Classification

Prediction DIABET Patients' Readmission

Agenda

- Objective
- Dataset
- Data cleaning
- Modelling
- Conclusion

OBJECTIVE

- Predict a patient who will be readmitted in 30 days after having treatment in hospital and discharged

DATASET

- **DIabetes 130 US Hospitals for years 1999-2008**
  - **UCI ML Repository**
- **50 VARIABLES**
- **OUTPUT VARIABLE has 3 different values**
  - No readmission;
  - A readmission in less than 30 days
  - A readmission in more than 30 days

variables

Features	
encounter_id	Unique identifier of an encounter
patient_nbr	Unique identifier of a patient
race	Caucasian, Asian, African American, Hispanic, and other
gender	male, female, and unknown/invalid
age	Grouped in 10-year intervals: [0, 10), [10, 20), ..., [90, 100)
weight	Weight in pounds
admission_type_id	Integer identifier corresponding to 9 distinct values, for example, emergency, urgent, elective, newborn, and not available
discharge_disposition_id	Integer identifier corresponding to 29 distinct values, for example, discharged to home, expired, and not available
admission_source_id	Integer identifier corresponding to 21 distinct values, for example, physician referral, emergency room, and transfer from a hospital
time_in_hospital	Integer number of days between admission and discharge
payer_code	Integer identifier corresponding to 23 distinct values, for example, Blue Cross\Blue Shield, Medicare, and self-pay
medical_specialty	Integer identifier of a specialty of the admitting physician, corresponding to 84 distinct values, for example, cardiology, internal medicine, family\general practice, and surgeon
num_lab_procedures	Number of lab tests performed during the encounter
num_procedures	Number of procedures (other than lab tests) performed during the encounter
num_medications	Number of distinct generic names administered during the encounter
number_outpatient	Number of outpatient visits of the patient in the year preceding the encounter

Variables-contd.

number_emergency	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Number of inpatient visits of the patient in the year preceding the encounter
diag_1	The primary diagnosis (coded as first three digits of ICD9); 848 distinct values
diag_2	Secondary diagnosis (coded as first three digits of ICD9); 923 distinct values
diag_3	Additional secondary diagnosis (coded as first three digits of ICD9); 954 distinct values
number_diagnoses	Number of diagnoses entered to the system
max_glu_serum	Indicates the range of the result or if the test was not taken. Values: “>200,” “>300,” “normal,” and “none” if not measured
A1Cresult	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
metformin	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
change	Indicates if there was a change in diabetic medications (either dosage or generic name). Values: “change” and “no change”
diabetesMed	Indicates if there was any diabetic medication prescribed. Values: “yes” and “no”
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
readmitted	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

Data cleaning

MISSING VALUES

	Total	Pct
weight	98569	96.9
medical_specialty	49949	49.1
payer_code	40256	39.6
race	2273	2.2
diag_3	1423	1.4
diag_2	358	0.4
diag_1	21	0.0

- EXtRACTED FROM THE MODEL

Missing values: Weight, medical_specialty,payer_code,

Patient ID: encounter_id, patient_nbr, admission_type_id, discharge_disposition_id, admission_source_id

Dropping variable all values are same: examide, citoglipton, metformin-rosiglitazone

Drop values with NaN

FEATURE ENGINEERING

- Readmission
  - 0 : Readmission > 30 or none
  - 1 : readmission <30
- Change
  - 0 : No
  - 1 : Change
- diabetesMed
  - 0: No
  - 1: Yes
- Age
  - [10,20,30 40,50, 60, 70, 80, 90,100]

