1.916530e+05	-1.000000	1.000000	0.000000	
25%	4.144580e+05	-1.000000	1.000000	0.000000	
50%	6.372630e+05	-1.000000	3.000000	13.000000	
75%	8.600680e+05	-1.000000	9.000000	17.000000	
max	1.082873e+06	3.000000	9.000000	21.000000	

- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns)

Sample

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	\
0	9626	2	1.0	10.0	NaN	NaN	
1	9628	-1	9.0	11.0	NaN	NaN	
2	143872	-1	1.0	6.0	NaN	NaN	
3	143873	1	1.0	8.0	NaN	NaN	
4	143874	-1	1.0	20.0	NaN	NaN	

Descriptive Stats

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	\...
count	191652.000000	191652.000000	145056.000000	145056.000000	
mean	95826.500000	0.344359	1.747525	11.352009	
std	55325.311233	1.391672	1.966334	6.275026	
min	1.000000	-1.000000	1.000000	0.000000	
25%	47913.750000	-1.000000	1.000000	8.000000	
50%	95826.500000	0.000000	1.000000	11.000000	
75%	143739.250000	2.000000	1.000000	16.000000	
max	191652.000000	3.000000	9.000000	21.000000	

The general population dataset (AZDIAS) will be used to create our unsupervised model (PCA and K-means). Then customers dataset will be mapped into both models in order to identify patterns and relation between customers groups.

2 - The second part ***Supervised Learning Model*** is composed of two csv file:

- Udacity_MAILOUT_052018_TRAIN.csv: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns)

Sample

```
   LNR      AGER_TYP  AKT_DAT_KL  ALTER_HH  ALTER_KIND1  ALTER_KIND2  \
0  1763            2         1.0       8.0          NaN          NaN
1  1771            1         4.0      13.0          NaN          NaN
2  1776            1         1.0       9.0          NaN          NaN
3  1460            2         1.0       6.0          NaN          NaN
4  1783            2         1.0       9.0          NaN          NaN
```

Descriptive Stats

```
                LNR      AGER_TYP    AKT_DAT_KL      ALTER_HH  ALTER_KIND1  \
count  42962.000000  42962.000000  35993.000000  35993.000000  1988.000000
mean   42803.120129      0.542922      1.525241     10.285556    12.606137
std    24778.339984      1.412924      1.741500      6.082610     3.924976
min        1.000000     -1.000000      1.000000      0.000000     2.000000
25%    21284.250000     -1.000000      1.000000      8.000000     9.000000
50%    42710.000000      1.000000      1.000000     10.000000    13.000000
75%    64340.500000      2.000000      1.000000     15.000000    16.000000
max    85795.000000      3.000000      9.000000     21.000000    18.000000
```

- Udacity_MAILOUT_052018_TEST.csv: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

Sample

```
                LNR      AGER_TYP    AKT_DAT_KL      ALTER_HH  ALTER_KIND1  \
count  42833.000000  42833.000000  35944.000000  35944.000000  2013.000000
mean   42993.165620      0.537436      1.518890     10.239511    12.534029
std    24755.599728      1.414777      1.737441      6.109680     3.996079
min        2.000000     -1.000000      1.000000      0.000000     2.000000
25%    21650.000000     -1.000000      1.000000      8.000000     9.000000
50%    43054.000000      1.000000      1.000000     10.000000    13.000000
75%    64352.000000      2.000000      1.000000     15.000000    16.000000
max    85794.000000      3.000000      9.000000     21.000000    18.000000
```

Descriptive Stats

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	\
0	1754	2	1.0	7.0	NaN	NaN	
1	1770	-1	1.0	0.0	NaN	NaN	
2	1465	2	9.0	16.0	NaN	NaN	
3	1470	-1	7.0	0.0	NaN	NaN	
4	1478	1	1.0	21.0	NaN	NaN	

We can see a lot of missing or non value datas that should be cleaned and preprocessed before implementing unsupervised and supervised models. All datasets described above should be treated in the same way. That means the final cleaned and preprocessed datasets should have the same columns length.

Cleaning and Preprocessing

Once our data is described and some of our features processed we are now ready to treat missing values. Arvato Financial Services has provided a real life dataset of demographics characteristics of customers for a general population located in Germany. Cleaning the data is deciding how to treat the missing values, for example if we drop or replace it. This is a key step which will influence our model performance in the next steps. In order to get a cleaned data, we made use of different preprocessing techniques to have a flawless data set. Data cleaning methods attempt to fill in missing values, smooth out noise, and correct inconsistencies in the data. In the last step we will normalize our data to get data values between 0 to 1 and also remove outliers.

It is important to note that all cleaning and preprocessing steps are applied on AZDIAS,Customers, Training and Test datasets in the same way. We will discuss about all steps later in Methodology section.

Exploratory Visualization

After cleaning and normalizing the data we'll use visualization tools such as histogram and correlation matrix heatmap to represent distribution and features correlation. These representations will help us to understand how features are distributed and related to each other.

The histogram plots below show the distribution of our dimensionality reduced dataset. Each plot represent a component out of sixteen components we have.

