تحلیل داده با پایتون

تألیف:
وحید قربانی

۱۴۰۳

سر شناسه	قربانی ، وحید، ۱۳۶۱.
عنوان و پدیدآور	تحلیل داده با پایتون / تألیف: وحید قربانی.
مشخصات نشر	تهران: نشر نور علم،۱۴۰۳.
مشخصات ظاهری	۳۴۸ ص.: مصور (بخشی رنگی)، جدول، نمودار (رنگی).
شابک	۹-۶۲۹-۱۶۹-۶۰۰-۹۷۸
موضوع	داده‌کاوی - Data mining
	ساختار داده‌ها - Data structures (Computer science)
	پایتون (زبان برنامه‌نویسی کامپیوتر) - Python (Computer program language)
رده بندی کنگره	QA ۷۶/۸
رده‌بندی دیویی	۰۰۶/۳۱۲

نشر نورعلم و پخش قلم سینا: تهران انقلاب خ۱۲فروردین پلاک۲۸۶ تلفن۶۶۴۰۵۸۸۰ و ۶۶۹۵۷۰۲۱

۰۹۱۲۳۳۳۴۲۲۹-۰۹۱۲۲۰۷۹۸۴۹ - ۰۹۱۲۳۴۶۲۸۱۱

وب سایت https://www.modiranketab.ir پیج اینستاگرام - @modiranketab

تحلیل داده با پایتون

تألیف: وحید قربانی

ناشر: نور علم

شابک: ۹-۶۲۹-۱۶۹-۶۰۰-۹۷۸

چاپ و صحافی: سورنا

نوبت چاپ: اول ۱۴۰۳

شمارگان:۲۰۰ جلد

قیمت: ۲۹۵۰۰۰ تومان

پیشگفتار

تحلیل داده در دنیای امروز از اهمیت بسیاری برخوردار است. تحلیل داده‌ها به افراد و سازمان‌ها کمک می‌کند تا تصمیم‌های بهتری بگیرند. در داده‌کاوی و به طور کلی در علم داده از یک متدولوژی به نام CRISP استفاده می‌شود. CRISP یک روش استاندارد برای تحلیل داده است که در فرآیند استخراج اطلاعات از داده‌ها و مدل‌سازی استفاده می‌شود. این روش از شش مرحله کلیدی تشکیل شده است که به ترتیب شامل شناخت کسب‌وکار، فهم داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و نهایتاً استقرار می‌شود.

به طور کلی، روش CRISP به سازمان‌ها در افزایش بهره‌وری از داده‌ها، افزایش قابلیت پیش‌بینی، بهبود فرایند تصمیم‌گیری، کاهش ریسک و هزینه‌ها و افزایش رقابت‌پذیری کمک می‌کند.

با تحلیل داده‌ها و استفاده از روش CRISP می‌توان روندها و الگوهای موجود در داده‌ها را شناسایی کرده و رویدادهای آینده را پیش‌بینی کرد. این امر به سازمان‌ها کمک می‌کند تا استراتژی‌ها و تصمیمات خود را بر اساس اطلاعات قابل اعتماد و مبتنی بر داده و پیش‌بینی‌های دقیق انجام داده و تصمیم‌های بهتری درباره مسائل مختلف مانند بازاریابی، مدیریت مشتریان، بهبود عملکرد و... بگیرند.

با تحلیل داده‌ها و شناسایی الگوها، سازمان‌ها می‌توانند از تصمیماتی که ممکن است باعث خطاها و ضررهای مالی شوند، جلوگیری نموده و از این طریق ریسک‌های مرتبط با تصمیمات نادرست را کاهش داده و در طولانی‌مدت بهبود عملکرد سازمان را تضمین نمایند. تحلیل داده به سازمان‌ها کمک می‌کند تا بازار رقابتی را بهتر بشناسند و استراتژی‌هایی را برای افزایش بازدهی، بهبود رقابت‌پذیری و جایگاه خود در بازار انتخاب کنند. هم چنین به سازمان‌ها این امکان را می‌دهد تا از داده‌های بزرگ و پیچیده بهره‌برداری کنند، روابط مخفی را شناسایی و اطلاعات ارزشمندی را استخراج نمایند که همین امر موجب اتخاذ تصمیمات استراتژیک و اجرای بهتر سیاست‌ها خواهد شد.

