IDEA 1

Data Set Description: Drug Review Dataset

Data location: https://www.kaggle.com/jessicali9530/kuc-hackathon-winter-2018

Data Size: 210k * 7

Columns: Unique ID, Drug, Condition, Review (text), Rating, Date, Usefulness Count

Question: Can we predict someone's drug rating?

Significance of the question: If people rate the drug higher, they are more likely to take it for chronic health issues. Side effect profiles can increase ratings or decrease ratings, independent of efficacy. Pharmaceutical companies can use these findings to aid in the develop new drugs.

Why I'm choosing this dataset: opportunity to practice NLP / shows interest in health issues.

IDEA 2:

Data Set Description: Fraudulent Health Claims

Data location:
https://www.kaggle.com/rohitrox/medical-provider-fraud-detection

Data Size: 50k * 27

Question: Can we predict which doctors have fraudulent practices?

Significance of the question: Save health insurance providers money by being able to identify criminal activity.

Why I'm choosing this dataset:

- Logistic regression
- Putting together different data sets
- Creating columns with sensible metrics to solve a problem.
- Using Claims Data

Columns: Multiple tables

1. Provider, Potential Fraud
2. Beneficiary Data
    a. BeneID
    b. DOB
    c. DOD
    d. Gender
    e. Race
    f. RenalDiseaseIndicator
    g. State
    h. County
    i. NoOfMonths_PartACov
    j. NoOfMonths_PartBCov
    k. ChronicCond_Alzheimer
    l. ChronicCond_Heartfailure
    m. ChronicCond_KidneyDisease
    n. ChronicCond_Cancer
    o. ChronicCond_ObstrPulmonary
    p. ChronicCond_Depression
    q. ChronicCond_Diabetes
    r. ChronicCond_IschemicHeart
    s. ChronicCond_Osteoporasis

    t. ChronicCond_rheumatoidarthritis
    u. ChronicCond_stroke
    v. IPAnnualReimbursementAmt
    w. IPAnnualDeductibleAmt
    x. OPAnnualReimbursementAmt
    y. OPAnnualDeductibleAmt
    z.
3. Inpatient Data
    a. BeneID
    b. ClaimID
    c. ClaimStartDt
    d. ClaimEndDt
    e. Provider
    f. InscClaimAmtReimbursed
    g. AttendingPhysician
    h. OperatingPhysician
    i. OtherPhysician
    j. AdmissionDt
    k. ClmAdmitDiagnosisCode
    l. DeductibleAmtPaid
    m. DischargeDt
    n. DiagnosisGroupCode
    o. ClmDiagnosisCode_1
    p. ClmDiagnosisCode_2
    q. ClmDiagnosisCode_3
    r. ClmDiagnosisCode_4
    s. ClmDiagnosisCode_5
    t. ClmDiagnosisCode_6
    u. ClmDiagnosisCode_7
    v. ClmDiagnosisCode_8
    w. ClmDiagnosisCode_9
    x. ClmDiagnosisCode_10
    y. ClmProcedureCode_1
    z. ClmProcedureCode_2
    aa. ClmProcedureCode_3
    bb. ClmProcedureCode_4
    cc. ClmProcedureCode_5
    dd. ClmProcedureCode_6
4. Outpatient Data
    a. BeneID
    b. ClaimID
    c. ClaimStartDt
    d. ClaimEndDt
    e. Provider
    f. InscClaimAmtReimbursed
    g. AttendingPhysician
    h. OperatingPhysician
    i. OtherPhysician
    j. ClmDiagnosisCode_1
    k. ClmDiagnosisCode_2

l. ClmDiagnosisCode_3
m. ClmDiagnosisCode_4
n. ClmDiagnosisCode_5
o. ClmDiagnosisCode_6
p. ClmDiagnosisCode_7
q. ClmDiagnosisCode_8
r. ClmDiagnosisCode_9
s. ClmDiagnosisCode_10
t. ClmProcedureCode_1
u. ClmProcedureCode_2
v. ClmProcedureCode_3
w. ClmProcedureCode_4
x. ClmProcedureCode_5
y. ClmProcedureCode_6
z. DeductibleAmtPaid
aa. ClmAdmitDiagnosisCode
bb.

-

IDEA 3:

Data Set Description: Energy Use

Datalocation:
https://www.kaggle.com/robikscube/hourly-energy-consumption#DUQ_hourly.csv

Question: Can we accurately predict trends in energy use for the past year?

Significance of the question: Provide energy information for the government for developing incentives for energy conservation.

Why I'm choosing this dataset:
- Time Series

Columns: Multiple tables, one for each provider, hourly data for up to 14 years for each.