Springboard Data Science Career Track

Ideas for Capstone Project 1

1. Predicting hospital readmission of patients with diabetes.

The response variable has three outcomes/classes: no readmission, readmission within 30 days and readmission beyond 30 days. Patients' personal information, laboratory test results and medications are recorded as dependent variables in the data set. The objective is to figure out which situations contribute to patients' readmissions and classify a patient's readmission status using dependent variable. There are 3 classes in readmission variable and most of the dependent variables are categorical (nominal). Those are challenges for this project.

Features:

Feature name	Type	Description and values	% missing
Encounter ID	Numeric	Unique identifier of an encounter	0%
Patient number	Numeric	Unique identifier of a patient	0%
Race	Nominal	Values: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Values: male, female, and unknown/invalid	0%
Age	Nominal	Grouped in 10-year intervals: 0, 10), 10, 20), …, 90, 100)	0%
Weight	Numeric	Weight in pounds.	97%
Admission type	Nominal	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available	0%
Discharge disposition	Nominal	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available	0%
Admission source	Nominal	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital	0%
Time in hospital	Numeric	Integer number of days between admission and discharge	0%
Payer code	Nominal	Integer identifier corresponding to 23 distinct values, for example, Blue Cross/Blue Shield, Medicare, and self-pay	52%

Medical specialty	Nominal	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab procedures	Numeric	Number of lab tests performed during the encounter	0%
Number of procedures	Numeric	Number of procedures (other than lab tests) performed during the encounter	0%
Number of medications	Numeric	Number of distinct generic names administered during the encounter	0%
Number of outpatient visits	Numeric	Number of outpatient visits of the patient in the year preceding the encounter	0%
Number of emergency visits	Numeric	Number of emergency visits of the patient in the year preceding the encounter	0%
Number of inpatient visits	Numeric	Number of inpatient visits of the patient in the year preceding the encounter	0%
Diagnosis 1	Nominal	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values	0%
Diagnosis 2	Nominal	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values	0%
Diagnosis 3	Nominal	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values	1%
Number of diagnoses	Numeric	Number of diagnoses entered to the system	0%
Glucose serum test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured	0%
A1c test result	Nominal	Indicates the range of the result or if the test was not taken. Values: “>8” if the result was greater than 8%, “>7” if the result was greater than 7% but less than 8%, “normal” if the result was less than 7%, and “none” if not measured.	0%
Change of medications	Nominal	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”	0%
Diabetes medications	Nominal	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”	0%

24 features for medications	Nominal	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage. Values: “up” if the dosage was increased during the encounter, “down” if the dosage was decreased, “steady” if the dosage did not change, and “no” if the drug was not prescribed	0%
Readmitted	Nominal	Days to inpatient readmission. Values: “<30” if the patient was readmitted in less than 30 days, “>30” if the patient was readmitted in more than 30 days, and “No” for no record of readmission.	0%

Questions:

1. What are the main risk factors for readmission?
2. Do basic personal information including race, gender, age and weight have associations with readmission?
3. What are the relationships between 24 medications?
4. Are hospitalized diabetes patients have higher risk of readmission than those without diabetes?
5. Does HbA1c measurement have largest effect on hospital readmission?
6. What are potential ways and barriers to reduce readmission?

Data Source:

Diabetes 130-US hospitals for years 1999-2008 Data Set
https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

2. Predicting Walmart’s department-wide sales for each store.

Walmart has by department weekly-sales data of 45 stores from 2010-02-05 to 2013-07-26. The objective is to predict department-wide sales for each store. Another challenge is to model the effects of promotional markdown on four largest holidays: Super Bowl, Labor Day, Thanksgiving, and Christmas.

Features:

Feature Name	Description
Store	the store number
Store Type	A, B, C
Store Size	the size of store
Dept	the department number
Date	the week
Weekly_Sales	sales for the given department in the given store
Temperature	average temperature in the region
Fuel_Price	cost of fuel in the region
MarkDown1-5	anonymized data related to promotional markdowns that Walmart is running. MarkDown data is only available after Nov 2011, and is not available for all stores all the time. Any missing value is marked with an NA.
CPI	the consumer price index
Unemployment	the unemployment rate
IsHoliday	whether the week is a special holiday week

Questions:

1. What are the patterns for the weekly sales across year for different stores?
2. What is the most profitable department for all stores?
3. What are the patterns of sales among three store types?
4. What are the effects of markdowns on the holiday weeks in the absence of complete/ideal historical data?
5. the feature importance?
6. Appropriate metrics on evaluating the models?

Data Source:

Walmart store sales forecasting
https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

3. Predicting the disease scores using genotypic data and comparing between regression and classification methods.

Gray leaf spot is a maize leaf disease. The data set has ordinal phenotypic data (manually scored from 1 to 5 in field) and 46,374 markers (genotypic data from laboratory, binary) for 272 maize lines. The objective is to predict gray leaf spot score of 272 maize lines using 46,374 markers. This is a $n << p$ problem. Feature engineering will be implemented to find important features for predicting. The response variable is ordinal and has multiple levels. A comparison between regression and classification methods on the predictions will be implemented.

Features:

Feature Name	Description
Line Name	Corn hybrid lines
Marker 1- 46,374	binary marker status

Questions:

1. Cluster the features?
2. What regularization method can be used to minimize overfitting?
3. n << p, training set and test set are small, how to evaluate the model?
4. What is the best method to predict, regression to get predicted decimal values or classification to different score classes?
5. What is difference between classifying ordinal outcomes and nominal outcomes?

Data Source:

Data maize and wheat
https://repository.cimmyt.org/handle/10883/4036?show=full