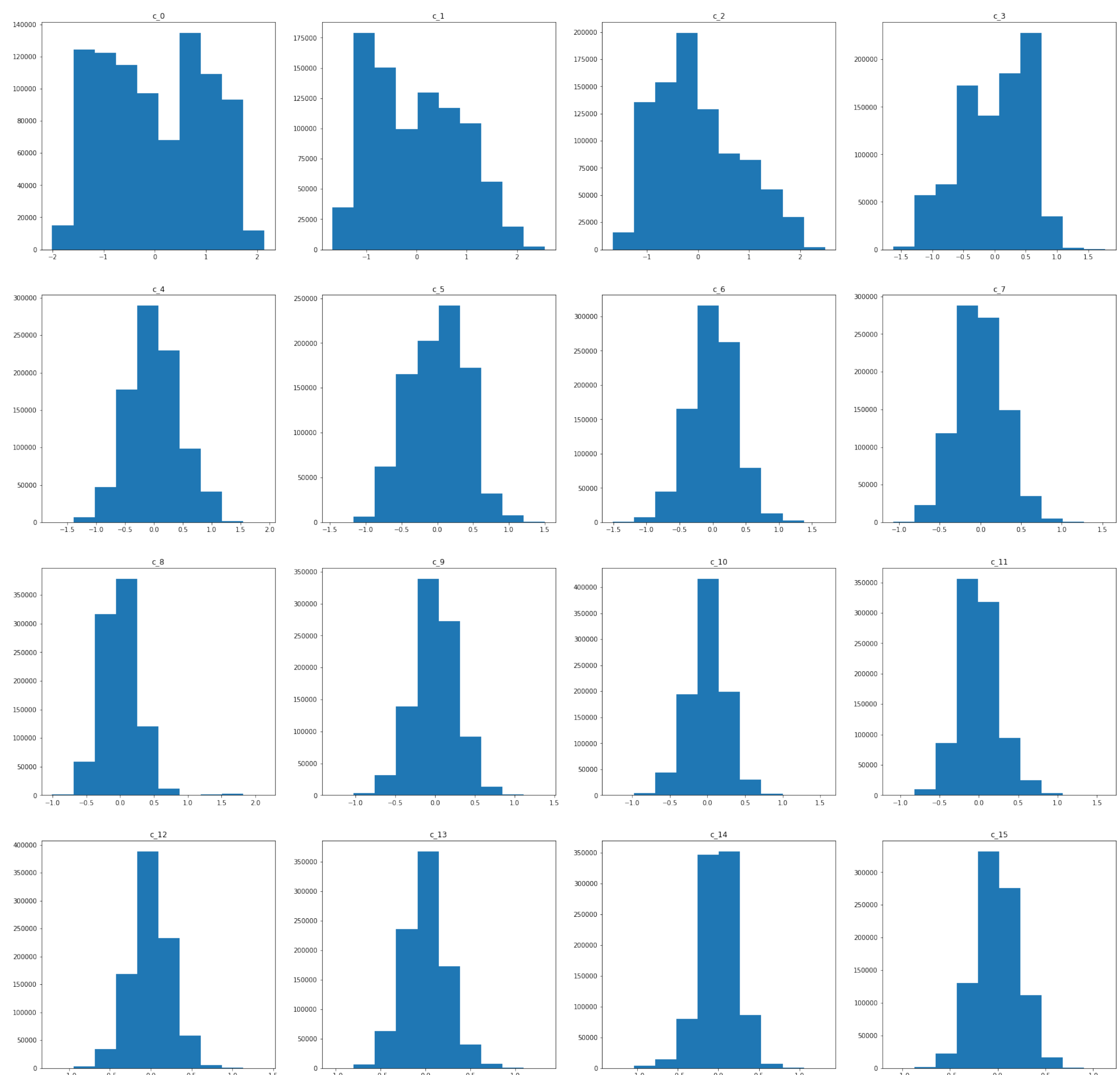


fig.1 – Data distribution

Heatmap is plotted using Seaborn package. Here we have used correlation heatmap to see how our components are connected to each other. However we observe that components are totally independent and it shows that our dimensionality reduction has played an important role in simplification of the data.

