Group Assignment: Applied Deep Learning – Dr. Philippe Blaettchen

Assignment overview

When using machine learning for classification, things are easiest if classes are "balanced" – that is, when the number of observations belonging to each of the classes are of the same order of magnitude. Unfortunately, this is often not the case. In this assignment, you will work with a dataset of car-insurance claims and try to classify claims into fraudulent (1) and non-fraudulent (0). There are more than 10,000 claims in the dataset, but only around 100 are fraudulent. Nevertheless, we want to create a model that helps the insurance provider target its investigation efforts. For this, we consider two options: (i) synthetically creating new data to make the dataset more balanced, and (ii) using an auto-encoder to represent "normal" (non-fraudulent) claims and applying it to distinguish fraudulent claims – a form of anomaly detection.

Task description

1. Briefly discuss why it is more difficult to find a good classifier on such a dataset than on one where, for example, 5,000 claims are fraudulent, and 5,000 are not. In particular, consider what happens when undetected fraudulent claims are very costly to the insurance company.
2. Load the dataset "Insurance_claims.csv" and clean it as appropriate for use with machine learning algorithms. A description of the features can be found at the end of this document.
3. Start by creating a (deep) neural network in TensorFlow and train it on the data. Using training and validation sets, find a model with high accuracy, then evaluate it on the test set. In particular, record both the accuracy and AUC. Briefly discuss what issues you observe based on the metrics.

Our first approach will be to create new (synthetic) data points representing fraudulent claims and remove some of the non-fraudulent claims. This will allow us to create a more balanced dataset. The state-of-the-art method we apply is SMOTE (Synthetic Minority Oversampling Technique).

4. The file "SMOTE.ipynb" explains the process in detail and shows how to change the dataset with an example. You can copy and adjust the code to make it work within your analysis. You can adjust the "sampling_strategy" parameters as you see fit, particularly if you want to fine-tune your model in part 5.
5. Create a new (deep) neural network and train it on your enhanced dataset. Use training and validation sets derived from the enhanced dataset to find a model with high accuracy. Evaluate your final model on a test set consisting only of original data. Again, record the accuracy and AUC. Briefly discuss the changes you would expect in the metrics and the actual changes you observe. Would you say that you are now doing better at identifying fraudulent claims?

Our second approach will be to use an autoencoder to learn what "normal" (non-fraudulent) data "looks like."

6. Using the original data, create a training and set that contains only non-fraudulent claims, as well as validation and test sets that contain non-fraudulent and fraudulent claims. Make sure to spread fraudulent claims evenly across validation and test sets.
7. Using TensorFlow, create an autoencoder, ensuring that the middle hidden layer has fewer neurons than your input has features. Use training and validation sets to find a model that represents its input data well. In particular, you will want to predict your validation set observations. For each observation, you can measure the difference between the original observations and the predicted one, using, for example, the mean squared error of all features of the observation. Plot the errors for all your validation set observations in a histogram - in a good model, this error should be much higher for fraudulent claims than non-fraudulent ones.
8. Use your trained autoencoder to predict the test set and define the corresponding losses. Create a histogram of your test set claims, clearly marking fraudulent and non-fraudulent claims. Discuss how you could use this to decide whether a transaction is fraudulent or not. Can you also derive an AUC in this approach - if yes, how does it perform compared to the previous approaches?

Transparency of deep learning approaches.

9. As you know, it is difficult to understand precisely why a neural network makes a specific prediction. Discuss why this might be problematic when the neural network prediction leads to a fraud investigation by the insurance company. What alternatives can you envision that make use of the techniques we have applied and allow for more interpretability and transparency?

Bonus task (for up to 5 extra points).

10. Use your synthetically extended dataset and train a simple model, such as logistic regression or a decision tree that allows you to interpret why fraud is suspected. Keep track of the accuracy and AUC on a test set made from original data only. How does your model perform compared to the previous models you have developed? Does your model allow you to answer a customer who asks, "why am I being investigated"?

Hints

- Don’t spend much time perfecting the model in part 3. What is most important at this point is to understand the issues faced by training a model with such an uneven dataset.
- For parts 5 and 7, start by creating minimum viable products. Only once everything runs should you go back to it and see how to improve your models. The performance of your models will matter for evaluation, but not as much as having a complete answer.
- The same goes for cleaning the dataset: combine rare categories, don’t overengineer new features, and instead start with only a subset of the existing ones. Then, once everything works smoothly, you can think about enhancing your dataset.

- In your autoencoder (part 7), make sure that your network can recreate your inputs. For example, if your output layer uses tanh activation, the outputs will be between -1 and 1. Remember to scale the inputs appropriately.

Materials to submit

- A Jupyter notebook that allows recreating your solutions
- The trained models you develop in parts 3, 5, and 7, as .h5-files
- Your written answers, either within the Jupyter notebook or in a separate .pdf-file. Make sure to create numbered sections within your notebook and separate .pdf-file corresponding to the task at hand.
- A PowerPoint slide deck of up to 5 slides comparing the different approaches you have developed and discussing the main challenges you faced. Several groups will be selected randomly to present their work in the next class. The selected groups will be informed by 8 pm on the day of submission. Please keep your presentation to under 5 minutes.

Assessment

Your submission will be evaluated against four criteria:

- appropriate use of concepts and frameworks discussed in class
- effectiveness of the proposed answer/solution
- originality and creativity of the proposed answer/solution
- organization and clarity of submitted materials

Appendix: Feature descriptions for Insurance_claims.csv

Claimant-specific information	PolicyholderNumber	Unique policy number for each policyholder
	FirstPartyVehicleNumber	Vehicle number
	PolicyholderOccupation	Occupation of the policyholder
	FirstPolicySubscriptionDate	Subscription date
	FirstPartyVehicleType	Type of vehicle (car, motorcycle, …)
	PolicyWasSubscribedOnInternet	1 if policy subscribed online; 0 otherwise
	NumberOfPoliciesOfPolicyholder	Number of subscribed policies
	FpVehicleAgeMonths	Age of car at time of incident (months)
	PolicyHolderAge	Age of policyholder
	FirstPartyLiability	Percentage of first party liability covered
Incident-specific information	ThirdPartyVehicleNumber	Vehicle number of third party if applicable
	InsurerNotes	Insurer notes about the incident (free text)
	LossDate	Date of the covered incident
	ClaimCause	Cause of the incident
	ClaimInvolvedCovers	Policy covers used by the claimant
	DamageImportance	Importance of damages as assessed by expert, in case assessment took place
	ConnectionBetweenParties	Connection between parties if known

	LossAndHolderPostCodeSame	1 if postcode of incident same as postcode of policyholder; 0 otherwise
	EasinessToStage	Indicator of easiness to stage the accident (computed by insurance company)
	ClaimWihoutIdentifiedThirdParty	1 if no other party involved; 0 otherwise
	ClaimAmount	The amount of money claimed
	LossHour	Hour of the incident
	NumberOfBodilyInjuries	Number of injured persons in the incident
	Fraud	1 in the case of fraud; 0 otherwise