Machine Learning in Medicine

Amir Souri
souri.amir111@gmail.com

December 2020

Contents

1. *Introduction*

Machine learning has become an attractive field to study that facilitates decision making. It is a means of pattern recognition without programming that can help to predict or classify based on the dataset at hand. Machine learning techniques are used in lots of area e.g. business [1], healthcare [2], and computer vision [3]. This project focus on one of the application of Machine Learning in healthcare.

1.1 Motivation

Put Machine Learning into practice by applying it on a real-world dataset in healthcare to achieve experience instead of only learning the theory and work on toy datasets. In this way, one will see other aspects of Machine Learning in practice such as, the data preparation that is key to get an accurate prediction or the need of data visualization to make it easier to identify trends, patterns, and outliers within large data sets. The dataset used in this project is, an imbalanced dataset. This kind of dataset are ubiquitous, such as those in medical diagnoses, intrusion detection, and credit ratings. Therefore, dealing with this kind of dataset is a rewarding practice.

1.2 Objectives

The main objective is, train a classifier based on the dataset at hand that predicts whether a given drug will be authorised by European Medicines Agency (EMA). The secondary objectives are as follows:

Understanding the data:
Having a dataset, try to increase the domain knowledge as much as possible. Understanding each feature, the target variable, etc.

Data preparation:
Garbage in, garbage out [4]. With this in mind, make data clean before fitting a model. Digging how to link everything together to achieve the outcome. Selecting the features to include in the model. Splitting the data to train and validation set.

Data visualization:
Exploring the data by building graphs. Make it possible to dig into the graphs at any time and answer any question someone may have about a given insight.

Data modeling:
Choosing a proper model for the problem based on the data and task. Training the model by the training set. Making sure the model is not overfitting or underfitting. Adapting the model to increase the accuracy.

Metrics:
Different evaluation metrics are used for different kinds of problems. Choosing the metric that is completely depends on the type of model and the dataset. Evaluating the model on the validation set using a fair metric.

2. *Data Preprocessing*

In this chapter, the structure and properties of the dataset, how to clean it, the way of visualizing it, the method used to feature selection are addressed. The plots are provided as well.

2.1 Dataset

The dataset is the European public assessment reports (EPAR) that is obtained from European Medicines Agency [5]. It is a full scientific assessment reports of medicines authorised at a European Union level. It is an XLSX file that contains 30 columns and 1571 rows.

Some columns are self explained. However, the columns and their definition are as follows:

1. Category. It indicates the given drug is used for humans or animals.
2. Medicine name. Drug name.
3. Therapeutic area. Disease area in which a drug is used.
4. International non-proprietary name (INN) / common name. The unique drug name that is globally recognized.
5. Active substance. The substance that causes the activity of a drug.
6. Product number. It is used to identify the individual medicine package (product name, form, strength and package size).
7. Patient safety. Patient safety is the absence of preventable harm to a patient during the process of health care and reduction of risk of unnecessary harm associated with health care to an acceptable minimum [6].
8. Authorisation status. The state of a drug (Authorised - Withdrawn - Refused - Suspended) which defines the classification classes.
9. ATC code. It is a unique code that encodes the organ a drug works on and how it works.

10. Additional monitoring. If the drug is being monitored even more intensively than other medicines. **11. Generic.** It indicates if the drug is a generic drug (copies of synthetic drugs) rather than the brand-name version.
12. Biosimilar. It indicates if the drug is a biosimilar (modeled after drugs that use living organisms as important ingredients) drug.
13. Conditional approval. It indicates if the drug is approved conditionally.
14. Exceptional circumstances. A type of marketing authorisation granted to medicines where the applicant is unable to provide comprehensive data on the efficacy and safety under normal conditions of use, because the condition to be treated is rare or because collection of full information is not possible or is unethical [7].
15. Accelerated assessment. It indicates if the drug is granted an accelerated assessment. If the CHMP decides the product is of major interest in public health it may be granted an accelerated assessment.
16. Orphan medicine. It indicates if the drug is used for rare diseases.
17. Marketing authorisation date.
18. Date of refusal of marketing authorisation.
19. Marketing authorisation holder/company name.
20. Human pharmacotherapeutic group.
21. Vet pharmacotherapeutic group. Veterinary pharmacotherapeutic group
22. Date of opinion.
23. Decision date.
24. Revision number.
25. Condition / indication.
26. Species. Animal Species.
27. ATCvet code. ATC code for animals.
28. First published.
29. Revision date.
30. URL.