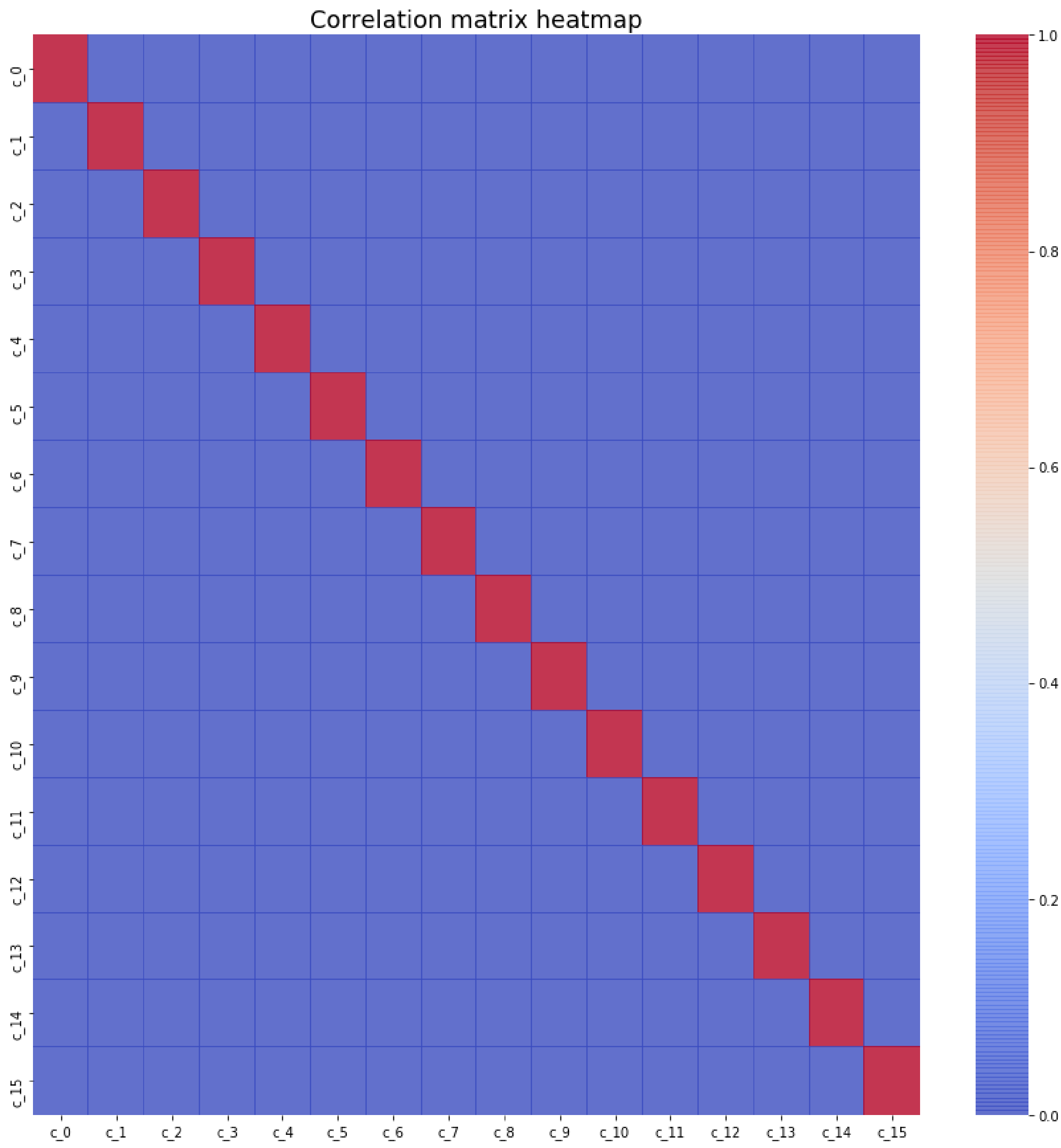


fig.2 – Feature Correlation Matrix

Algorithms and Techniques

One of the most important part of a machine learning work flow is the algorithms that we choose to get an output from our model. For each step of this project we'll implement different models. **Principal Components Analysis** for dimensionality reduction, **K-means clustering** used to group customers in clusters, **LinearLearner** for binary classification and **Logistic Regression** used as a benchmark, are well known algorithms used in modern case studies that we'll discuss about them here below.

1. Principal Components Analysis (PCA)

Principal Component Analysis (PCA) is a technique which uses sophisticated mathematical principles to transforms a number of possibly correlated variables into a smaller number of variables called principal components. The origins of PCA lie in multivariate data analysis. One of the most important and perhaps its most common use is as to reduce dimensionality of large data sets. The large size of our datasets and features would be difficult to use through creating a clustering model. To prevent this, we'll use Principal Components Algorithms (PCA) to reduce the dimensionality of our preprocessed data. Our model is trained with azdias dataset, thereafter we choose the n_components (the number of components we want to keep) to retain regarding the explained variance (visualizing with elbow graph in figure 3). Once our azdias PCA model is ready we map the customers dataset into the azdias pca model. the PCA model used is imported from Scikitlearn package.

The following graph represent the number of components vs the explained variance. The goal is to reduce dimensionality as much as possible and capture minimum 80% of the explained variance.

