Prueba Técnica iFood

Oscar Chavarriaga

Tratamiento de datos

Se realizó un análisis exploratorio inicial para establecer cuáles variables podrían ser descartadas, no se tuvieron en cuenta:

- examide: medicamento que no fue suministrado a nadie
- citoglipton: medicamento que no fue suministrado a nadie
- weight: más del 90% de valores perdidos
-

Tratamiento de datos

Para los medicamentos, se realizó una transformación de los valores a 1 o 0, dependiendo si el medicamento fue suministrado o no a un paciente. Esto, debido a que los casos en los que al medicamento se le aumentó o disminuyó la dosis fueron muy pocos.

Tratamiento de datos

Una observación por paciente!!!

Esto debido a que la mayoría de las variables que tenemos disponibles (exceptuando las demográficas) están asociadas directamente a la visita del paciente. La proporción de pacientes repetidos es considerable XXX, esto tendría un efecto en el cuál le estaríamos dando más importancia a los valores de las variables demográficas a los cuales están asociados más clientes repetidos.

Nota: Cuando conservamos una observación por paciente puede que nos quedemos con aquella observación en la que tiene estado readmitido=0.

Tratamiento de datos

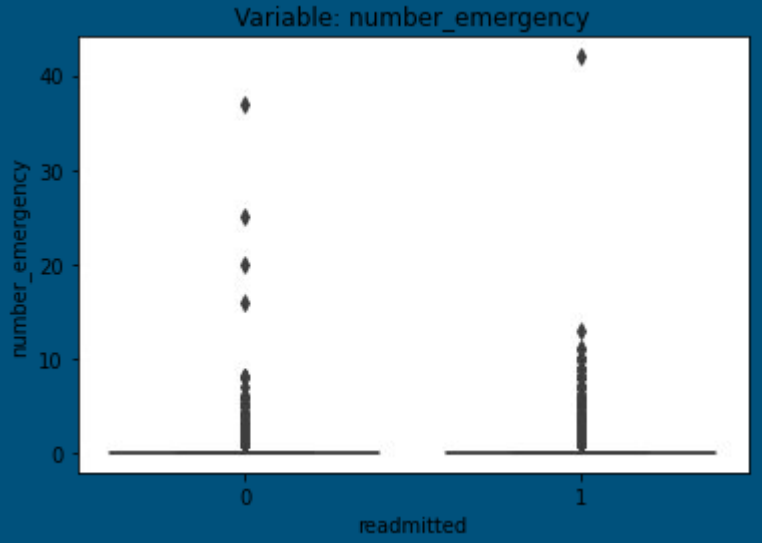
Por el momento se decidió quitar las variables categóricas de los diagnosticos: diag_1 (848 posibles valores), diag_2 (923 posibles valores), diag_3 (954) debido a su alta cardinalidad. Existen muchísimos posibles diagnósticos, para los cuáles si deseamos tenerlos en cuenta de manera desagregada, tendríamos un total de 2725 características que se incluirían en el modelo.

Con la ayuda de un experto se podría identificar cuáles son los diagnósticos (o agruparlos) que puedan ser de más ayuda para identificar si una visita es una readmisión o no.