2.2 Data cleaning

The Excel file contained information (8 columns) about EMA that was skipped during the loading. After loading the file the number of rows (instances) reduced to 1562. A lots of columns were removed since they were redundant for this study e.g *'Date of opinion', 'Decision date'*. A few of them were removed since they did not seem to be a feature (input/attribute)

e.g *'Product number', 'ATC code'*. I selected nine columns to be the feature set besides one column as the response value (independent variable). They are listed below. The selection of the input feature set was based on my preknowledge. The lack of an expert in the domain knowledge is manifest here.

Input feature set: Category, Patient safety, Additional monitoring, Generic, Biosimilar, Conditional approval, Exceptional circumstances, Accelerated assessment, Orphan medicine.

Response value: Authorisation status.

The *data* did not contain any duplicate or missing values which is a fortunate. The variables that the input feature set and response value take are as follow:

Category: ['Human' 'Veterinary']
Patient safety: ['no' 'yes']
Additional monitoring: ['no' 'yes']
Generic: ['no' 'yes']
Biosimilar: ['no' 'yes']
Conditional approval: ['no' 'yes']
Exceptional circumstances: ['no' 'yes']
Accelerated assessment: ['no' 'yes']
Orphan medicine: ['no' 'yes']
Authorisation status: ['Authorised' 'Withdrawn' 'Refused' 'Suspended']

They are all categorical data that should be transform into the suitable numeric values. To do this, the *Human* was replaced with 1, *Veterinary* with 0, all *no* with 0, and all *yes* with 1.
The instance labeled as *Suspended* was removed since it is only one sample and it won't impact the result. The *Authorised* was replaced with 1 and the rest (*Refused, Withdrawn*) with 0 to be considered as the *Unauthorized* class label. Therefore, it became a binary classification problem.

2.3 Data visualization

All the features are binary variables. Frequency distribution and relative frequency distribution of the target variable is shown in Figure 2.1 and 2.2 respectively. The relative frequency distribution has represented in percent-

age.