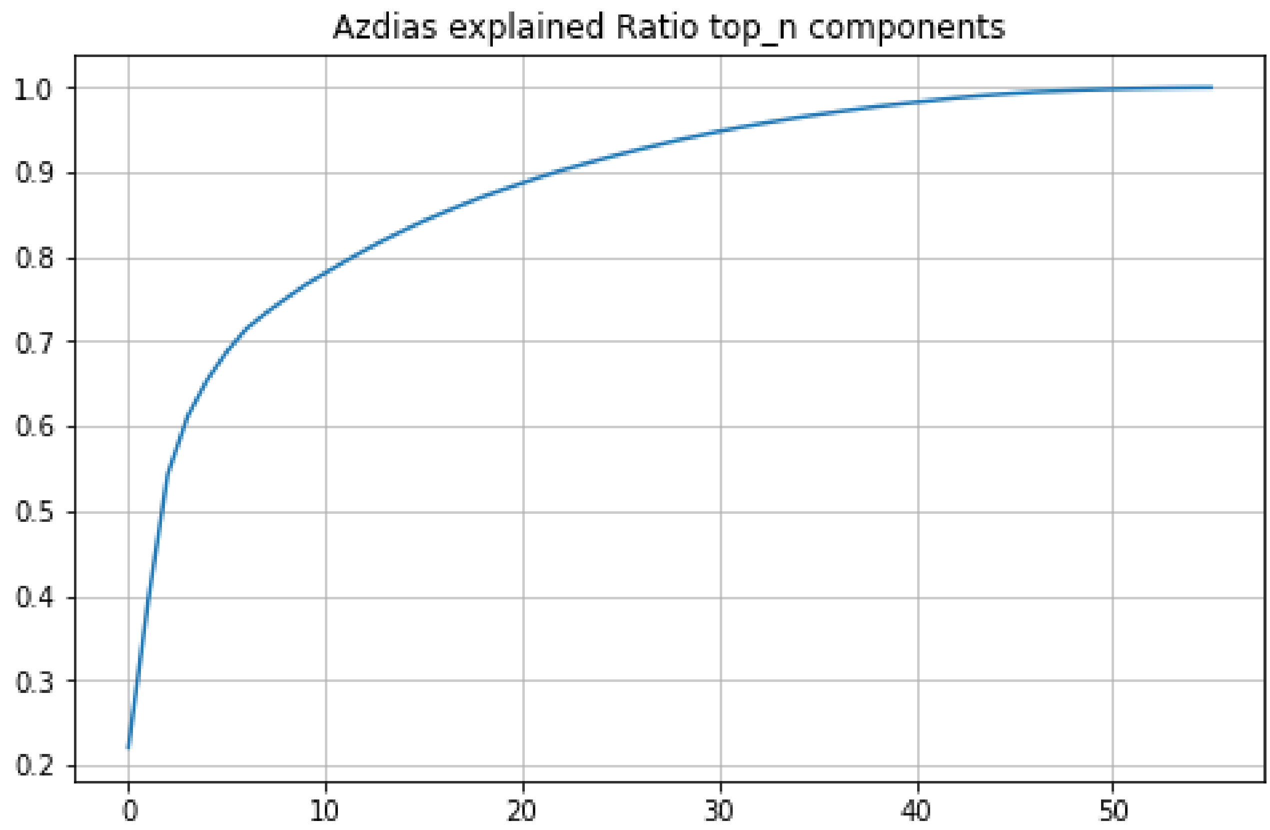


fig.3 – PCA components vs explained variance

In this context we chose n_components = 16 for an explained variance of 84.21%. Once we have our model ready the customers dataset is mapped into the azdias pca model and tranformed and reduced to 16 components.

2. K-means Clustering

In this project, we'll use the unsupervised clustering algorithm, k-means, to segment general population using their PCA components, which are in the transformed DataFrame we just created. Then we will use this model to create segments for our customers dataset. The goal is to identify groups of individuals that have similar demographic characteristics and then relate this analysis to our goal which is to identify individuals which are more likely to respond to the mail-order list marketing campaign.

K-means create clusters, regardless of the actual existence of any structure in the data. When using K-means clustering, we are making an hypothesis of some structure among the objects. We should note that just because clusters can be found does not validate their existence. Only with strong conceptual support and good visualization clusters could potentially be meaningful and relevant. In the following parts we describe how we choose the k value and then how we create our model. In this context we used different visualization tools to represent clusters and components correlation.

However the first step in this section will be to determine the optimal k value. For this purpose we create a plot representing k value vs the sum of squared distance to clusters centers.

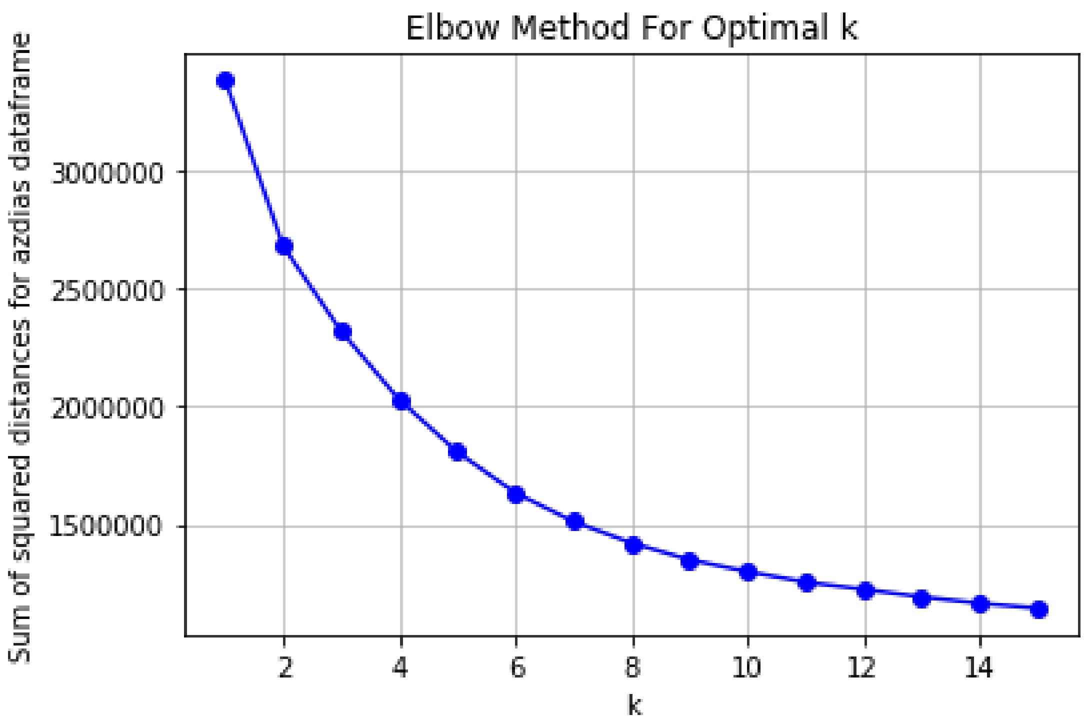


fig.4 – Optimal k elbow graph

In the plot the optimal k for this dataset is between 4 and 5. we chose k=4 for our model.

In the **Supervised Learning** section we will use this model for clustering of our reduced supervised dataset and try to regroup customers that are labeled 1 inside a specific cluster.