DIAG_1, DIAG_2, DIAG_3

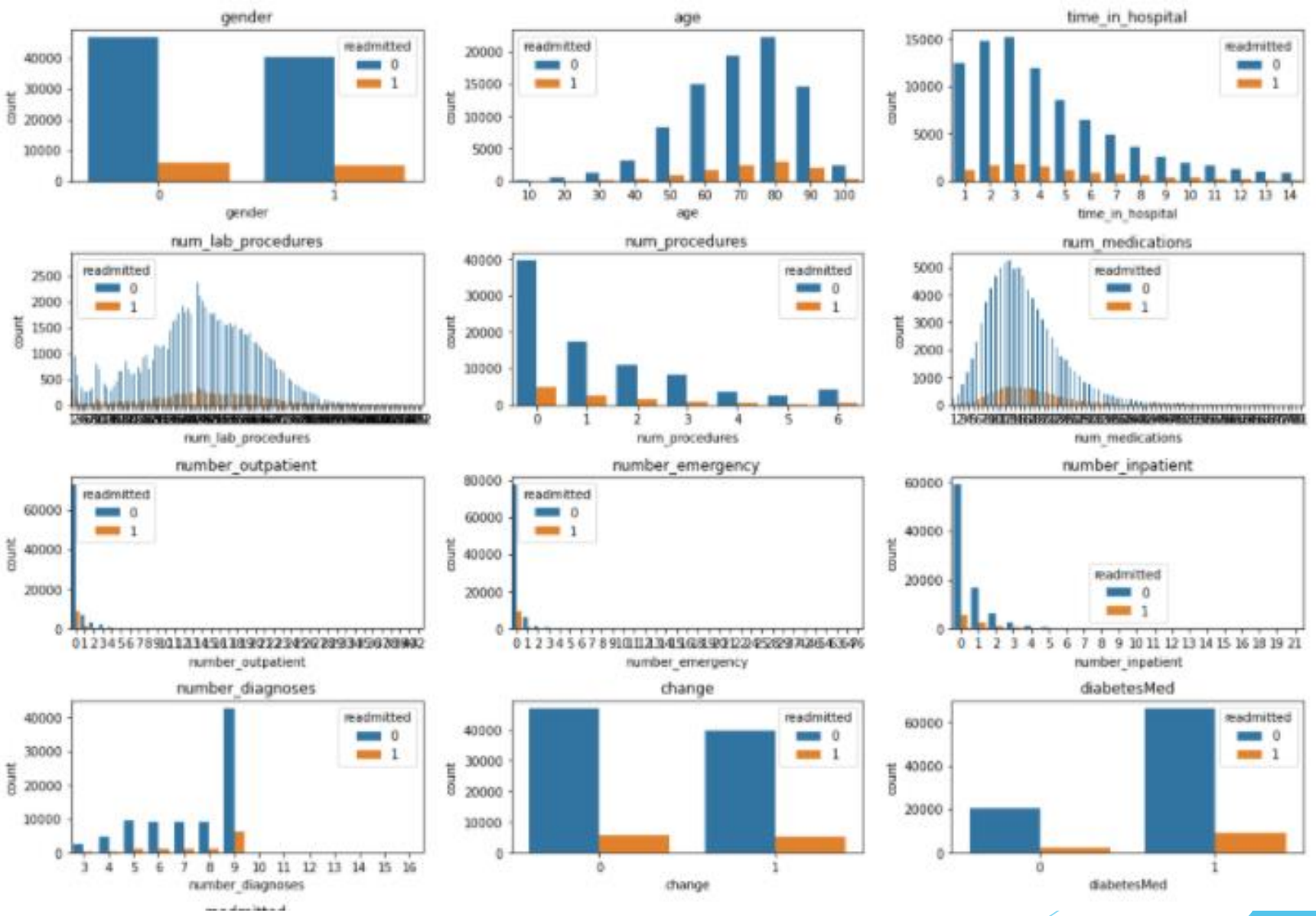
- ICD9 CODES FOR DİSEASES,
- CİRCULATORY 390–459, 785
- RESPİRATORY 460–519, 786
- DİGESTİVE 520–579, 787
- DİABETES 250.XX
- INJURY 800–999
- MUSCULOSKELETAL 710–739
- GENİTOURİNARY 580–629, 788
- NEOPLASMS 140–239
- OTHERS (17.3%)