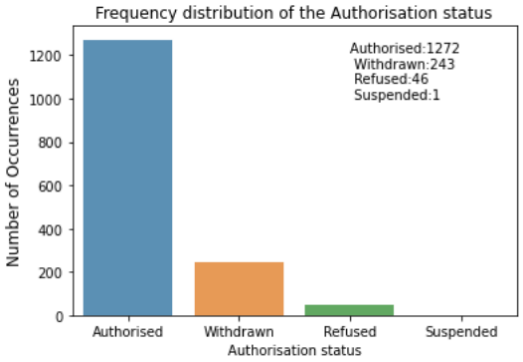


Figure 2.1: Frequency distribution

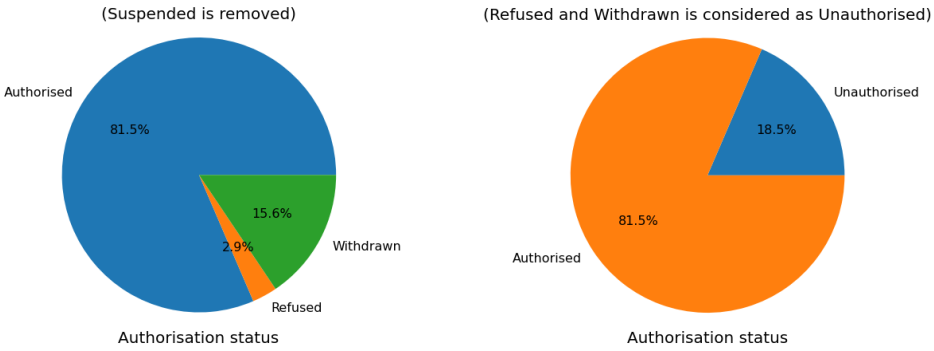


Figure 2.2: Relative frequency distribution

For quantitative analysis of the relationship between multiple binary variable each feature is plotted using a helper function, *plot_cat_feature*. It takes a data frame, a feature variable, and a target variable afterward normalizes the feature variable and plot it against the target variable. The plots are shown in Figure 2.3. Although the plots are useful to see how the variables are distributed, it is heavy-duty to observe the correlation between variables. Therefore, it is much easier to look at the Pearson's correlation coefficient that is shown in Figure 2.4. Strictly speaking, the Phi Coefficient, aka Matthews Correlation Coefficient. In fact, a Pearson's correlation coefficient estimated for two binary variables will return the phi coefficient [1].