3. Linear Learner (SageMaker)

Linear Learner is Amazon SageMaker builtin supervised learning algorithm used for solving either classification or regression problems. Here we give a high dimension multiclass dataframe as input and the output is a binary one dimensional dataframe where the label must be either 0 or 1. This algorithm allows us to explore a large number of models and choose the best, which optimizes either continuous objectives such as *mean square error*, *cross entropy loss*, *absolute error*, etc., or discrete objectives suited for classification such as *F1* measure, *precision@recall*, *accuracy*. However, in this project we will focus on optimizing recall and compensating imbalance using defined parameters implemented in this algorithm.

4. Logistic Regression

Logisitc Regression is a statistical method used in machine learning and implemented here as a binary classifier. Here we will use this algorithm to explain the relationship between multiclass variables dataset and our binary label described as 0 for people predicted not to be in mail-order list and 1 for people predicted be in the mail-order list.

III. Methodology

Data Preprocessing

The dataset provided needs to be cleaned and preprocessed before feeding to PCA model.
To make it possible we proceed as follows:

1. Identifying and dropping non values columns
2. Convert remaining non values data to numeric values

3. Replacing unknown and none data with NaN
4. Analyzing NaN values in dataset
5. Dropping columns with more than 20% of NaN values for azidas
6. Replacing remaining NaNs with -1
7. Checking data type and cleaned values
8. Normalizing the data

Implementation

The implementation process can be split into three main stages:

1. Dimensionality reduction with PCA
2. K-means Clustering
3. Linear Learner (Sage Maker)

1. Dimensionality reduction with PCA

During the first stage, the PCA was trained on the preprocessed data. This was done in a Jupyter notebook on Udacity work space "Bertelsmann/Arvato Project Workspace" (titled "Arvato Project Workbook.ipynb"), and can be further divided into the following steps:

1. Explore cleaned data attributes
2. Creating PCA model
3. Data Variance
4. Data variance vs dimensionality reduction
5. Component Makeup
6. Components Histogram
7. Correlation matrix heatmap

2. K-means clustering

Now we'll ready to implement our k-means model. Before training we have to determine the k value, then we'll use the reduced dimension data to train our k-mean model.

This section is break out to the following steps:

1. Determining the optimal number of clusters for k-means clustering
2. Creating K-means model
3. Predicting customers labels
4. Visualization
5. Natural groupings

3. Linear Learner (Sage Maker)

We'll have access to a third dataset with attributes from targets of a mail order campaign. We'll use the previous analysis to build a Linear Learner model that predicts whether or not each individual will respond to the campaign.
To build our supervised model using Amazon SageMaker we will divide this section to the following steps:

1. Load preprocessed Data from S3
2. Splitting the data
3. Imbalanced training data
4. Create a LinearLearner Estimator
5. Convert data into a RecordSet format
6. Evaluating Model

4. Benchmark

Once our Linear Learner model is implemented and its performance is measured, it's time to create our benchmark model. The goal is to compare our binary classifier and its metrics to our benchmark model. In this context we will use Logisitc Regression provided by Scikitlearn to create our model and see if it outperforms our Linear Learner model or not.

1. Preparing the data
2. Model Development and Prediction
3. Metrics
4. Oversample minority class
5. Compare to our model

IV. Results

Model Evaluation and Validation

The LinearLearner is used as a binary classifier. This SageMaker algorithm allows us to focus on the minority class accuracy trying to maximize True Positives and minimize False Negative (Recall). Moreover the Amazon SageMaker platform allows us to deploy the model and create an API in order to put the model in production environment. However instead tuning hyperparameters the result is not satisfying:

```
LinearLearner(role=role,
              train_instance_count=1,
              train_instance_type='ml.c4.xlarge',
              predictor_type='binary_classifier',
              output_path=output_path,
              sagemaker_session=sagemaker_session,
              epochs=20,
              binary_classifier_model_selection_criteria=
              'precision_at_target_recall',
              target_precision=0.8,
              positive_example_weight_mult='balanced')
```

```
prediction (col)   0.0   1.0

actual (row)
0.0               4247  6354
1.0                 23   117
```

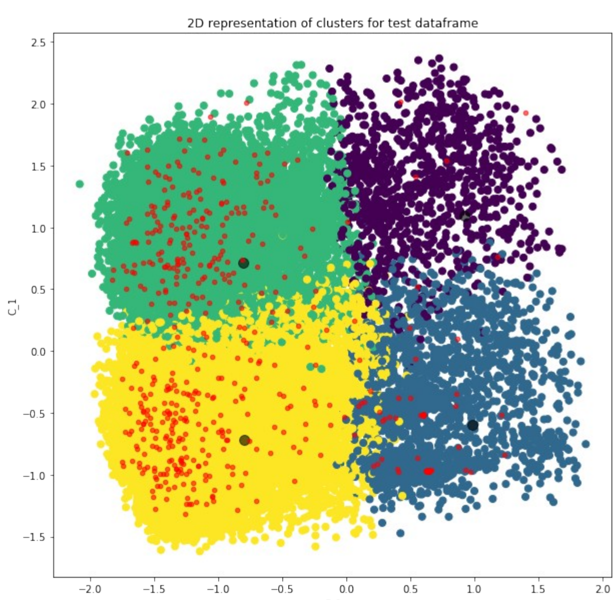
```
Recall:     0.836
Precision:  0.018
Accuracy:   0.406
f1_score:   0.035
```

