Trabalho Prático Fase 2 - Treino e Comparação de Modelos

Nuno Gomes*, Ricardo Ramos†, Rafael Carvalho‡

DEETC - MEIC
Instituto Superior Engenharia de Lisboa
Lisboa, Portugal
*A18364@alunos.isel.ipl.pt, †A46638@alunos.isel.ipl.pt, ‡A47663@alunos.isel.ipl.pt

***Resumo*—O trabalho prático foi dividido em duas partes: A primeira, referente à formulação de possíveis questões e respostas que se poderiam formular a partir de um conjunto de dados e, consequentemente, ao pré-processamento desse mesmo conjunto de dados para viabilizar respostas a essas questões. Daqui resultou a formulação do problema objetivo de proceder à correta classificação de um município como industrial ou residencial com base na informação do seu consumo energético, assim como, através da predição utilizando um regressor, prever o consumo energético em Lisboa face às condições meteorológicas. A segunda parte completa o trabalho prático, tendo por objetivo a aplicação e compreensão de métricas de performance sobre os modelos de aprendizagem automática para classificação e regressão, e a influência da aplicação de técnicas de amostragem dos dados sobre os resultados obtidos, visando as questões do problema objetivo.**

Neste documento demonstram-se os resultados obtidos aplicando diferentes técnicas de amostragem a diversos modelos de aprendizagem automática para classificação, assim como para o modelo regressor, reforçando a inviabilidade de aplicar amostragem neste caso.

***Index Terms*—Megadados, Amostragem, Aprendizagem Automática, Classificação, Regressão**

I. Introdução

Numa era tecnológica onde predomina a abundância de dados, a utilização de modelos de aprendizagem automática viu um rápido crescimento, permitindo extrair informação que outrora escapariam a abordagens analíticas. Desta forma, tornou-se necessário obter e preparar os dados para fornecê-los a algoritmos de aprendizagem automática, para que estes possam tomar decisões e obter soluções para problemas definidos em requisitos funcionais.

Com vista a compreender a modelação de dados e descobrir problemas que possam ser respondidos pelos mesmos, na primeira fase do trabalho o grupo formulou duas questões:

1) "É possível inferir se um município é predominantemente industrial ou residencial com base na evolução do consumo energético ao longo do dia?
2) "Conseguimos prever o consumo energético atendendo às condições meteorológicas previstas para uma determinada data?"

Com as duas questões em mente, foram formulados dois conjuntos de dados através de diferentes métodos de *Feature Selection* e *Feature Reduction* para aplicar a diversos algoritmos de aprendizagem automática, utilizando diversos tipos de amostragem para balanceamento dos dados de treino.

Na aprendizagem automática, a escolha do modelo certo para um determinado conjunto de dados é complexa, uma vez que não depende apenas do tipo de algoritmo mas também da escolha do modelo que equilibra o desempenho em termos de objetivo final, eficiência computacional e a sua interpretabilidade. A abordagem mais típica na comparação de modelos envolve começar por modelos mais simples, chamados de *baseline models*, e compará-los com modelos mais complexos, procurando sempre manter o modelo mais simples e com melhor compromisso entre desempenho e eficiência [1].

O desempenho dos modelos pode ser medido com um conjunto de métricas, que permitem monitorizar e avaliar a *performance* dos modelos durante a fase de treino e de teste. Atendendo que os problemas de *Machine Learning* podem ser reduzidos a classificação ou regressão, existem diferentes métricas associadas a cada um. Para a regressão, uma vez que o *output* é contínuo, as métricas associadas

são calculadas com base na distância entre o previsto (*predicted*) e o considerado verdadeiro (*ground truth*). Para classificação, compara-se o desempenho num problema discreto, com métricas tipicamente associadas ao número de previsões corretas (*accuracy*), o rácio entre verdadeiros casos positivos e os classificados (*precision*), entre outras, a serem mostradas. [2]

Na segunda fase do trabalho prático pretende-se treinar modelos de classificação e regressão que respondam às questões formuladas, utilizando técnicas de amostragem na seleção de instâncias, *oversampling* e *undersampling*. Serão comparadas criticamente as diferentes métricas de avaliação de *performance* dos modelos. Além disso, objetiva-se aproveitar a infra-estrutura computacional da ferramenta SPARK para minimizar o *overhead* introduzido durante a fase de treino dos modelos.