[1]https://en.wikipedia.org/wiki/Phi_coefficient

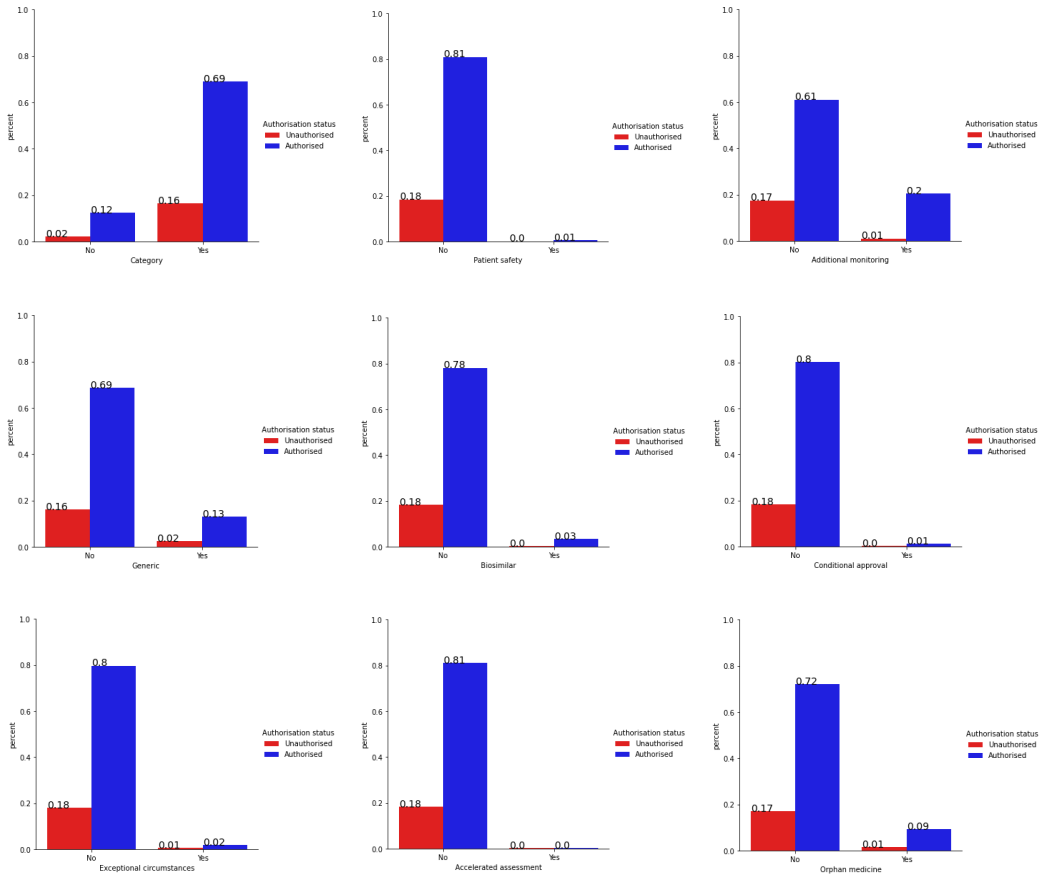


Figure 2.3: Quantitative analysis for imbalanced data

	Category	Patient safety	Additional monitoring	Generic	Biosimilar	Conditional approval	Exceptional circumstances	Accelerated assessment	Orphan medicine
Category	1.000000	0.007411	0.215631	0.089022	0.080306	0.048171	0.028438	-0.173990	0.143256
Patient safety	0.007411	1.000000	0.083851	-0.032299	-0.014825	-0.008893	-0.011865	-0.005466	-0.026446
Additional monitoring	0.215631	0.083851	1.000000	-0.208688	0.197723	0.196305	0.185246	-0.037518	0.236377
Generic	0.089022	-0.032299	-0.208688	1.000000	-0.082570	-0.049529	-0.066087	-0.030442	-0.147295
Biosimilar	0.080306	-0.014825	0.197723	-0.082570	1.000000	-0.022733	-0.030333	-0.013972	-0.067607
Conditional approval	0.048171	-0.008893	0.196305	-0.049529	-0.022733	1.000000	-0.018195	-0.008381	0.228596
Exceptional circumstances	0.028438	-0.011865	0.185246	-0.066087	-0.030333	-0.018195	1.000000	0.165746	0.258430
Accelerated assessment	-0.173990	-0.005466	-0.037518	-0.030442	-0.013972	-0.008381	0.165746	1.000000	-0.024925
Orphan medicine	0.143256	-0.026446	0.236377	-0.147295	-0.067607	0.228596	0.258430	-0.024925	1.000000

Figure 2.4: Correlation within imbalanced data

2.4 Feature selection

Before feature selection the data must be split to a train set and held-out validation set to prevent data leakage, hence the model will not bias the performance analysis. Therefore, the cleaned dataset was divided into two different sets so that 25% of the whole dataset considered as the validation set and the rest as the training set. A *random_state* (seed) was set for reproducible output across multiple calls.

Selecting a subset of relevant features from the input feature set to be used for training the model has several advantages. Training with a proper selected subset, the accuracy will be the same or slightly different ($\pm$) than using the all input feature set. Instead, it reduces the complexity of the model and reduces the consumption of memory and time. Moreover, it might reduce overfitting.

There are many different techniques for feature selection. One can use a well-known method, Pearson's correlation coefficient (Phi Coefficient) to select the most correlated features with the target variable to gain the subset of relevant features. e.g. the features marked with a red rectangle in Figure 2.5. If two variables are highly correlated, one of them can be used to predict the other one. Therefore, one generally looks for features that are highly correlated with the target variable.

Although the feature selection is often straightforward when working with real-valued data, such as using the Pearson's correlation coefficient, but can be challenging when working with categorical data, [Machine Learning Mastery] [8]. There is two more robust methods that can be used to feature scoring when dealing with categorical predictor and target variable, namely *chi-squared statistic* and the *mutual information statistic*.