- OTHER VARIABLES ARE CONVERTED TO DUMMIES
- POLYNOMIAL FEATURES
- RECURSIVE FEATURE ELIMINATION

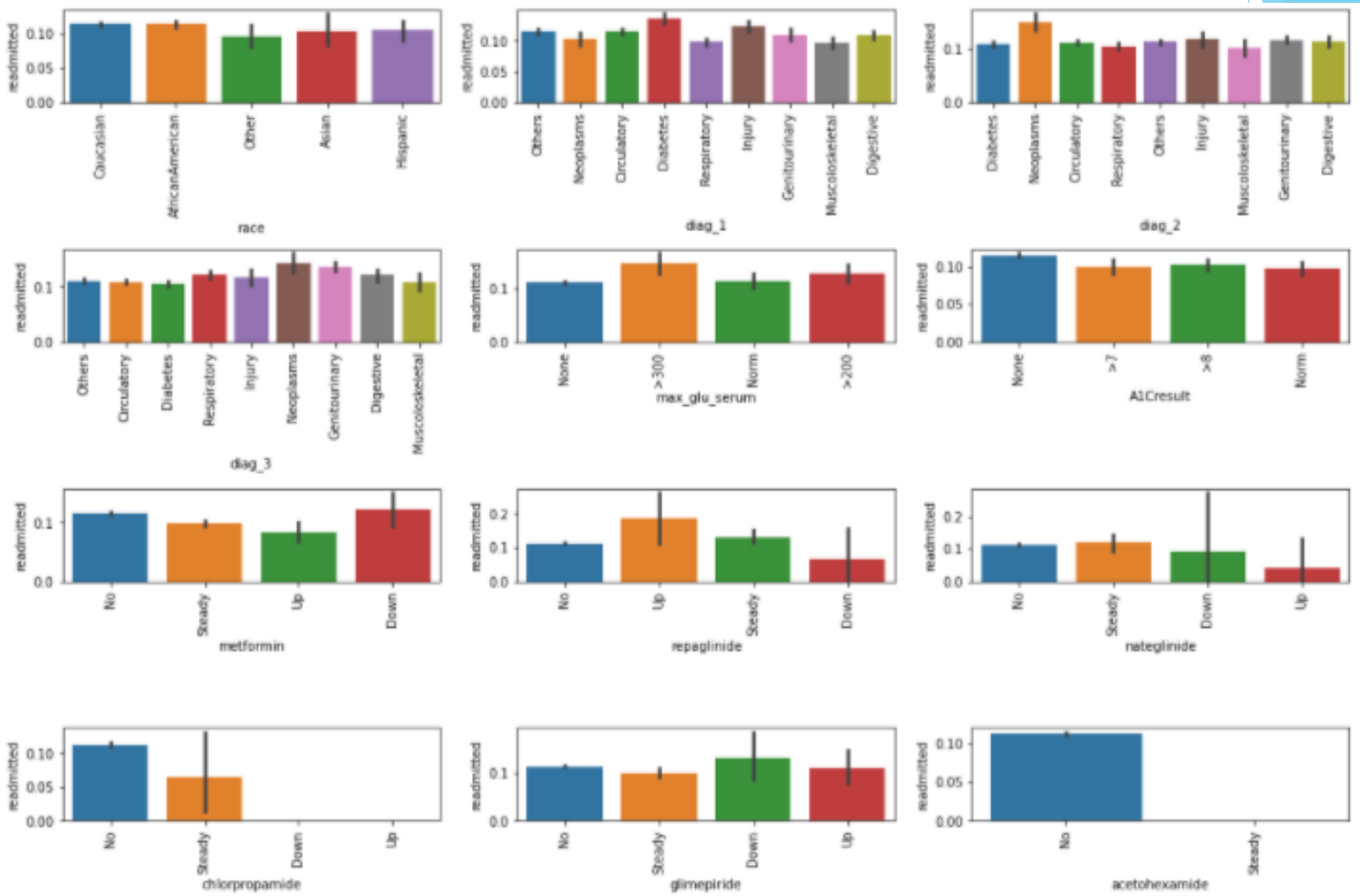
Number of features after cleaning data before modeling

DATA CLEANING			
	Before	After	After Dummies
Variables	50	39	120
Rows	101.766	98.052	98.052

