

# A社データセットの概要

## 1. データベースについて

- 電気通信事業会社A社の過去のデータが含まれたデータベースファイル「telecom.sqlite3」が配布されている
- カラム名とその説明が明らかになっている（「4-3. カラムの説明」参照）
- 以下の情報が格納された2種類のテーブルが含まれている
  1. 顧客情報（Client）
  2. 利用履歴（Record）
- データセットの特徴
  - 合計100000レコード×約100カラムのサイズの大きいテーブルデータ
    - EDAをしっかり行い、データについて理解する必要がある
    - コンペで扱ったデータと似たような性質を持つ
  - 最初にSQLを用いてデータを取得する
    - サンプルコードを配布しておりますので、どなたでもお気軽に取り組むことができます
- こんな方におすすめ
  - コンペで培ってきたEDAやモデリング力を試したい
  - SQLの復習をしたい

## 2. データベースの利用

### A. SQLiteのインストールができている場合

1. SQLiteを用いて、データベースに接続
2. 各自、データベースを観察
3. データ分析、事業提案に必要なデータをcsv形式で出力
4. 各自必要なpythonモジュール（pandasなど）でデータ分析

## B. SQLiteのインストールができていない場合

1. pythonの標準ライブラリを用いて、Aと同様のことを実行する

## C. 何から手をつければ良いか分からない場合

1. 参考資料として配布されているsample codeを使用する
2. 全データをcsv形式で出力
3. 各自必要なpythonモジュール（pandasなど）でデータ分析

## 3. カラムの説明

A社の担当者にデータについて問い合わせたところ、以下の説明が送られてきました。現段階では正式なコンサルティング契約は未締結であり、情報保護規程の厳しいA社からこれ以上の詳しい情報開示を受けることはできません。

カテゴリカル変数の各種記号や略語の意味など不明な点は多々ありますが、不明な点があること自体も業務上の与条件として分析を進めてください。

カラム名	内容の説明
rev_Mean	Mean monthly revenue (charge amount)
mou_Mean	Mean number of monthly minutes of use
totmrc_Mean	Mean total monthly recurring charge
da_Mean	Mean number of directory assisted calls
ovrmou_Mean	Mean overage minutes of use
ovrrev_Mean	Mean overage revenue
vceovr_Mean	Mean revenue of voice overage
datovr_Mean	Mean revenue of data overage
roam_Mean	Mean number of roaming calls
change_mou	Percentage change in monthly minutes of use vs previous three month average
change_rev	Percentage change in monthly revenue vs previous three month average
drop_vce_Mean	Mean number of dropped (failed) voice calls
drop_dat_Mean	Mean number of dropped (failed) data calls

blck_vce_Mean	Mean number of blocked (failed) voice calls
blck_dat_Mean	Mean number of blocked (failed) data calls
unan_vce_Mean	Mean number of unanswered voice calls
unan_dat_Mean	Mean number of unanswered data calls
plcd_vce_Mean	Mean number of attempted voice calls placed
plcd_dat_Mean	Mean number of attempted data calls placed
recv_vce_Mean	Mean number of received voice calls
recv_sms_Mean	NaN
comp_vce_Mean	Mean number of completed voice calls
comp_dat_Mean	Mean number of completed data calls
custcare_Mean	Mean number of customer care calls
ccrndmou_Mean	Mean rounded minutes of use of customer care calls
cc_mou_Mean	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls
inonemin_Mean	Mean number of inbound calls less than one minute
threeway_Mean	Mean number of three way calls
mou_cvce_Mean	Mean unrounded minutes of use of completed voice calls
mou_cdat_Mean	Mean unrounded minutes of use of completed data calls
mou_rvce_Mean	Mean unrounded minutes of use of received voice calls
owylis_vce_Mean	Mean number of outbound wireless to wireless voice calls
mouowylisv_Mean	Mean unrounded minutes of use of outbound wireless to wireless voice calls
iwylis_vce_Mean	NaN
mouiwylisv_Mean	Mean unrounded minutes of use of inbound wireless to wireless voice calls
peak_vce_Mean	Mean number of inbound and outbound peak voice calls
peak_dat_Mean	Mean number of peak data calls
mou_peav_Mean	Mean unrounded minutes of use of peak voice calls
mou_pead_Mean	Mean unrounded minutes of use of peak data calls
opk_vce_Mean	Mean number of off-peak voice calls
opk_dat_Mean	Mean number of off-peak data calls
mou_opkv_Mean	Mean unrounded minutes of use of off-peak voice calls
mou_opkd_Mean	Mean unrounded minutes of use of off-peak data calls

