دانشگاه شهید بهشتی

دانشکده علوم پایه

پروژه درس یادگیری ماشین

Bank-Marketing

استاد : دکتر حسین حاجی ابوالحسن

دانشجو : محمد عضدی

مقطع: کارشناسی ارشد

گرایش: علوم داده ها

تیر ماه ۱۳۹۹

فهرست :

مقدمه و طرح مسئله :

کمپین های بازار بانکی امروزه بیشتر به عقاید انسان متخصص در مورد انتخاب مشتریان بالقوه متکی هستند. این روش وقت گیر است و فاقد دقت است. همانطور که بانک ها اطلاعات بسیار و معامله های مشتری ساختار یافته و دقیق سوابق در دست دارند ، مطلوب است برای تصمیم گیری مبتنی بر داده ها سیستم هایی با ضریب موفقیت بالا در تبلیغات ساخته شود.

مدل ها و تکنیک های یادگیری ماشین پتانسیل بسیار خوبی برای نشان دادن قدرت خود در چنین تنظیمات مشکل را دارند. این گزارش پروژه نشان می دهد که الگوریتم های یادگیری ماشین چگونه می توانند بخصوص در چنین تنظیمات مشکل بصورت عملی استفاده شود.

این پروژه در مورد اجرای روش های یادگیری ماشین بر روی داده هاست(داده ها از سایت kaggle.com گرفته شده است). در این پروژه مجموعه داده (Data set) در نظر گرفته شده شامل داده های یک بانک در کشور پرتغال است که یک کمپین جدید جهت سرمایه گذاری راه اندازی کرده وقصد دارد با توجه به داده هایی که از افراد دارد پیشگویی کند:

که آیا افراد در این کمپین شرکت می کنند یا خیر.

اطلاعات مربوط به فعالیتهای بازاریابی مستقیم یک موسسه بانکی پرتغال است. کارزارهای بازاریابی مبتنی بر تماسهای تلفنی بود. اغلب ، بیش از یک تماس با همان مشتری لازم بود ، برای دسترسی به این محصول ("بله") یا نه ("نه") مشترک اگر دسترسی داشته باشید.

تاریخچه:

چشم انداز تاریخی از بازاریابی و تکامل فناوری اطلاعات را در این صنعت فشرده اطلاعاتی ارائه می دهد ، راه هایی را برای تحقیقات آینده که از این منظر ناشی می شود ، پیشنهاد می کند. با استفاده از یادگیری ماشین در ده سال اخیر، بانکها توانایی پیشگویی برای جذب مشتریان بالقوه در بانکها را بدست آورده است. از جمله استفاده از شبکه های عصبی و مدلهای کارآمد یادگیری ماشین در بازاریابی امور بانکی باعث کم شدن هزینه های تبلیغات و سرعت بیشتر شده است. اکثر بانکهای موفق امروزی از این روشهای داده کاوی و پردازش داده ها استفاده می کنند که با درصد درستی بالایی همراه است.

شرح داده ها:

این داده شامل ۲۱ ویژگی(ستون) و ۴۱۱۸۸ مشاهده(سطر) است. که مجموع کل داده ۴۱۱۸۸ است و حال ویژگی ها را اندکی توضیح می دهیم:

مهمترین داده های ما اعم از سن، شغل، وضعیت تاهل(طلاق-ازدواج-مجرد-مشخص نیست)، آموزش(بیسواد- ابتدایی-متوسطه-حرفه-دانشگاه-مشخص نیست)، وام، وام مسکن، وام شخصی است.

یک سری از داده ها که مرتبط با آخرین تماس با کمپین فعلی:

تماس: نوع ارتباط با مخاطب (طبقه بندی شده: "تلفن همراه" ، "تلفن")

ماه: آخرین ماه تماس سال (طبقه بندی شده: ' 'jan، 'feb'، 'mar'، ....، 'nov'، 'dec'')

روز: آخرین روز تماس هفته (طبقه بندی شده: 'mon' ، 'سه' ، ' 'wed، 'thu'، 'fri'')

مدت زمان: آخرین مدت زمان تماس ، چند ثانیه (عددی).