زبان برنامه‌نویسی پایتون به عنوان یکی از زبان‌های برنامه‌نویسی محبوب و قدرتمند در زمینه علوم داده شناخته می‌شود. به دلایل زیادی از جمله سادگی، فهم و خوانایی بالا، دارا بودن بستر وسیع، داشتن کتابخانه‌ها و ابزارهای قدرتمندی که به طور خاص برای تحلیل داده‌ها و علوم داده

طراحی شده و توسعه یافته‌اند، جامعیت و قابل اجرا بودن بر روی اکثر سیستم عامل‌ها، وجود منابع آموزشی غنی و پشتیبانی از برنامه‌نویسان و توسعه‌دهندگان و سازگاری و هماهنگی با سایر ابزارهای مورد استفاده در حوزه علوم داده آن را به یک ابزار بسیار مناسب برای تحلیل داده تبدیل نموده که در حال حاضر توسط افراد و سازمان‌های زیادی مورد استفاده قرار می‌گیرد.

با مطالعه این کتاب خواهید آموخت که چگونه با به‌کارگیری زبان برنامه‌نویسی قدرتمند پایتون و با استفاده از چهار کتابخانه پرکاربرد آن یعنی NumPy، Pandas، Matplotlib و Seaborn، به شناخت و فهمی از داده‌ها برسید، آن‌ها را پاک‌سازی، تبدیل، بصری‌سازی، و تحلیل نمایید.

در طول این کتاب، با مفاهیم اصلی تحلیل داده آشنا خواهید شد و با استفاده از متدها و پارامترهای موجود در این چهار کتابخانه، می‌توانید داده‌های خود را تفسیر کنید و نتایج دقیق و قابل اعتمادی را به دست آورید. هم‌چنین، نمونه‌های عملی و تمرین‌هایی ارائه شده است تا بتوانید مهارت‌های خود را تقویت نموده و در طول مسیر یادگیری، تجربه‌های جدیدی کسب نمایید.

کتاب پیش‌رو شامل مطالبی درباره متدولوژی CRISP در علم داده، کتابخانه NumPy، کتابخانه Pandas، کتابخانه Matplotlib و کتابخانه Seaborn است.

در مقدمه گام‌های متد CRISP به تفصیل آمده است. در فصل اول کتاب، به کتابخانه NumPy پرداخته شده است که امکان انجام عملیات جبری بر روی آرایه‌ها، توابع ریاضی و عملیات ضرب ماتریس‌ها را فراهم می‌کند.

در فصل دوم، کتابخانه Pandas معرفی می‌گردد که برای شناخت داده‌ها، آماده سازی داده‌ها و عملیات مختلف بر روی داده‌ها از این کتابخانه استفاده می‌شود. شما با ساختارهای داده مانند سریز و دیتافریم در Pandas آشنا می‌شوید و یاد می‌گیرید چگونه داده‌ها را از منابع مختلف مانند SQL و فایل‌های اکسل بخوانید و با آن‌ها کار کنید.

در فصل سوم، کتابخانه Matplotlib معرفی می‌شود که برای ترسیم و نمایش داده‌ها در قالب نمودارها از آن استفاده می‌شود. شما با انواع نمودارها مانند histplot، countplot و scatterplot آشنا می‌شوید و یاد می‌گیرید چگونه داده‌ها را در قالب پلات نشان دهید و شکل و رنگ پلات‌ها را تنظیم نمایید. در فصل چهارم، کتابخانه Seaborn معرفی می‌شود که برای ترسیم داده‌ها بر

پایه کتابخانه Matplotlib است. این کتابخانه امکان ترسیم انواع نمودارها مانند rugplot، displot و heatmap را فراهم می‌کند.