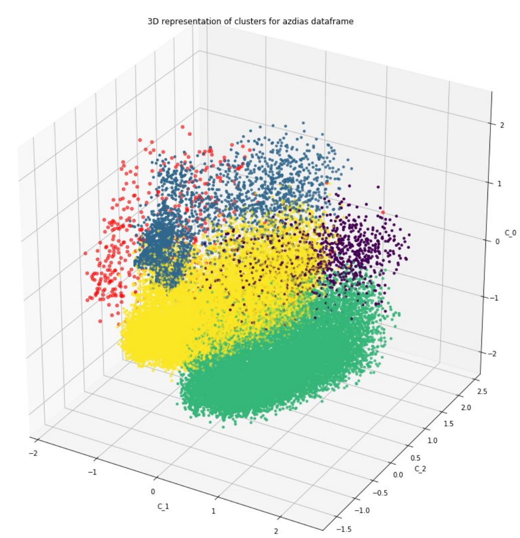
Justification

Both models created previously K-means and LinearLearner have been tested using Test dataset (Udacity_MAILOUT_052018_TEST.csv). The K-means clustering model shows good generalizing qualities with unseen data. In the following scatter plots we have represented the clusters and also the positive labeled data of training set in red color. We can observe that more than 80% of positive labeled data are regrouped in cluster 2 and 3.

```
concat_train

# Cluster count for label 1
concat_train[concat_train['RESPONSE']==1]['Cluster'].value_counts()
3    237
2    193
1     95
0      7
Name: Cluster, dtype: int64
```

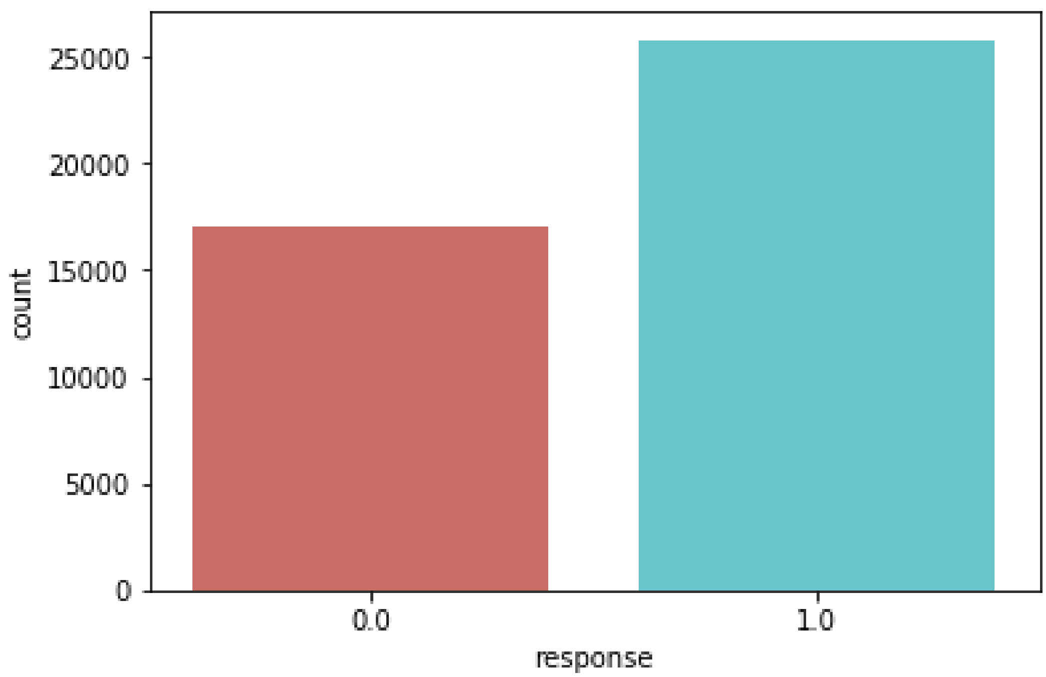




However the Linear model shows low performance as we can see here below:

```
1.0    25811

0.0    17022
dtype: int64
```



The results as we can see are not correlated with the trained model.

Finally we compare our model to Benchmark. Here below we are comparing results between two models:

	Recall	Precision	Accuracy	F1
Linear Learner	0.836	0.018	0.406	0.035
Logisitic Regression	0.671	0.021	0.589	0.041

V. Conclusion

Free-Form Visualization

Visualization is an important tool in data analysis and machine learning which helps us to get an overview of the data and also to see the improvements that we made during the project. In the following figures below we can observe how the azdias PCA model has simplify the data (figure 5 and 6). In the following steps, in the other figure (figure 7) we have represented a 2D and 3D plots of clusters made it by azdias K-means model where we can differentiate different clusters.