	Category	Patient safety	Additional monitoring	Generic	Biosimilar	Conditional approval	Exceptional circumstances	Accelerated assessment	Orphan medicine	Authorisation status
Category	1.000000	-0.003622	0.222309	0.084216	0.080652	0.045997	0.045149	-0.156739	0.146493	-0.033441
Patient safety	-0.003622	1.000000	0.048399	-0.030513	-0.013854	-0.007901	-0.010609	-0.004704	-0.025163	0.003625
Additional monitoring	0.222309	0.048399	1.000000	-0.208868	0.207651	0.187946	0.206533	-0.034845	0.249481	0.175101
Generic	0.084216	-0.030513	-0.208868	1.000000	-0.082009	-0.046771	-0.062800	-0.027843	-0.148957	0.038734
Biosimilar	0.080652	-0.013854	0.207651	-0.082009	1.000000	-0.021235	-0.028513	-0.012641	-0.067630	0.056940
Conditional approval	0.045997	-0.007901	0.187946	-0.046771	-0.021235	1.000000	-0.016261	-0.007210	0.188075	0.032474
Exceptional circumstances	0.045149	-0.010609	0.206533	-0.062800	-0.028513	-0.016261	1.000000	0.080927	0.289048	0.009990
Accelerated assessment	-0.156739	-0.004704	-0.034845	-0.027843	-0.012641	-0.007210	0.080927	1.000000	-0.022961	-0.002301
Orphan medicine	0.146493	-0.025163	0.249481	-0.148957	-0.067630	0.188075	0.289048	-0.022961	1.000000	0.076298
Authorisation status	-0.033441	0.003625	0.175101	0.038734	0.056940	0.032474	0.009990	-0.002301	0.076298	1.000000

Figure 2.5: Correlation between features and the target variable

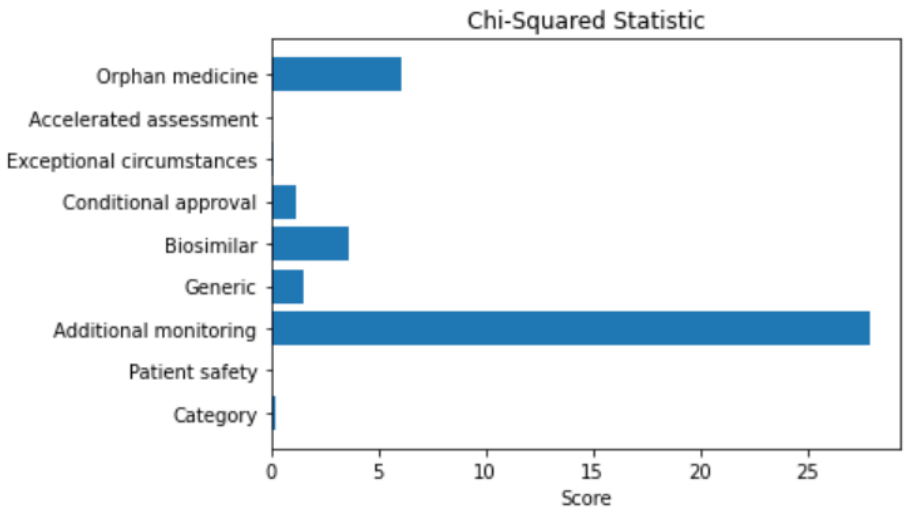


Figure 2.6: Scores of features using Chi-Squared

Moreover, [Khaldy MAI, Kambhampati C] [9] discusses and experiments on resampling imbalanced class and the effectiveness of feature selection methods on both imbalanced and balanced (resampled) class. Their experiments show that the resampling data has highly improved only the information gain method (the calculation is referred to as mutual information [2]). All other methods reduces their performance, or produced the same performance as the original data (imbalanced class). Therefore, the *chi-squared* is used for feature scoring in this project.

A wrapper function (select_cat_features_chi) is provided to calculate *chi-squared statistic* and plot the scores. It takes a feature set and a target variable to plot all scores. The higher score, the stronger dependency. The plot is shown in Figure 2.6.

One can select k highest scored features to be the subset of feature. The question is how to choose k. It should be chosen such that the selected subset of features achieves a classification accuracy that is as good or ideally better than the accuracy achieved by the feature set. The accuracy refers to the performance that a classifier trained on the train set can reach in prediction based on the validation set.

The accuracy of the model trained by all the features was 0.60. Selecting the 6 highest scores features the accuracy became 0.59. Adding more features only improved the accuracy by 0.01 that is trivial.

The selected subset of feature is *['Additional monitoring', 'Orphan medicine', 'Biosimilar', 'Generic', 'Conditional approval', 'Category']* .