نکته مهم: این ویژگی به شدت بر روی هدف خروجی تأثیر می گذارد (به عنوان مثال ، اگر مدت زمان = ۰ باشد ، y = 'no'). با این وجود ، مدت زمان قبل از انجام تماس مشخص نیست. همچنین ، پس از پایان تماس y کاملاً مشخص است. بنابراین ، این ورودی فقط باید برای اهداف معیار گنجانده شود و در صورتی که قصد داشتن یک مدل پیش بینی واقع گرایانه باشد ، باید از آن صرف نظر کرد.

ویژگی های زمینه های اجتماعی و اقتصادی :

Emp.var.rate: نرخ تغییر اشتغال - شاخص سه ماهه (عددی)

Cons.price.idx: شاخص قیمت مصرف کننده - شاخص ماهانه (عددی)

Cons.conf.idx: شاخص اعتماد مصرف کننده - شاخص ماهانه (عددی)

Euribor3m: نرخ ۳ ماه euribor - شاخص روزانه (عددی)

Nr.employment: تعداد کارمندان - شاخص سه ماهه (عددی)

Variable	Description
age	numeric, age of client
job	categorical, type of job (admin, unknown, unemployed, management, housemaid, entrepreneur, student, blue-collar, self-employed, retired, technician, services)
marital	Categorical, marital status (married, divorced, single. Here "divorced" states the both divorced or widowed)
education	categorical (unknown, secondary, primary and tertiary)
default	binary, customer credit is in default (yes, no)
balance	numeric, average yearly balance (in euros)
housing	binary, status of housing loan (yes, no)
loan	binary, clients personal loan (yes, no)
contact	categorical, contact communication type (unknown, telephone, cellular)
day	numeric, the last contact day of the month range (1-31)
month	categorical, last contact month of the year
duration	numeric, last contact duration (in seconds)
campaign	numeric, number of contacts performed during this campaign
pdays	numeric, number of days that passed by after the client was last contacted from a previous campaign
previous	numeric, number of contacts which are made before this campaign
poutcome	categorical, result or outcome of the previous marketing campaign (unknown, other, failure, success)
y	binary, (desired target) client subscribed a term deposit or not

شرح روش :

در طی این پروژه که در فضای ژوپیتر نوت بوک انجام میشود، ابتدا یک سری دیداری کردن و پیش پردازش داده ها را انجام میدهیم که بیشتر به صورت نمودارهای فراوانی نشان داده میشود. ودر ادامه راستای کاهش بعد قدم برمیداریم.

با بررسی بیشتر داده ها و متغیرها، تصمیم به حذف یا اضاف کردن بعضی از متغیرها میگیریم و در نهایت با اعمال مدلهای زیر بروی داده ها میپردازیم:

1: Logestic regression

2: Decision tree

3: Random Forest

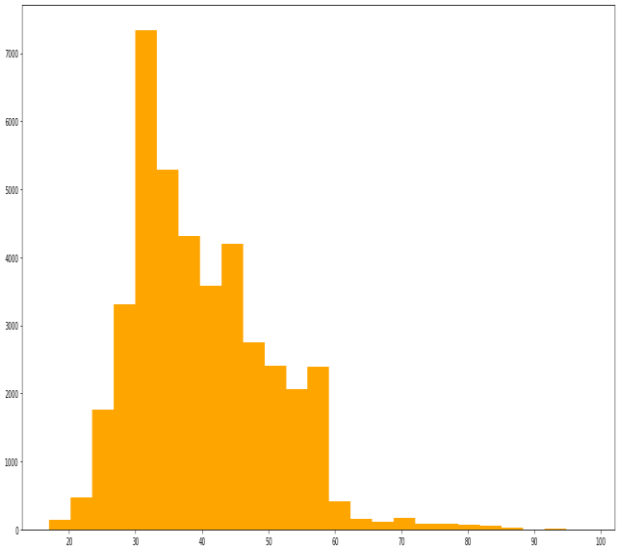
البته میتوان از مدلهای دیگری مانند NAIVE BAYSE ، SVM ، KNN و یا شبکه های عصبی استفاده کرد و خروجی به میزان نسبی خوبی از داده ها در پیشگویی مان گرفت.