امیدوارم این کتاب به عنوان یک راهنمای جامع در مبحث تحلیل داده با پایتون برای خوانندگان، مفید واقع شده و به افزایش دانش و مهارت آنها در این حوزه کمک کند.

وحید قربانی

VahidGhorbani@hotmail.com
http://www.linkedin.com/in/VahidGhorbani
https://www.youtube.com/@DataLand01

فهرست

مقدمه

متدولوژی CRISP[1] در علم داده

می‌خواهیم با متدی به نام CRISP آشنا شویم. ابتدا علم داده را تعریف می‌کنیم.

علم داده[2] چیست؟

در حوزه‌های مختلف علم و یا در کسب و کار سوال‌ها و چالش‌هایی به وجود می‌آید که خبرگان و صاحبان آن حوزه و کسب و کار نمی‌توانند به آن پاسخ دهند. علم داده با استفاده از داده‌های موجود در همان حوزه علمی و کسب و کار به این سوال‌ها پاسخ می‌دهد.

برای اینکه پروژه‌های علم داده نظم پیدا کند از یک متدولوژی به نام CRISP استفاده می‌کنیم. برای پروژه‌های داده‌کاوی نیز از همین روش استفاده می‌نماییم.

گام‌ها در متد CRISP

گام اول: شناخت کسب و کار[3]: در گام اول لازم است فهم و درک درستی نسبت به آن حوزه و کسب و کار پیدا کنیم. از آن‌جایی که کسب و کارها پیچیدگی‌های زیادی دارند معمولا شناخت درست و دقیق آن‌ها مستلزم صرف زمان زیادی است. برای حل این مشکل می‌توانیم از خبرگان و صاحبان آن حوزه و کسب و کار کمک بگیریم. این افراد به عنوان مشاور[4] در کنار دانشمند داده[5] حضور دارند و به او کمک می‌کنند تا بتواند مسائل را درست متوجه شود، مسائل درستی را مطرح کند و برای رسیدن به راه حل آن مسائل مسیر درستی نیز انتخاب نماید.

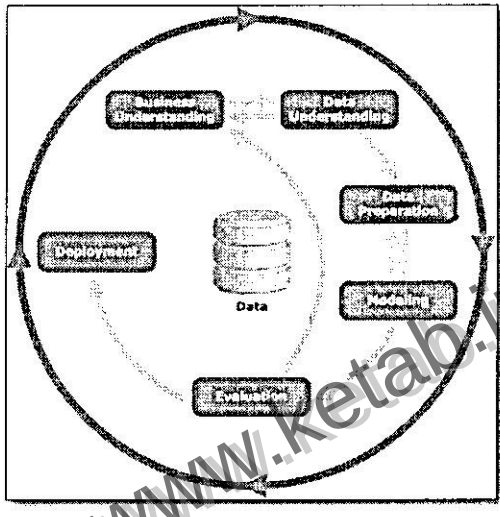
گام دوم: شناخت داده‌ها[6]: در این گام تلاش می‌کنیم تا فهم و درک درستی نسبت به داده‌ها پیدا کنیم. یعنی اگر یک مجموعه داده[7] در اختیار داشته باشیم، لازم است همه ستون‌های آن را بشناسیم. برای مثال در مجموعه داده‌های کشتی تایتانیک ستون Survived یعنی نجات یافتن یا غرق شدن مسافر و ستون Embarked یعنی مسافر در کدام ایستگاه سوار کشتی شده است.