[2]https://machinelearningmastery.com/information-gain-and-mutual-information/

3. *Methodology*

In this chapter, all the models trained in this project and the optimal model are listed. How to split the data, how to select the optimal model, how the model trained and the approaches applied to overcome imbalanced data, the methods for feature selection, and which metrics to use are all discussed. Also, the final results are reported.

3.1 Model

REPLASMENT FALSE

This project leveraged Scikit-Learn [10] and TensorFlow [11] to experiment with four distinguish models as follows:

- Logistic Regression
- Random Forest Classifier
- Support Vector Classification
- Neural Network

The dataset was divided into two parts. One part (training set) was used for training the models and the other one (validation set) to test the generalization ability. The model that is the most accurate on the validation set is the best one. This process is called cross-validation [12].

Although the models have almost the same accuracy the Logistic Regression model trained with down-sampled data is the best one with optimal complexity in this project.

The accuracy of all the models on the train and validation set is provided in table 3.1 to check overfitting and underfitting. All the models except the first two in the table 3.1 seem to be right. The first two models are a little tricky.

Apart from the fact that they are biased models, their validation accuracy are higher than their training accuracy. It usually means the training data is harder to model than the validation data. If the seed was set e.g. to 85 their training and validation accuracy would be 0.83 and 0.78 respectively.

Accuracy		
Model	training	validation
Logistic Regression (imbalanced)	0.814	0.818
Random Forest Classifier (imbalanced)	0.815	0.821
Logistic Regression (Down-sampled)	0.603	0.596
Logistic Regression (Up-sampled)	0.638	0.596
Support Vector Classification (Down-sampled)	0.608	0.588
Neural Network (imbalanced data)	0.62	0.6

Table 3.1: Model Accuracy

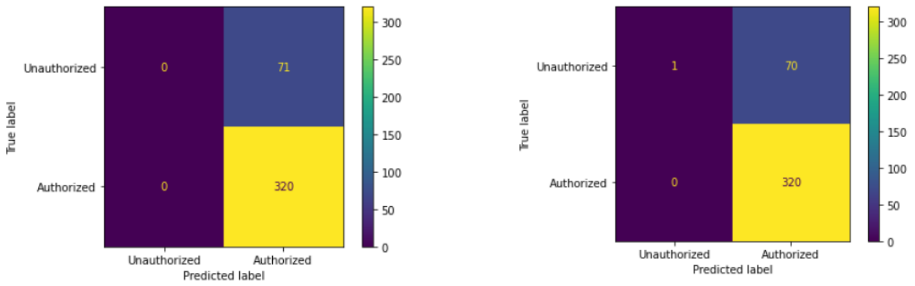
3.2 Training Processing

A Logistic Regression model fit on the given training set. The accuracy of the validation was 0.82. That is great for real-world data! Since the accuracy is not always a reliable metric, the confusion matrix was plotted which is shown in Figure 3.1a. As you noticed, the model classified all the instances as class *Authorised*. Let's take a step back and look at the data that was used to train the model. The number of *Authorised* instances was 1272 and for *Unauthorised*, 289. It is imbalanced dataset (classes are not represented equally) hence leads to get a biased classifier.

The following are a series of approaches and decisions one can carry out to overcome the imbalanced dataset. [13].

- Collecting more data
- Try different algorithms
- Re-sample the dataset

Collecting more data is not possible in this project since all the data that (MEA) has been published [5] is used. What's more, collecting more data is a time-consuming process and is not feasible in some cases. Moving to the next approach, a Random Forest Classifier trained with the imbalanced dataset. It only managed to classify one *Unauthorised* instance correctly. The confusion matrix is shown in figure 3.1b.