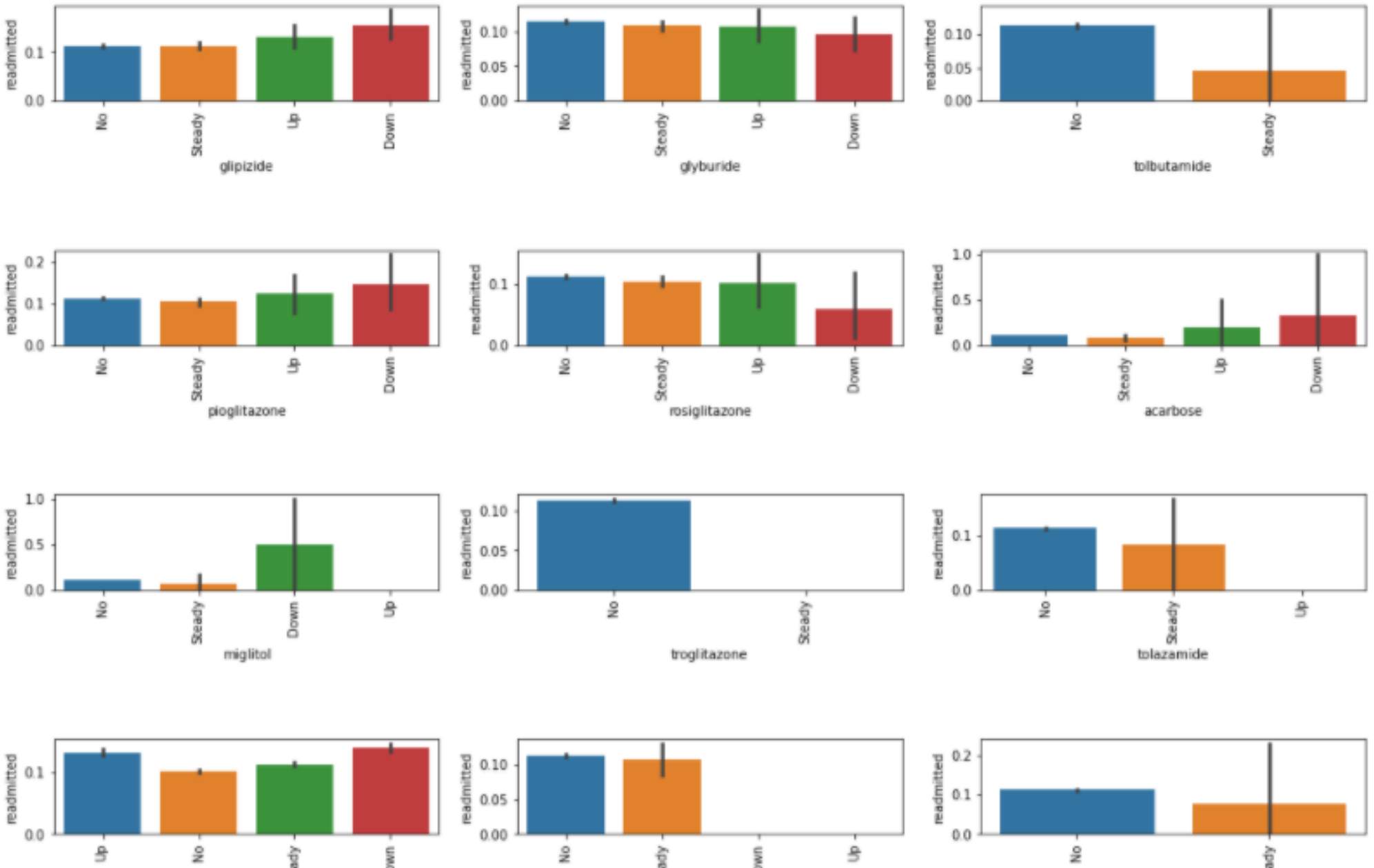
Continuous variables



Categorical variables -1



Categorical variables -2



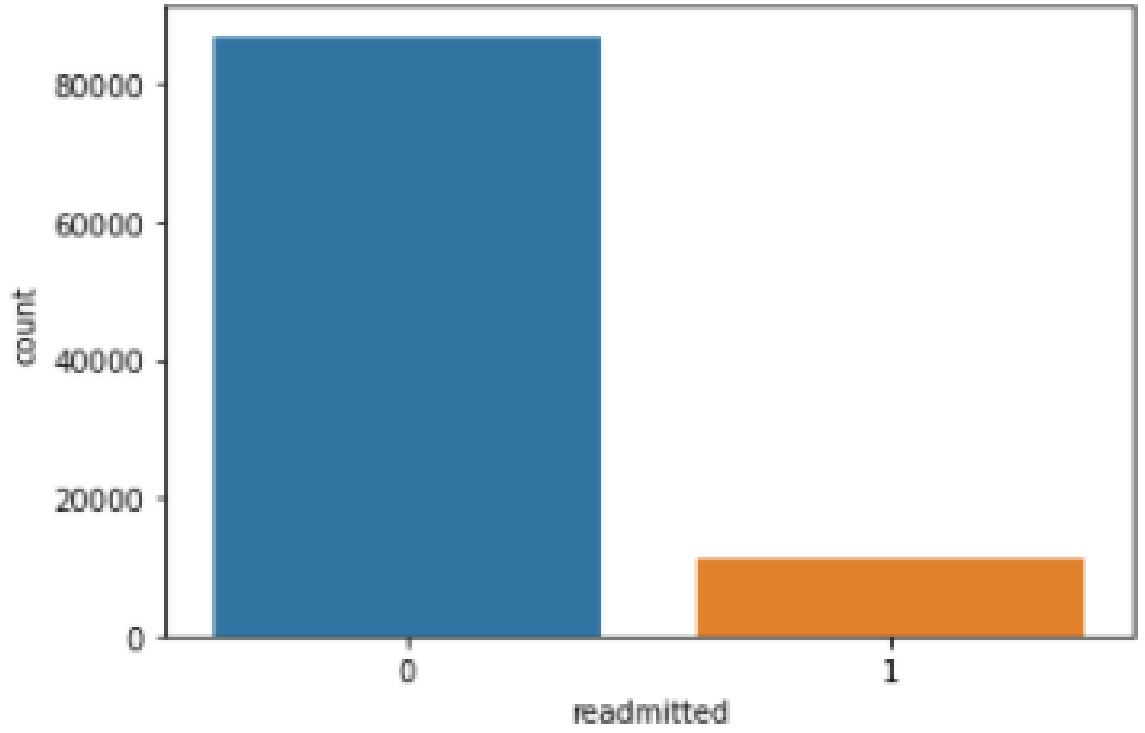
Modelling - Logistic Regression

Train Dataset- Classification Report			
	Precision	Recall	F1-score
0	0.89	1.00	0.94
1	0.49	0.02	0.03
accuracy	0.89		

Test Dataset- Classification Report			
	Precision	Recall	F1-score
0	0.89	1.00	0.94
1	0.47	0.01	0.03
accuracy	0.89		

Imbalanced Data Problem

```
Patients readmitted in 30 days ratio %11.29
Patients nicht readmitted in 30 days ratio    : %88.71
```



Modelling - SMOTE

Train Dataset- Classification Report			
	Precision	Recall	F1-score
0	0.88	0.98	0.92
1	0.97	0.86	0.91
accuracy	0.92		

Test Dataset- Classification Report			
	Precision	Recall	F1-score
0	0.88	0.98	0.93
1	0.98	0.86	0.92
accuracy	0.92		

Feature selection

- Currently model have 210 variables
- Scikitlearn Polynomial Features
- Scikitlearn feature_selection –RFE

RFE			
N	25	20	15
Accuracy	0.92	0.89	0.88
Features	diag_1_Genitourinary', 'diag_1_Muscoloskeletal', 'diag_1_Neoplasms', 'diag_1_Respiratory', 'diag_2_Circulatory', 'diag_2_Diabetes', diag_2_Digestive', diag_2_Genitourinary', 'diag_2_Injury', 'diag_2_Muscoloskeletal', 'diag_2_Others', 'diag_2_Respiratory', diag_3_Circulatory', 'diag_3_Diabetes', 'diag_3_Digestive', 'diag_3_Genitourinary', 'diag_3_Injury', 'diag_3_Muscoloskeletal', 'diag_3_Neoplasms', 'diag_3_Others', 'diag_3_Respiratory', 'insulin_Down', 