---

[1] The CRoss Industry Standard Process for Data Mining (CRISP-DM)
[2] Data science
[3] Business Understanding
[4] Business Experts
[5] Data Scientist
[6] Data Understanding
[7] Dataset

از طرفی باید نسبت به هر یک از ستون‌ها در مجموعه داده‌ها شناخت داشته باشیم که چه تعداد داده مفقود[1]، داده دور افتاده[2] و Noise Data وجود دارد.

به عبارتی در مرحله شناخت داده‌ها باید یک دید آماری نسبت به کل داده‌ها پیدا کنیم. نوع داده‌ای آن‌ها را بدانیم. داده‌هایی که نیاز به تغییر و تبدیل دارند و داده‌های Valid را شناسایی کنیم.



تصویر ۱-۱. گام‌ها در متدلوژی CRISP

برای اینکه بتوانیم به این شناخت از مجموعه داده‌ها برسیم، بهترین روش، استفاده از کتابخانه‌های پایتون مانند Pandas، Matplotlib و Seaborn است. (کتابخانه Seaborn توسط کتابخانه Matplotlib توسعه پیدا کرده است).

یکی از مواردی که در شناخت داده‌ها به ما کمک می‌کند بصری‌سازی داده‌ها[3] است. اگر دیتا را بصری کنیم می‌توانیم در کمترین زمان ممکن به چندین مورد در شناخت داده برسیم ازجمله نرمال بودن داده‌ها یعنی آیا توزیع نرمال در این داده‌ها وجود دارد؟ با ترسیم یک نمودار به نام

---

[1] Missing data
[2] Outlier data
[3] Data Visualization

Histogram Plot درکتابخانه Matplotlib می‌توانیم به این هدف برسیم. پس برای بصری‌سازی داده‌ها از کتابخانه Matplotlib استفاده می‌کنیم.

برای اینکه نوع داده‌ها را بشناسیم، گزارش‌های آماری در مورد داده‌ها داشته باشیم، داده‌های دور افتاده، داده‌های مفقود، Noise data و داده‌هایی که Valid نیستند را پیدا کنیم از کتابخانه Pandas استفاده می‌کنیم.

ممکن است گاهی یک رفت و برگشت بین شناخت داده‌ها و شناخت کسب و کار انجام شود. برای مثال ممکن است مساله‌ای مطرح کنیم ولی داده‌ای برای بررسی آن مساله وجود نداشته باشد. در چنین شرایطی باید مساله را تغییر دهیم و یا با اقلام اطلاعاتی دیگری سعی کنیم مساله را حل نماییم. پس ممکن است بارها و بارها رفت و برگشت بین داده‌ها و کسب و کار اتفاق بیافتد.

گام سوم: آماده‌سازی داده[1]: در این مرحله داده‌ها را تغییر می‌دهیم. وقتی در مرحله شناخت داده‌ها، برای مثال متوجه می‌شویم چه تعداد داده گمشده داریم، در مرحله آماده‌سازی داده‌ها تصمیم می‌گیریم که به جای داده‌هایی که مقدارشان وجود ندارد چه مقداری قرار دهیم، این داده‌ها را چگونه جایگزین و آماده کنیم. همچنین اگر بخواهیم یک ستون را به دو ستون یا برعکس دو ستون را به یک ستون تبدیل کنیم یا نوع داده‌ای یک ستون را تغییر دهیم مثلا مقادیر رشته‌ای را به عددی تبدیل کنیم همه این موارد را در مرحله آماده‌سازی داده‌ها و با استفاده از کتابخانه Pandas انجام می‌دهیم.

حدود ۶۰ تا ۷۰ درصد وقت یک دانشمند داده صرف دو مرحله شناخت داده و آماده‌سازی داده می‌شود. پس این دو مرحله خیلی مهم هستند.

گام چهارم: مدل‌سازی[2]: در این مرحله از مجموعه الگوریتم‌های یادگیری ماشین[3] مانند Sklearn استفاده می‌کنیم. برای این کار الگوریتم‌های زیادی وجود دارند که بهینه شده و وارد مرحله یادگیری عمیق شده‌اند.