drop_blk_Mean	Mean number of dropped or blocked calls
attempt_Mean	Mean number of attempted calls
complete_Mean	Mean number of completed calls
callfwdv_Mean	Mean number of call forwarding calls
callwait_Mean	Mean number of call waiting calls
churn	Instance of churn between 31-60 days after observation date
months	Total number of months in service
uniqsubs	Number of unique subscribers in the household
actvsubs	Number of active subscribers in household
new_cell	New cell phone user
crclscod	Credit class code
asl_flag	Account spending limit
totcalls	Total number of calls over the life of the customer
totmou	Total minutes of use over the life of the customer
totrev	Total revenue
adjrev	Billing adjusted total revenue over the life of the customer
adjmou	Billing adjusted total minutes of use over the life of the customer
adjqty	Billing adjusted total number of calls over the life of the customer
avgrev	Average monthly revenue over the life of the customer
avgmou	Average monthly minutes of use over the life of the customer
avgqty	Average monthly number of calls over the life of the customer
avg3mou	Average monthly minutes of use over the previous three months
avg3qty	Average monthly number of calls over the previous three months
avg3rev	Average monthly revenue over the previous three months
avg6mou	Average monthly minutes of use over the previous six months
avg6qty	Average monthly number of calls over the previous six months
avg6rev	Average monthly revenue over the previous six months
prizm_social_one	Social group letter only
area	Geogrpahic area
dualband	Dualband
refurb_new	Handset: refurbished or new
hnd_price	Current handset price

phones	Number of handsets issued
models	Number of models issued
hnd_webcap	Handset web capability
truck	Truck indicator
rv	RV indicator
ownrent	Home owner/renter status
lor	Length of residence
dwlltype	Dwelling Unit type
marital	Marital Status
adults	Number of adults in household
infobase	InfoBase match
income	Estimated income
numbcars	Known number of vehicles
HHstatin	Premier household status indicator
dwllsize	Dwelling size
forgrntvl	Foreign travel dummy variable
ethnic	Ethnicity roll-up code
kid0_2	Child 0 - 2 years of age in household
kid3_5	Child 3 - 5 years of age in household
kid6_10	Child 6 - 10 years of age in household
kid11_15	Child 11 - 15 years of age in household
kid16_17	Child 16 - 17 years of age in household
credited	Credit card indicator
eqpdays	Number of days (age) of current equipment
Customer_ID	NaN

以下は、各特徴量の内容の簡潔な解説です：

1. **rev\_Mean**: 月間平均収益(課金額)
2. **mou\_Mean**: 月間平均利用時間
3. **totmrc\_Mean**: 月間平均総月額従量課金
4. **da\_Mean**: ディレクトリアシスト呼び出しの月間平均回数
5. **ovrmou\_Mean**: 超過利用の月間平均分
6. **ovrrev\_Mean**: 超過収益の月間平均額
7. **vceovr\_Mean**: 音声超過の平均収益
8. **datovr\_Mean**: データ超過の平均収益
9. **roam\_Mean**: ローミング呼び出しの月間平均回数
10. **change\_mou**: 月間利用時間の前三か月平均との比較の変化率
11. **change\_rev**: 月間収益の前三か月平均との比較の変化率
12. **drop\_vce\_Mean**: 失敗した音声通話の月間平均数
13. **drop\_dat\_Mean**: 失敗したデータ通話の月間平均数
14. **blk\_vce\_Mean**: ブロックされた音声通話の月間平均数
15. **blk\_dat\_Mean**: ブロックされたデータ通話の月間平均数
16. **unan\_vce\_Mean**: 答えられなかった音声通話の月間平均数
17. **unan\_dat\_Mean**: 答えられなかったデータ通話の月間平均数
18. **plcd\_vce\_Mean**: 試みられた音声通話の月間平均数
19. **plcd\_dat\_Mean**: 試みられたデータ通話の月間平均数
20. **recv\_vce\_Mean**: 受け取った音声通話の月間平均数
21. **recv\_sms\_Mean**: 受け取ったSMSの月間平均数
22. **comp\_vce\_Mean**: 完了した音声通話の月間平均数
23. **comp\_dat\_Mean**: 完了したデータ通話の月間平均数
24. **custcare\_Mean**: 顧客ケア通話の月間平均数
25. **ccrndmou\_Mean**: 顧客ケア通話の丸められた利用時間の月間平均
26. **cc\_mou\_Mean**: 顧客ケア通話の丸められていない利用時間の月間平均
27. **inonemin\_Mean**: 1分未満の着信通話の月間平均数
28. **threeway\_Mean**: 3者通話の月間平均数
29. **mou\_cvce\_Mean**: 完了した音声通話の丸められていない利用時間の月間平均