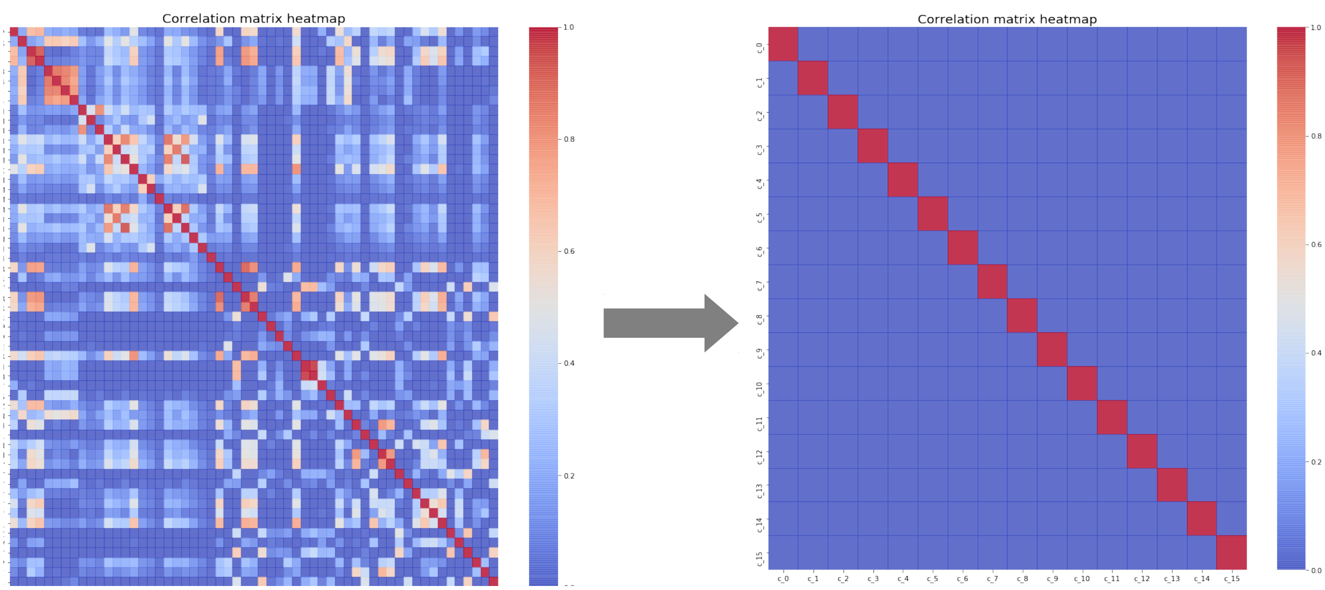


fig.5 – Correlation Heatmap: before and after PCA

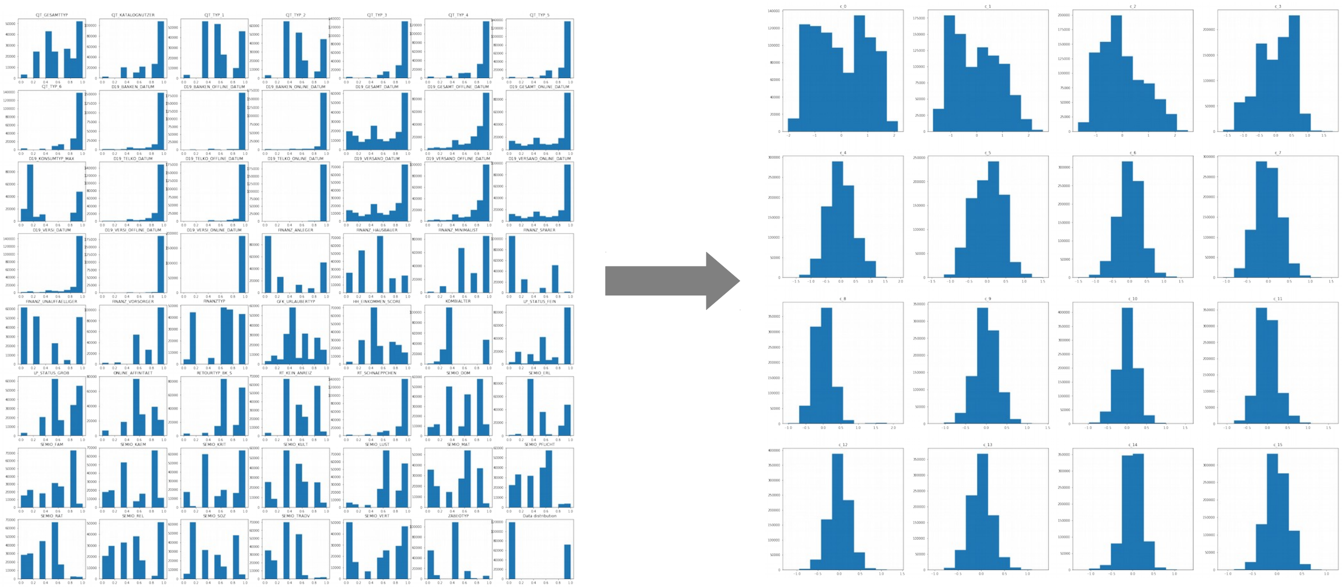


fig.6 – Features Histogram: before and after PCA

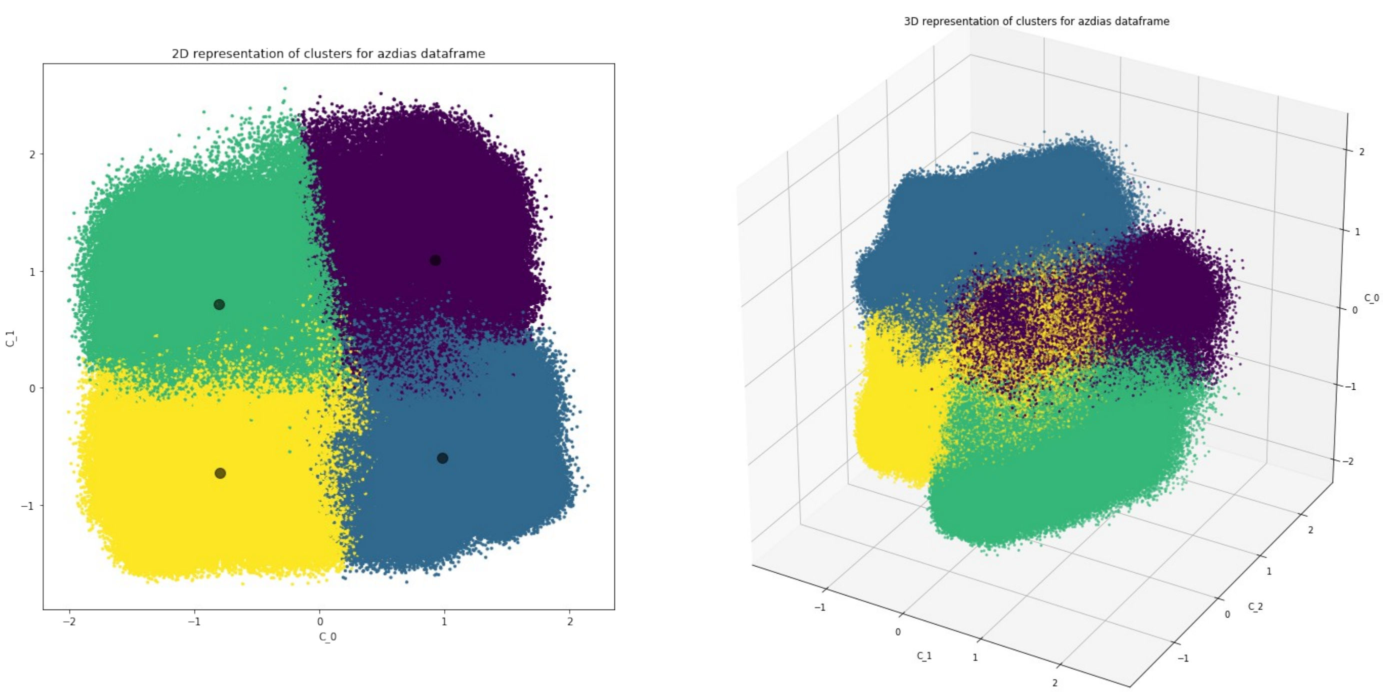


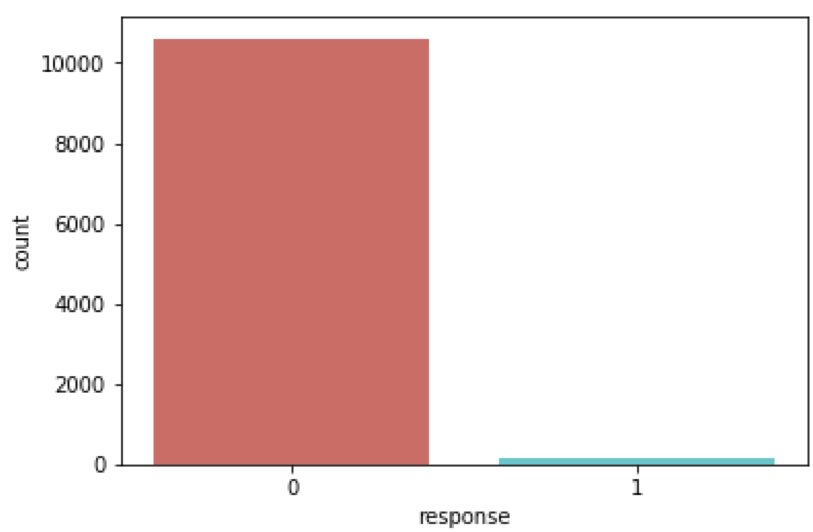
fig.7 – 2D and 3D scatter plot of K-means clusters

Refinement

In the previous sections we discussed about how we have implemented our classifier and the outputs. In this section we will describe how we have improved our model to reach out the final model. The LinearLearner model parameters was first implemented as follows:

```
LinearLearner(role=role,
              train_instance_count=1,
              train_instance_type='ml.c4.xlarge',
              predictor_type='binary_classifier',
              output_path=output_path,
              sagemaker_session=sagemaker_session,
              epochs=20
              )
```

This model has been trained and deployed successfully. The accuracy obtained was about 95%. However the recall shows low results (about 5%). This low performance is caused by the class imbalance observed in the data. The following representation shows the minority class (labeled 1) vs the majority (labeled 0) :



In this project the goal is to create a model capable to predict if a customer is likely to be in the mail list (class 1) or not (class 0). Therefore we will set *binary classifier model selection criteria* parameter to *precision at target recall* in order to maximize True Positives and minimize False Negatives. In the other hand we will manage the imbalance setting *positive example weight mult* to *balanced*. Once these parameters are set the model created will be compatible to our data and the output will be significantly improved (results are presented in section Model Evaluation and Validation).

Reflection

In this project we tried different machine learning techniques and algorithm to implement a robust model capable to predict if a customer has the potential to respond positively to mail-order marketing campaign or not. However after evaluation we can observe that our models are not performing well and some improvement has to be made. In the other hand we have not be able to establish a relation between our clustering model and the supervised data. In this context, it will be interesting to test other algorithms such as Convolutional Neural Network (CNN) and try to tune this model with different architectures.

REFERENCES

Udacity Machine Learning Case Study (2017). Fraud Detection
Shruti Saxena (2018). Precision Vs Recall, *TowardsDataScience.com*
Franklin, Joel N. (1968). Matrix Theory. Dover Publications.
Siotani, Minoru (1964). Tolerance regions for a multivariate normal population
Czar Robero (2018). K- means Clustering Tutorial
Andrew Aziz (2017). Customer Segmentation based on Behavioural Data in E-marketplace
Joseph F. Hair, Jr. William C. Bl, Barry J. Babin, Rolph E. Anders (2014). Multivariate Data Analysis