---

[1] Data Preparation
[2] Modeling
[3] Machine Learning

در این مرحله نیز بین آماده‌سازی داده‌ها و مدل‌سازی دادها رفت و برگشت وجود دارد. برای مثال وقتی که داریم الگوریتم را پیاده‌سازی می‌کنیم متوجه می‌شویم یک ویژگی کم داریم، باید برگردیم به مرحله آماده‌سازی داده، داده‌ها را آماده کنیم. یا وقتی مدل در حال ساخته شدن است با پیغام خطا مواجه می‌شویم و متوجه می‌شویم که نوع داده‌ای یک ستون را باید تغییر دهیم. بعضی الگوریتم‌ها مانند الگوریتم خوشه‌بندی[1] مقادیر Text نمی‌پذیرند. هنر یک دانشمند داده انتخاب الگوریتم مناسب برای مدل‌سازی داده و انتخاب پارامترهای مناسب است. مرحله مدل‌سازی محبوب‌ترین مرحله است و حدود ۱۰ تا ۱۵ درصد وقت دانشمند داده را می‌گیرد.

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,jfhsdjfgdshfgds,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38.1,0,PC
17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
8,0,3,"Palsson, Master. Gosta Leonard",male,2,3,1,349909,21.075,,S
9,1,3,"Johnson, Mrs. Oscar W (Elisabeth Vilhelmina
Berg)",female,27,0,2,347742,11.1333,,S
10,1,2,"Nasser, Mrs. Nicholas (Adele Achem)",female,14,1,0,237736,30.0708,,C
11,1,3,"Sandstrom, Miss. Marguerite Rut",female,4,1,1,PP 9549,16.7,G6,S
12,1,1,"Bonnell, Miss. Elizabeth",female,58,0,0,113783,26.55,C103,S
13,0,3,"Saundercock, Mr. William Henry",male,20,0,0,A/5. 2151,8.05,,S
14,0,3,"Andersson, Mr. Anders Johan",male,39,1,5,347082,31.275,,S
15,0,3,"Vestrom, Miss. Hulda Amanda Adolfina",female,14,0,0,350406,7.8542,,S
16,1,2,"Hewlett, Mrs. (Mary D Kingcome) ",female,55,0,0,248706,16,,S
17,0,3,"Rice, Master. Eugene",male,2,4,1,382652,29.125,,Q
18,1,2,"Williams, Mr. Charles Eugene",male,,0,0,244373,13,,S
19,0,3,"Vander Planke, Mrs. Julius (Emelia Maria
Vandemoortele)",female,31,1,0,345763,18,,S
20,1,3,"Masselmani, Mrs. Fatima",female,,0,0,2649,7.225,,C
21,0,2,"Fynney, Mr. Joseph J",male,35,0,0,239865,26,,S
22,1,2,"Beesley, Mr. Lawrence",male,34,0,0,248698,13,D56,S
23,1,3,"McGowan, Miss. Anna ""Annie""",female,15,0,0,330923,8.0292,,Q
24,1,1,"Sloper, Mr. William Thompson",male,28,0,0,113788,35.5,A6,S
25,0,3,"Palsson, Miss. Torborg Danira",female,8,3,1,349909,21.075,,S
26,1,3,"Asplund, Mrs. Carl Oscar (Selma Augusta
Emilia Johansson)",female,38,1,5,347077,31.3875,,S
```


[1] Clustering