(a) Logistic Regression model (b) Random Forest Classifier

Figure 3.1: confusion matrix trained with imbalanced data

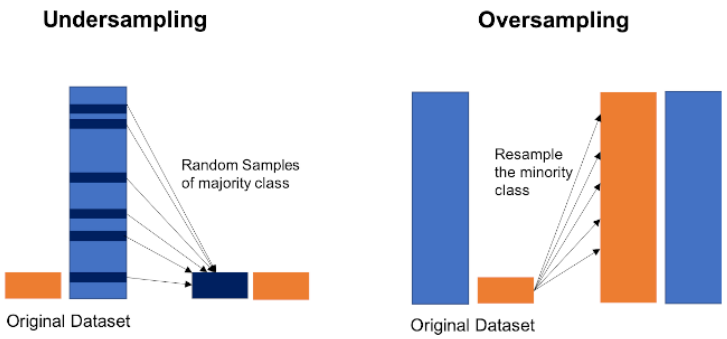


Figure 3.2: The different types of resampling. Image downloaded from here.

The next approach is, *re-sample the dataset* that is the most used approach in industries. In total 289 (number of minority class) instances was re-sampled from the majority (authorized) class without replacement. The minority class was combined with the re-sampled (down-sampled) majority class to achieve a down-sampled training set. Afterwards, it was shuffled to avoid any element of bias/patterns. For oversampling (Up-sampling) there were taken 1272 (number of majority class) samples from minority (unauthorized) class with replacement. The majority class was combined with the re-sampled (up-sampled) minority class to achieve a up-sampled training set. In the end, it was shuffled.

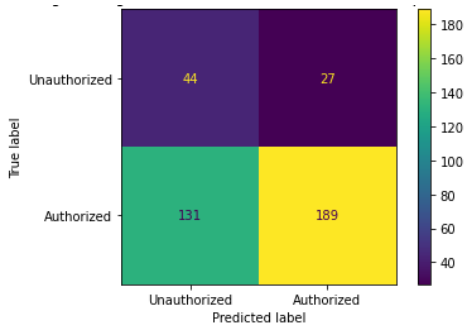
Furthermore, a support vector classifier fits to the down-sampled data to check if it improves the performance. Finally, a neural network trained with the imbalanced data to show that it is possible with *Keras* to get an unbiased model trained even with imbalanced data by passing *Keras weights* for each class through a parameter. These will cause the model to "pay more attention" to examples from an under-represented class [14].

3.3 Metric

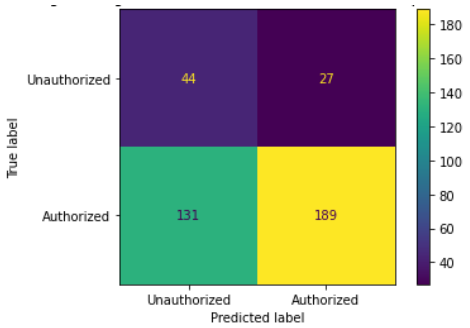
Three metrics measured the models performance, namely, accuracy, specificity (aka True Negative Rate), sensitivity (aka Recall or True Positive Rate). The wrapper function called *model_performance* takes feature vectors (inputs), corresponding target values (labels), and the model to return the metrics values. Besides, the confusion matrix for each model was plotted. The confusion matrices and metric values are shown in Figure 3.3 and Table 3.4 respectively. All the result are based on the validation (test) set.

Model	Accuracy	Specificity	Sensitivity
Logistic Regression (Down-sampled)	0.596	0.62	0.591
Logistic Regression (Up-sampled)	0.596	0.62	0.591
Support Vector Classifier (Down-sampled)	0.588	0.62	0.581
Neural Network (imbalanced data)	0.6	0.62	0.59

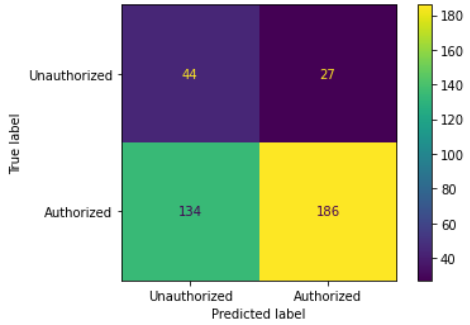
Table 3.4: Performance of the models



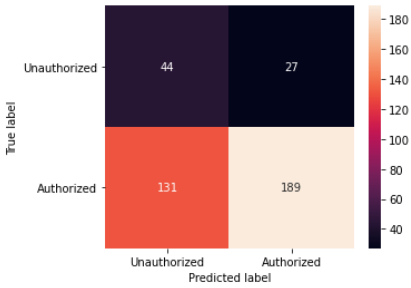
(a) Logistic Regression (Down-sampled)