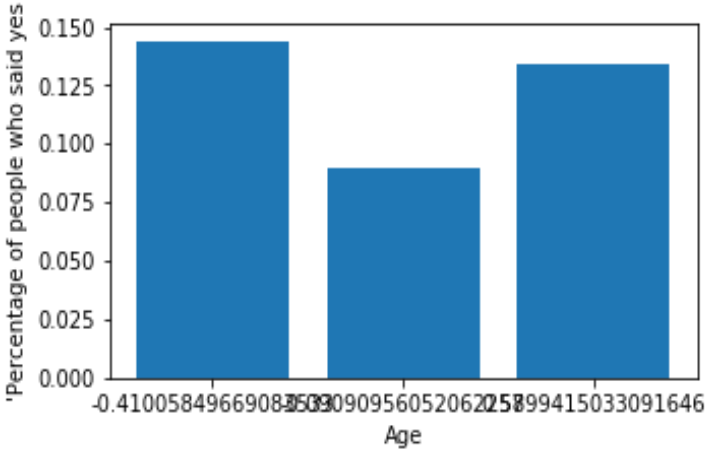
اجرا و شرح مدل:

در ابتدا برای پیش پردازش داده ها از بصری سازی و روش های داده کاوی که شامل پاکسازی داده ها و داده های گمشده استفاده میکنیم که به ما کمک میکنه تا تحلیل و دید خوبی از مجموعه داده ها داشته باشیم.

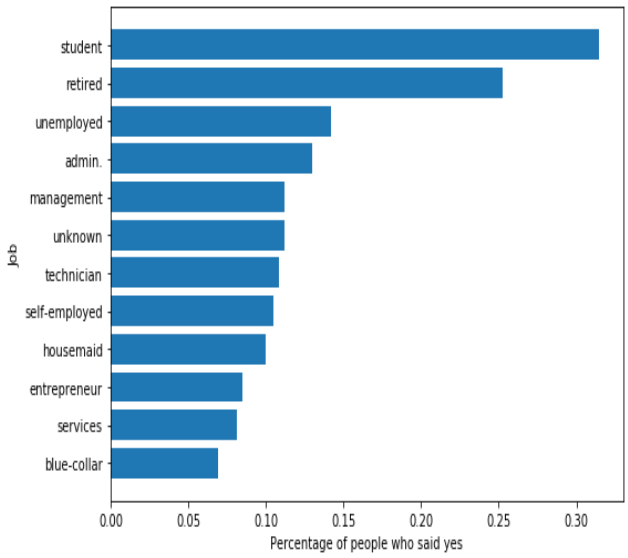
نمودار فراوانی سن و درصد افرادی که بر حسب سن شرکت کرده اند و جوابشان مثبت بوده است را در زیر میبینیم :

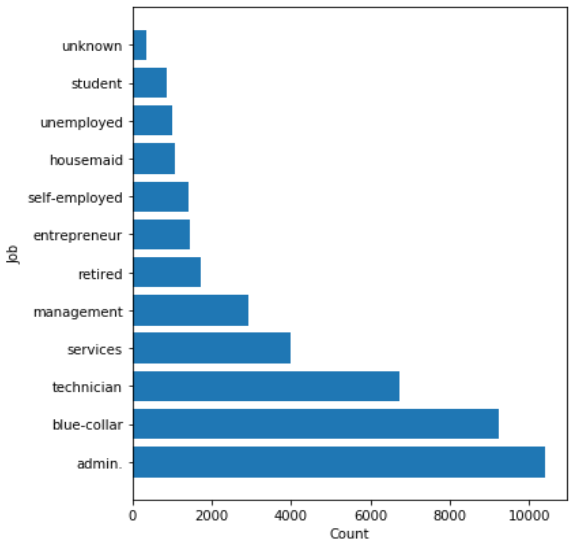
نمودار فراوانی سن را در زیر میبینیم که بیشتر شرکت کنندگان در چه رنج سنی هستن:



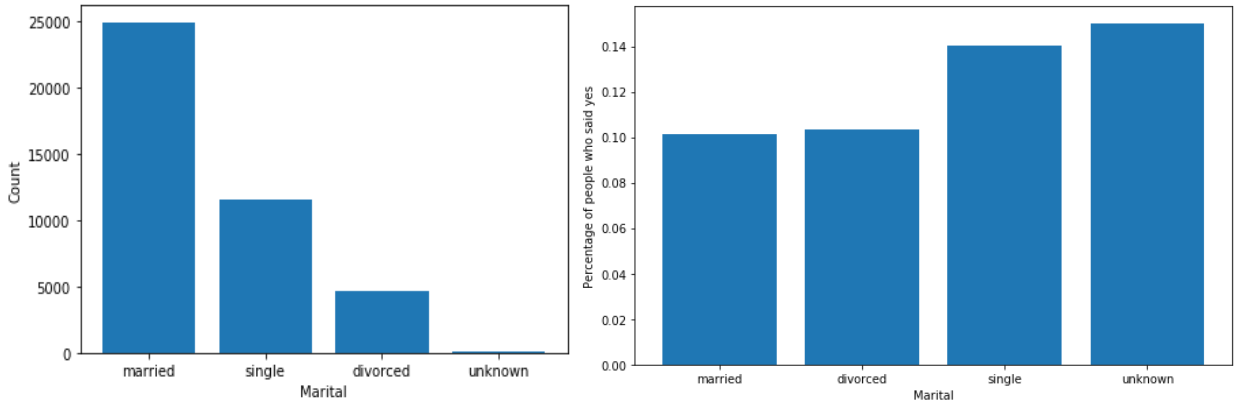


نمودار فراوانی شغل را در زیر میبینیم که بیشتر شرکت کنندگان در چه شغلهایی جواب بعله داده اند :

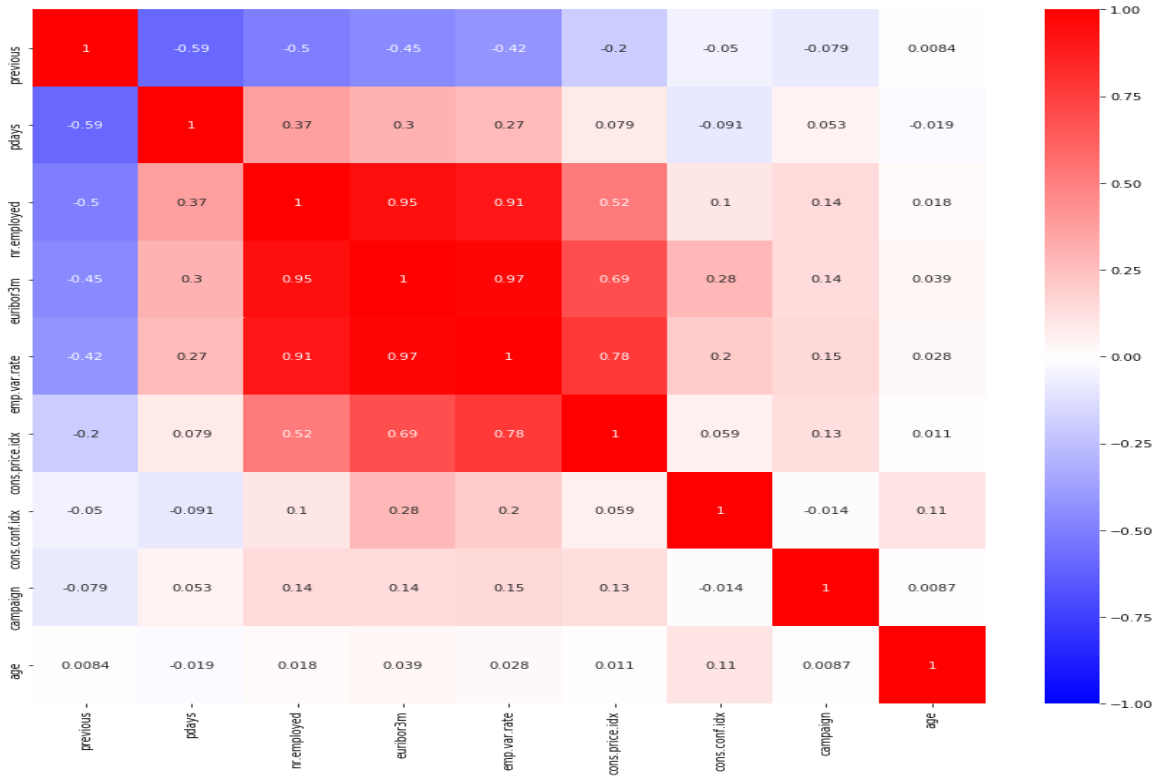




نمودار فراوانی وضعیت تاهل را در زیر میبینید که با توجه به این که ناشناخته ها مانند مجردها رفتار می کند می توانیم آنها را برای کاهش ابعادی در یک متغیر قرار دهیم :