- 30. **mou\_cdat\_Mean**: 完了したデータ通話の丸められていない利用時間の月間平均
- 31. **mou\_rvce\_Mean**: 受け取った音声通話の丸められていない利用時間の月間平均
- 32. **owylis\_vce\_Mean**: ワイヤレスからワイヤレスへの発信音声通話の月間平均数
- 33. **mouowylisv\_Mean**: ワイヤレスからワイヤレスへの発信音声通話の丸められていない利用時間の月間平均
- 34. **iwylis\_vce\_Mean**: ワイヤレスからワイヤレスへの着信音声通話の月間平均数
- 35. **mouiwyylisv\_Mean**: ワイヤレスからワイヤレスへの着信音声通話の丸められていない利用時間の月間平均
- 36. **peak\_vce\_Mean**: ピーク時の音声通話の月間平均数
- 37. **peak\_dat\_Mean**: ピーク時のデータ通話の月間平均数
- 38. **mou\_peav\_Mean**: ピーク時の音声通話の丸められていない利用時間の月間平均
- 39. **mou\_pead\_Mean**: ピーク時のデータ通話の丸められていない利用時間の月間平均
- 40. **opk\_vce\_Mean**: オフピーク時の音声通話の月間平均数
- 41. **opk\_dat\_Mean**: オフピーク時のデータ通話の月間平均数
- 42. **mou\_opkv\_Mean**: オフピーク時の音声通話の丸められていない利用時間の月間平均
- 43. **mou\_opkd\_Mean**: オフピーク時のデータ通話の丸められていない利用時間の月間平均
- 44. **drop\_blk\_Mean**: ドロップまたはブロックされた通話の月間平均数
- 45. **attempt\_Mean**: 試みられた通話の月間平均数
- 46. **complete\_Mean**: 完了した通話の月間平均数
- 47. **callfwdv\_Mean**: 転送通話の月間平均数
- 48. **callwait\_Mean**: 待ち通話の月間平均数
- 49. **churn**: 観測日から31-60日後の離脱のインスタンス
- 50. **months**: サービス利用期間の総月数
- 51. **uniqusubs**: 世帯内のユニークな加入者数
- 52. **actvsubs**: 世帯内のアクティブな加入者数
- 53. **new\_cell**: 新しい携帯電話ユーザーかどうか
- 54. **crclscod**: 信用クラスコード
- 55. **asl\_flag**: アカウントの支出制限の有無
- 56. **totcalls**: 顧客の生涯通話回数
- 57. **totmou**: 顧客の生涯利用時間
- 58. **totrev**: 総収益
- 59. **adjrev**: 顧客の生涯課金に基づいた調整済み総収益
- 60. **adjmou**: 顧客の生涯利用時間に基づいた調整済み総収益
- 61. **adjqty**: 顧客の生涯通話回数に基づいた調整済み総収益
- 62. **avgrev**: 顧客の生涯平均月間収益
- 63. **avgmou**: 顧客の生涯平均月間利用時間
- 64. **avgqty**: 顧客の生涯平均月間通話回数
- 65. **avg3mou**: 直近3か月の平均月間利用時間
- 66. **avg3qty**: 直近3か月の平均月間通話回数
- 67. **avg3rev**: 直近3か月の平均月間収益
- 68. **avg6mou**: 直近6か月の平均月間利用時間
- 69. **avg6qty**: 直近6か月の平均月間通話回数

- 70. avg6rev: 直近6か月の平均月間収益
- 71. prizm\_social\_one: ソーシャルグループの文字列
- 72. area: 地理的なエリア
- 73. dualband: デュアルバンド
- 74. refurb\_new: 携帯電話の状態(再生産または新品)
- 75. hnd\_price: 現在の携帯電話の価格
- 76. phones: 発行された携帯電話の数
- 77. models: 発行された携帯電話のモデル数
- 78. hnd\_webcap: 携帯電話
- 79. hnd\_webcap: 携帯電話のウェブ機能の有無
- 80. truck: トラックの指標
- 81. rv: RVの指標
- 82. ownrent: 住宅所有者/賃借人のステータス
- 83. lor: 居住期間
- 84. dwlltype: 住居のタイプ
- 85. marital: 結婚ステータス
- 86. adults: 世帯内の大人の数
- 87. infobase: InfoBaseの一致
- 88. income: 推定収入
- 89. numbcars: 既知の車両数
- 90. HHstatin: プレミアム家庭ステータスインジケータ
- 91. dwllsize: 住居のサイズ
- 92. forgntvl: 外国旅行のダミー変数
- 93. ethnic: 民族の統合コード
- 94. kid0\_2: 世帯内の0-2歳の子供の数
- 95. kid3\_5: 世帯内の3-5歳の子供の数
- 96. kid6\_10: 世帯内の6-10歳の子供の数
- 97. kid11\_15: 世帯内の11-15歳の子供の数
- 98. kid16\_17: 世帯内の16-17歳の子供の数
- 99. creditcd: クレジットカード指標
- 100. eqpdays: 現在の機器の日数(年齢)
- 101. Customer\_ID: 顧客ID