II. Problemas na fase de modelação

Fornecidos dados de consumos energéticos para múltiplos códigos postais, em determinadas horas do dia, bem como a evolução das condições meteorológicas ao longo do dia, em Lisboa, o grupo definiu as duas perguntas formuladas anteriormente no capítulo da Introdução.

Na primeira fase do trabalho, foram aplicadas as técnicas de redução de dimensionalidade com o objetivo de evitar problemas causados pela *Curse Of Dimensionality*, onde a aprendizagem dos modelos é afetada pelo elevado conjunto de características dos dados e pela sua natureza qualitativa ou quantitativa discreta/contínua dos atributos.

A. Classificação de municípios

Do problema de classificação de municípios, foram extraídas, do conjunto de dados original, instâncias para diversos códigos postais, classificados manualmente e, por isso, suscetíveis a erro humano, como sendo predominantemente "Industriais"ou "Residenciais".
Após redução de dimensionalidade, verificou-se um cenário de classificação pouco complexo, ou seja, o fator determinante para distinguir o tipo de município pode ser analisado quase sem auxilio de algoritmos de aprendizagem automática, comparando diretamente o atributo referente ao consumo energético. Não obstante à fácil classificação, e com foco em compreender as diversas técnicas de amostragem e comparação de modelos de aprendizagem automática para os diversos conjuntos de dados, desenvolveu-se e compararam-se as diferentes técnicas de amostragem: *oversampling* e *undersampling* bem como os resultados obtidos do treino de diferentes classificadores.

B. Previsão de consumo energético

Para abordar a previsão do consumo energético com base nas condições meteorológicas previstas para um determinado dia preparou-se um conjunto de dados através de técnicas de *Feature Selection* para o treino de diferentes modelos de regressão.

III. Escolha do modelo de aprendizagem

Para a geração do modelo de aprendizagem, é necessário determinar o número de instâncias de cada classe nos conjuntos de dados preparados para ambas as questões. Desta forma, é possível compreender que técnicas de amostragem devem ser aplicadas, ou que modelos são adequados para responder às questões colocadas com base nos conjuntos de dados disponíveis.

Na classificação de municípios, o objetivo é resolver um problema de classificação binária, onde o número de instâncias manualmente categorizados como "Industrial"ou "Residencial"difere no conjunto de treino e teste numa proporção aproximada de 40/60. Assim, o uso de técnicas de *oversampling* e *undersampling* pode ser aplicado sem prejudicar a *performance* dos algoritmos.

Tabela I: Instâncias de cada classe no conjunto de dados de treino

CLASS	Instâncias
Industrial	80232
Residencial	117443

Pretende-se então comparar o desempenho de modelos de classificação como: *Random Forest*, *Decision Trees* e *Logistic Regression*.

Para a segunda questão, a previsão do consumo energético é feita treinando modelos de regressão supervisionados. Num cenário supervisionado, as *class labels* definidas na fase anterior são o atributo referente às condições meteorológicas.