'insulin_No', insulin_Steady', 'insulin_Up'	diag_1_Circulatory', 'diag_1_Diabetes', 'diag_1_Digestive', 'diag_1_Genitourinary', 'diag_1_Injury', 'diag_1_Muscoloskeletal', 'diag_1_Neoplasms', 'diag_1_Others', 'diag_1_Respiratory', 'diag_3_Circulatory', 'diag_3_Diabetes', 'diag_3_Digestive', 'diag_3_Injury', 'diag_3_Muscoloskeletal', 'diag_3_Others', 'diag_3_Respiratory', 'insulin_Down', 'insulin_No', 'insulin_Steady','insulin_Up'	diag_2_Muscoloskeletal', 'diag_2_Respiratory', 'diag_3_Circulatory', 'diag_3_Diabetes', 'diag_3_Digestive', 'diag_3_Genitourinary', 'diag_3_Injury', diag_3_Muscoloskeletal', 'diag_3_Neoplasms','diag_3_Others', 'diag_3_Respiratory', 'insulin_Down', 'insulin_No','insulin_Steady', 'insulin_Up'

Comparison of accuracy score and KNN and Logistic Regression (LR) models

Accuracy Results

	KNN, k=4	LR
Accuracy	0.883	0.92

CONCLUSION

- We went to the best of our ability to clean up the data. We have tried to feature elimination and used RFE and we have determined that second and third level diagnose results are the best predictive features. In our data, we have faced the imbalanced data and used the SMOTE method to eliminate the imbalance problem. We modeled the data with logistics regression and K-Neigbours-Network (KNN). We have predicted any patient who will be readmitted in 30 days after having treatment in hospital and discharged. The test scores are compared and in our model logistic regression gave 0,92 and KNN gave 0,88 respectively.