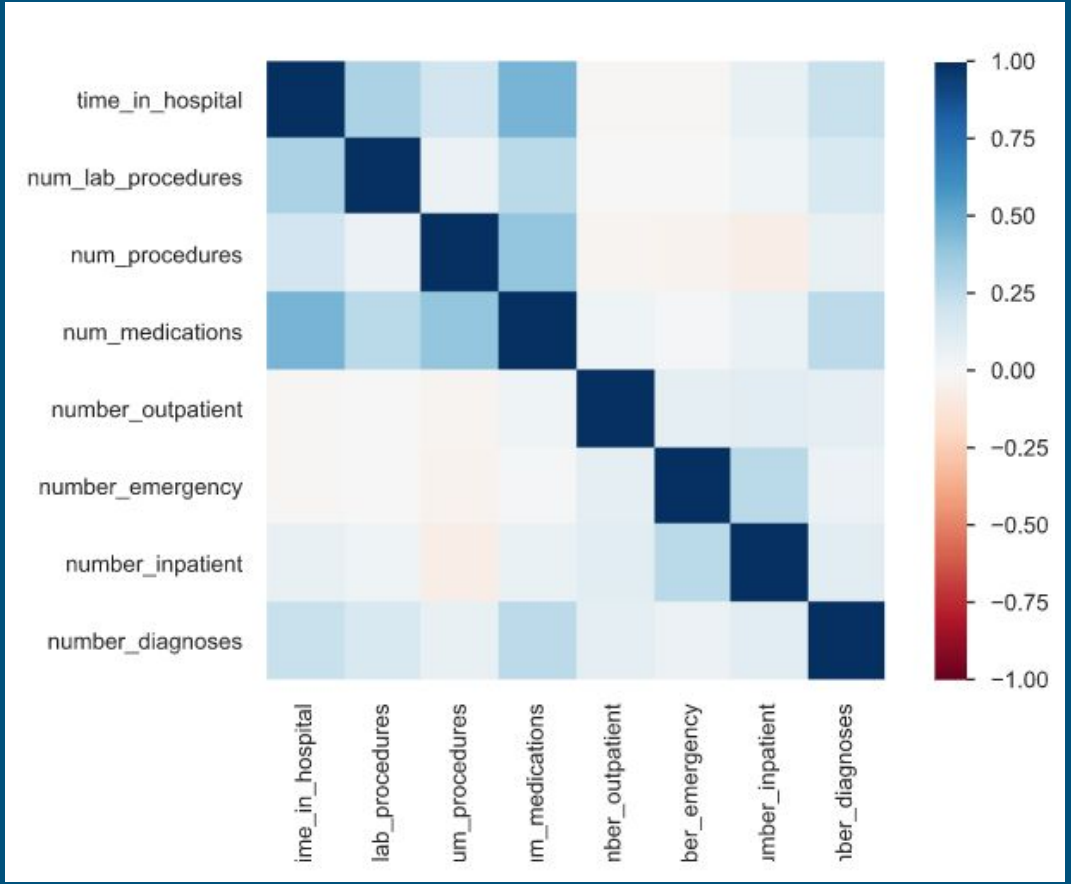
Tratamiento de datos

Las siguientes variables contienen outliers pero es debido a que tienen gran cantidad de ceros, una eliminación de atípicos podría llevar a una reducción de la variable a valores únicos en cero.

- number_outpatient (83.6% ceros)
- number_emergency (88.8% ceros)
- number_inpatient (66.5% ceros)



Matriz de correlaciones



Modelamiento - Baseline

```
from sklearn.dummy import DummyClassifier
dummy_clf = DummyClassifier(strategy="uniform")
dummy_clf.fit(X_train_scaled, y_train)
dummy_clf_train_predicted = dummy_clf.predict(X_train_scaled)
dummy_clf_test_predicted = dummy_clf.predict(X_test_scaled)
```

```
Accuracy - train:0.5054066147134494 - test:0.5026845637583892
Precision - train:0.4042006479722936 - test:0.40233105457805673
Recall - train:0.5075757575757576 - test:0.502168135403553
f1_score - train:0.4500279868151005 - test:0.4467396714783475
```

Modelamiento - RandomForest

```
{'n_estimators': [450, 483, 516, 550],
'max_features': ['sqrt'],
'max_depth': [40, 46, 53, 60],
'min_samples_split': [20, 25, 30, 40],
'min_samples_leaf': [15, 20, 30],
'bootstrap': [True]}
```

```
{'n_estimators': 483,
 'min_samples_split': 40,
 'min_samples_leaf': 20,
 'max_features': 'sqrt',
 'max_depth': 50,
 'bootstrap': True}
```

```
Accuracy - train:0.66339162534024 39 - test:0.64121923937360 18
Precision - train:0.7257561372575614 - test:0.6568376068376068
Recall - train:0.250233819678264 14 - test:0.21499510421037907
f1_score - train:0.37215286712800366 - test:0.32395405206028033
```

Importancia de variables