Heat Map: Visualizing Correlation



بررسی نتایج:

1: Logestic regression

1.1: train set accuracy 0.910437292616472

1.2: test set accuracy 0.9101723719349356

1.3: Confusion matrix and classification-report for training set

```
[[32030   861]
 [ 2459  1719]]
```

	precision	recall	f1-score	support
no	0.93	0.97	0.95	32891
yes	0.67	0.41	0.51	4178
accuracy			0.91	37069
macro avg	0.80	0.69	0.73	37069
weighted avg	0.90	0.91	0.90	37069

1.4: Confusion matrix and classification-report for test set

```
array([[3549,  108],
       [ 262,  200]], dtype=int64)
```

	precision	recall	f1-score	support
no	0.93	0.97	0.95	3657
yes	0.65	0.43	0.52	462
accuracy			0.91	4119
macro avg	0.79	0.70	0.73	4119
weighted avg	0.90	0.91	0.90	4119

2: Decision tree

2.1: train set accuracy 0.928376810812269

2.2: test set accuracy 0.9045884923525127

2.3: Confusion matrix and classification-report for training set

```
array([[31817,  1074],
       [ 1581,  2597]], dtype=int64)
              precision    recall  f1-score   support

          no       0.95      0.97      0.96     32891
         yes       0.71      0.62      0.66      4178

    accuracy                           0.93     37069
   macro avg       0.83      0.79      0.81     37069
weighted avg       0.93      0.93      0.93     37069
```

2.4: Confusion matrix and classification-report for test set

```
array([[3481,  176],
       [ 217,  245]], dtype=int64)
              precision    recall  f1-score   support

          no       0.94      0.95      0.95      3657
         yes       0.58      0.53      0.55       462

    accuracy                           0.90      4119
   macro avg       0.76      0.74      0.75      4119
weighted avg       0.90      0.90      0.90      4119
```

3: Random Forest

3.1: train set accuracy 0.9537619034772991

3.1: test set accuracy 0.9133284777858703

3.3: Confusion matrix and classification-report for training set

```
array([[32454,   437],
       [ 1277,  2901]], dtype=int64)
              precision    recall  f1-score   support

          no       0.96      0.99      0.97     32891
         yes       0.87      0.69      0.77      4178

    accuracy                           0.95     37069
   macro avg       0.92      0.84      0.87     37069
weighted avg       0.95      0.95      0.95     37069
```

3.4: Confusion matrix and classification-report for test se

```
array([[3527,  130],
       [ 227,  235]], dtype=int64)
              precision    recall  f1-score   support

          no       0.94      0.96      0.95      3657
         yes       0.64      0.51      0.57       462

    accuracy                           0.91      4119
   macro avg       0.79      0.74      0.76      4119
weighted avg       0.91      0.91      0.91      4119
```

هدف از انتخاب مدل ، یافتن معماری شبکه با بهترین خصوصیات تعمیم ، یعنی مواردی است که خطاهای موجود در موارد انتخاب شده مجموعه داده را به حداقل می رساند. در پایان پروژه بعد از اعمال مدلهای بالا و خروجی گرفتن از داده برای انتخاب بهترین مدل خود برای بررسی و پیش بینی داده هایمان استفاده میکنیم. که در نهایت مدل random forest بهترین مدل انتخابی ما خواهد بود.

منابع:

Understanding Machine Learning: From Theory to Algorithms, by Shai Shalev-Shwartz and Shai Ben-David

Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd Edition) by Aurelien Geron

Bank Marketing with Machine Learning-Zewei Chu-April 19, 2015

link:

https://www.kaggle.com/henriqueyamahata/bank-marketing

Dataset from：http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

https://www.neuraldesigner.com/learning/examples/bank-marketing-campaign

https://www.researchgate.net/publication/340788220_Mining_a_Marketing_Campaigns_Data_of_Bank

https://nycdatascience.com/blog/student-works/machine-learning/machine-learning-retail-bank-marketing-data/