گام پنجم: درستی مدل[1]: بعد از اینکه مدل ساخته شد آن را مورد تحلیل و بررسی قرار می‌دهیم و مقدار صحت یا درستی مدل را به دست می‌آوریم. معمولا در مجموعه داده‌هایی که انتخاب می‌کنیم اگر داده‌ها درست باشند، الگوریتمی که برای مدل‌سازی انتخاب کردیم درست باشد و پارامترهای مناسبی انتخاب کرده باشیم، میزان درستی مدل بالای ۸۰ درصد می‌شود. در دنیای واقعی احتمال وقوع هر پدیده‌ای ۵۰ درصد است. پس اگر درستی مدل را ۵۰ درصد به دست آوریم در واقع هیچ کاری انجام نداده‌ایم. از طرفی درستی مدل ۱۰۰ درصد نخواهد شد مگر اینکه تعدادجامعه آماری و نمونه‌ها کم باشد و ورودی‌ها شباهت خیلی زیادی به جامعه آماری داشته باشند. در چنین شرایطی می‌توانیم نزدیک به ۱۰۰درصد یا دقیقا ۱۰۰ درصد درستی مدل داشته باشیم. در مجموعه داده‌هایی که تعداد موردها در مدل‌سازی زیاد باشد به دست آوردن درستی مدل کار پیچیده‌ای می‌شود. گاهی ممکن است آن قدر میزان درستی مدل پایین باشد که مجبور شویم به مرحله اول یعنی شناخت کسب و کار برگردیم و مراحل را دوباره تکرار کنیم. به همین دلیل است که مدیران علاقه‌ای به پروژه‌های حوزه داده‌کاوی، یادگیری ماشین و علم داده ندارند. اما اگر پروژه جواب دهد آنگاه وارد مرحله بعد می‌شویم.

نرم افزارهایی که تولید می‌کنیم و استفاده می‌کنیم را سامانه‌های توصیه گر یا DSS می‌نامیم.

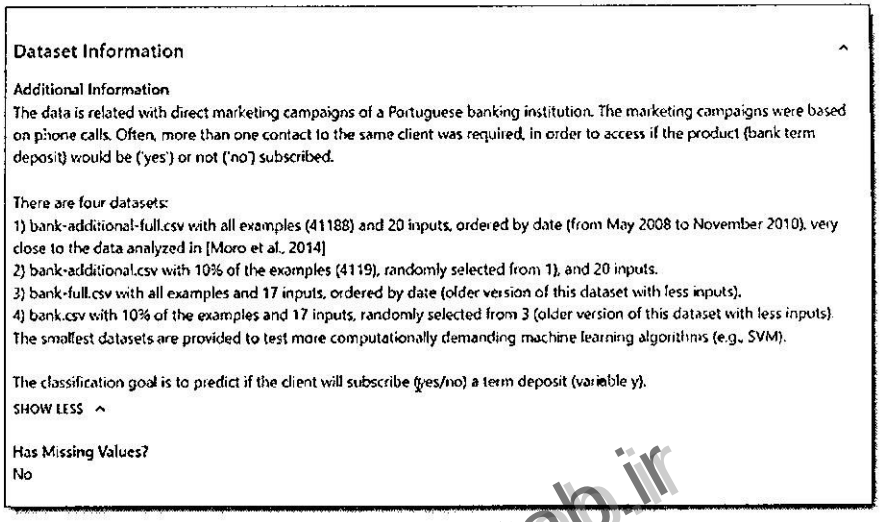
اعضای تیم در متد CRISP: یک مثلث را تشکیل می‌دهند که اضلاع آن متخصص کسب و کار[2]، متخصص فنی[3] و Statistics Experts هستند.

وقتی یک مجموعه داده داریم و می‌خواهیم روی آن کار تحلیلی انجام دهیم، در درجه اول باید یک شناختی روی آن پیدا کنیم. پس به یک Data Dictionary نیاز داریم. یعنی ضروری است یک سری اطلاعات راجع به مجموعه داده‌ها و هر یک از ستون‌ها کسب کنیم. Data Dictionary را می‌توانیم از وب سایت‌هایی مانند Kaggle، GitHub و غیره نیز جستجو کنیم.

---

[1] Evaluation
[2] Business Experts
[3] Technical Experts

در وب سایت Kaggle گفته شده است که هر مجموعه داده برای چه کاری مناسب است. در این وب سایت کدهای پایتون را نیز می‌توانیم ببینیم.


Dataset Information

Additional Information
The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed.

There are four datasets:
1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]
2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs.
3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs).
4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs).
The smallest datasets are provided to test more computationally demanding machine learning algorithms (e.g., SVM).

The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y).
SHOW LESS

Has Missing Values?
No

تصویر ۱-۲

از وب سایت UCI Machine Learning Repository اطلاعاتی در مورد مجموعه داده‌های بازاریابی بانکی[1] به دست می‌آوریم.