Observando o número de instâncias para cada classe, no conjunto de treino obtido à custa do *dataset* reduzido utilizando a métrica *Fisher's Ratio* representado na Tabela III verificam-se classes

Tabela II: Métricas de *performance* dos modelos de Classificação de Municípios

Random Forest							
	False Positive Rate	Accuracy	Kappa	Pos Pred Value	Neg Pred Value	F1 Score	Area under ROC
No Sampling	0,23	0,90	0,79	0,78	**0,99**	0,86	0,88
Oversampling	**0,07**	**0,96**	**0,92**	**0,93**	0,98	**0,95**	**0,96**
Undersampling	0,08	**0,96**	**0,92**	0,92	**0,99**	**0,95**	**0,96**
Logistic Regression							
	False Positive Rate	Accuracy	Kappa	Pos Pred Value	Neg Pred Value	F1 Score	Area under ROC
No Sampling	0,29	0,80	0,58	0,71	**0,86**	0,74	0,79
Oversampling	**0,21**	**0,81**	**0,61**	**0,80**	0,82	**0,77**	**0,81**
Undersampling	0,22	0,80	0,60	0,78	0,82	0,76	0,80
Decision Trees							
	False Positive Rate	Accuracy	Kappa	Pos Pred Value	Neg Pred Value	F1 Score	Area under ROC
No Sampling	0,03	0,98	0,95	0,97	0,98	0,97	0,98
Oversampling	**0,02**	**0,99**	**0,98**	**0,99**	**0,99**	**0,99**	**0,99**
Undersampling	**0,02**	**0,99**	**0,98**	**0,99**	**0,99**	**0,99**	**0,99**

bastante sub representadas, pelo que a aplicação de técnicas de *undersampling* resulta na remoção de bastante informação do conjunto de treino, e *oversampling* resulta na introdução de *bias*.
Para este problema, abordam-se os modelos supervisionados: *Random Forest*, *Linear Regression*, *Decision Trees* e *Gradient Boosted Trees*.

Tabela III: Instâncias do conjunto de treino

CLASS	Instâncias
Clear	1189
Overcast	278
Partially Cloudy	3292
Rain, Overcast	10
Rain, Partially Cloudy	52

Estes modelos de aprendizagem automática são disponibilizados pelo Apache Spark [4] via interface para a linguagem de programação R, *sparklyr* . O uso desta biblioteca permite distribuir computações, aumentando a eficiência computacional das tarefas a desempenhar.

IV. Demonstração e comparação de resultados

Para efeitos de conclusão de validação dos conjuntos de dados elaborados na primeira fase, separaram-se os conjuntos de dados em treino e teste, numa proporção de 2/3 e 1/3, respetivamente.

A. Classificação de Municípios - Classificador

Destacando o número de instâncias de cada classe, mostrado na Tabela III, reforça-se a necessidade da utilização de técnicas de amostragem para realizar o balanceamento das classes do conjunto de dados. Estas técnicas não só permitem que o modelo tenha melhor *performance*, como em determinadas situações podem ser aplicadas para reduzir o conjunto de dados de treino, encontrando um sub-conjunto representativo dos dados originais que permita uma aprendizagem computacionalmente eficiente.

Neste problema, treinaram-se os modelos mencionados anteriormente, obtendo os resultados demonstrados na Tabela II, destacando as melhores métricas para cada um dos modelos a negrito.
O modelo *Random Forest* com *oversampling* mostra boa *performance* particularmente reduzido a taxa de falsos positivos e melhorando a precisão. Com um *F1 Score* de 0.95, apresenta-se um bom compromisso entre a precisão e o *recall* (identificação correta de positivos).
Da mesma forma, o modelo *Logistic Regression* também demonstra melhores resultados com *oversampling*. Contudo, durante o treino dos modelos, verificaram-se inconsistências quanto à *performance* dos modelos *Random Forest* e *Logistic Regression*, mostrando serem altamente sensíveis ao fator de aleatoriedade introduzido pela *seed*.
Este problema pode verificar-se devido a fatores como escassez de dados, ou *overfit* do modelo. Em ambos os cenários, a introdução de mais dados significativos poderiam resolver a sensibilidade à aleatoriedade da *seed*.