(b) Logistic Regression (Up-sampled)



(c) Support Vector Classifier (Down-sampled)



(d) Neural Network (imbalanced data)

Figure 3.3: Confusion matrices

4. *Conclusion*

This experiment shows that the Logistic Regression trained with (Down-sampled) data is the optimal prediction model. It has the highest accuracy and lowest complexity. However, the specificity and sensitivity were computed to check if one of the models could classify more *Unauthorized* instances. This kind of model is preferred in this project since it is more useful to predict the drugs that are not going to be authorized.

The results might slightly vary if you run the code with a different seed. It is due to randomness. The portion of the training and validation set are randomly selected. Therefore, a certain seed was used to reproduce the same training and validation set. It applies to re-sampling. For the Neural Network model, you will get a slightly different result due to its stochastic nature.

4.1 Future Work

Future research should consider the potential effects of feature set more carefully. Find an expert in the domain knowledge to get information about the *Revision number* and *Therapeutic area* e.g to what extent they effect the decision and how to interpret them. Discuss with her/him about including them in the features set. Moreover, it is worth finding another dataset and combine it with the dataset at hand. One possibility is to request to access the dataset from FDA [1].

[1] https://www.fda.gov/

References

[1] "Joyce Chiu". *The Many Business Applications of Machine Learning*. Dec. 2019. URL: https://www.datacamp.com/community/blog/machine-learning-tracks.

[2] *Top 10 Applications of Machine Learning in Healthcare*. URL: https://www.flatworldsolutions.com/healthcare/articles/top-10-applications-of-machine-learning-in-healthcare.php.

[3] "Jason Brownlee". *9 Applications of Deep Learning for Computer Vision*. Mar. 2019. URL: https://machinelearningmastery.com/applications-of-deep-learning-for-computer-vision/.

[4] *Concept of (GIGO)*. URL: https://en.wikipedia.org/wiki/Garbage_in,_garbage_out.

[5] *European public assessment reports (EPARs)*. URL: https://web.archive.org/web/20181223182250/https://www.ema.europa.eu/sites/default/files/Medicines_output_european_public_assessment_reports.xlsx.

[6] "World Health Organization". *Patient safety*. URL: https://www.who.int/patientsafety/en/.

[7] "European Medicines Agency". *Exceptional circumstances*. URL: https://www.ema.europa.eu/en/glossary/exceptional-circumstances.

[8] Jason Brownlee. *How to Perform Feature Selection with Categorical Data*. 2019. URL: https://machinelearningmastery.com/feature-selection-with-categorical-data/.

[9] Kambhampati C Khaldy MAI. *Resampling Imbalanced Class and the Effectiveness of Feature Selection Methods for Heart Failure Dataset*. 2018. URL: https://medcraveonline.com/IRATJ/resampling-imbalanced-class-and-the-effectiveness-of-feature-selection-methods-for-heart-failure-dataset.html.

[10] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[11] Martin Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: `http://tensorflow.org/`.

[12] Ethem Alpaydın. *Introduction to Machine Learning Third Edition*. MIT-Press, 2014, pp. 39–40. ISBN: 978-0-262-02818-9.

[13] Numal Jayawardena. *How to Deal with Imbalanced Data*. 2020. URL: `https://towardsdatascience.com/how-to-deal-with-imbalanced-data-34ab7db9b100`.

[14] TensorFlow tutorials. *Classification on imbalanced data*. URL: `https://www.tensorflow.org/tutorials/structured_data/imbalanced_data`.