در این وب سایت به جای عبارت Data Dictionary از عبارت Dataset Information استفاده می‌شود. با مطالعه توضیحات مربوط به این مجموعه داده متوجه می‌شویم که این داده‌ها مربوط به کمپین بازاریابی مستقیم است. از طریق شماره تماس یا به صورت حضوری افراد را متقاعد می‌کنند که از آنها وام بگیرند. نتایج کمپین‌های گذشته را نیز به ما نشان می‌دهند. در مورد این کمپین برای مثال گفته شده ۴۱۰۰۰ نفر از افراد جامعه نمونه شده‌اند. این ۴۱ هزار داده، اطلاعات افراد با ویژگی‌های مختلف است. همین‌طور نشان می‌دهد داده‌هایی که ما در اختیار داریم ۱۰ درصد از کل داده‌ها است. این مجموعه داده ۱۷ ورودی و یک خروجی دارد.

از این منبع داده‌ای با هدف Classification می‌توانیم استفاده کنیم. برای هرکدام از ستون‌ها توضیحاتی نوشته شده است که به بعضی از مهم‌ترین آن‌ها اشاره می‌نماییم:

---

[1] Bank Marketing

سن افراد (Age)

شغل (Job): مدیر (admin)، کارآفرین (entrepreneur)، خانه‌دار (housemaid)، بازنشسته (retired)، کارگر (blue-collar)، تکنسین (technician)، بیکار (unemployed)، برای خودش کار می‌کند (self-employed)؛

توضیحات مربوط به وضعیت تاهل (marital): جدا شده (divorced)، متاهل (married)، مجرد (single)؛

توضیحات مربوط به تحصیلات (education): تحصیلات مقدماتی (high school)، بی سواد (illiterate)، دوره حرفه‌ای (professional)، تحصیلات دانشگاهی (university)؛

Variables Table

Variable Name	Role	Type	Demographic	Description
age	Feature	Integer	Age	
job	Feature	Categorical	Occupation	type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
marital	Feature	Categorical	Marital Status	marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
education	Feature	Categorical	Education Level	(categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
default	Feature	Binary		has credit in default?
balance	Feature	Integer		average yearly balance
housing	Feature	Binary		has housing loan?
loan	Feature	Binary		has personal loan?
contact	Feature	Categorical		contact communication type (categorical: 'cellular','telephone')
day_of_week	Feature	Date		last contact day of the week

تصویر ۱-۳

Default: یعنی برای ما مسجل شده که این شخص وام می‌گیرد. اعتبار سنجی کردیم و مطمئنیم که این شخص می‌تواند وام بگیرد. این ستون از نوع Binary یعنی true و false یا صفر و یکی است؛

Balance: میانگین درآمد سالانه بر مبنای euros؛

Housing: مشخص می‌کند این شخص برای خانه وام گرفته است یا خیر؛

نوع برقراری ارتباط با مشتری (Contact): تلفن (telephone)، cellular؛

Duration: یعنی آخرین تماسی که با مشتری داشتیم چند ثانیه صحبت کردیم؛

Pdays: یعنی از آخرین تماس با این شخص چند روز گذشته است؛

Previous: چند بار کلا با این مشتری تماس گرفتیم؛

پاسخ دادن به تماس(Putcome): پاسخ داده است (success)، پاسخ نداده است (nonexistent).