O modelo que aparenta ser mais robusto é o ***Decision Trees***, contudo apresenta um resultado pouco usual, em que tanto em situações de *oversampling* e *undersampling* as métricas de *performance* têm o mesmo valor. Mesmo aplicando um conjunto de teste não balanceado, os resultados são consistentemente os mesmos, este tópico requer uma análise

Tabela IV: Métricas de *performance* dos modelos de Regressão do consumo energético

	Random Forest	Linear Regression	Decision Trees	Gradient Boosted Trees
rmse	1,552	2,48	1,496	**1,092**
mse	2,41	6,151	2,238	**1,191**
r2	0,855	0,63	0,865	**0,928**
mae	1,203	1,959	1,12	**0,757**

mais pormenorizada num trabalho futuro para o entendimento destes resultados.

B. Previsão de consumo energético - Regressor

Remetendo ao número de instâncias de cada classe representado na Tabela III, compreende-se que é possível aplicar modelos de aprendizagem supervisionada, contudo, não são expectáveis bons resultados face à disparidade de elementos de cada classe. Mesmo aplicando técnicas de amostragem, para o balanceamento do conjunto de treino, no caso do *oversampling*, seria necessário introduzir demasiada informação repetida, aumentando o *bias*, e reduzindo a expressividade dos dados originais. Para *undersampling*, o conjunto de dados ficaria de tal forma reduzido que não seria possível obter resultados robustos.

A tabela IV apresenta as métricas de *performance* dos quatro modelos utilizados para fazer o consumo energético. Estas métricas são a raiz quadrada do erro médio (RMSE - *Root Mean Square Error*), o erro médio quadrático (MSE - *Mean Square Error*), o coeficiente de determinação ($R^2 - R\ Squared$) e o erro médio absoluto (MAE - *Mean Absolute Error*). Mostra-se que o algoritmo ***Gradient Boosted Trees*** apresenta os melhores resultados para todas as métricas:

- **RMSE**: O *Gradient Boosted Trees* tem o RMSE menor, com 1.092, o que indica um menor erro entre os valores previstos e os verdadeiros.
- **MSE**: Com 1.191, demonstra mais precisão entre valores previstos e reais.
- $\mathbf{R^2}$: Com o valor de 0.928, demonstra que 92.8% da variância do consumo energético é contemplada na regressão deste modelo.
- **MAE**: Com o menor valor, demonstra que este modelo obtém os menores erros absolutos comparativamente aos restantes.

No extremo oposto, o pior desempenho acontece no modelo *Linear Regression*, com os piores valores para todas as métricas, demonstrando ser o modelo menos preciso e menos fiável. Coincide que é também o modelo mais simples, e por isso, pode ser considerado o *baseline* nesta situação, servindo como comparação para modelos mais complexos.

V. Conclusão

A segunda fase do trabalho prático tinha por objetivo treinar e analisar métricas de *performance* de diferentes modelos de aprendizagem automática para dois problemas definidos na primeira fase: Classificação de municípios como "Industrial"ou "Residencial"e previsão do consumo energético com base nas condições meteorológicas na região de Lisboa. Em simultâneo, pretendia-se comparar a influência de diferentes tipos de amostragem, onde se observou que para cenários em que algumas classes estão bastante sub-representadas, podem não contribuir para melhores resultados dada a introdução de muita informação repetida, ou redução massiva do número de instâncias representativas.

Na classificação, o *Decision Trees* apresentou os melhores resultados para qualquer tipo de amostragem, contudo, tal como mencionado, alguns resultados requerem uma análise mais detalhada dado que as métricas de *performance* são idênticas para situações de *oversampling* e *undersampling*. Já os outros modelos neste caso, apresentam forte dependência do fator de aleatoriedade introduzido pela *seed*, o que demonstra pouca robustez, algo que pode ser melhorado no futuro ao adicionar mais dados relevantes.

O modelo regressor que melhor resultado apresentou foi o *Gradient Boosted Trees*, com as melhores métricas de *performance*. Quanto aos piores, o *Random Forest* mostrou que um cenário supervisionado com classes sub-representadas leva a maus resultados, e um modelo simples como o *Linear Regression* também fica aquém dos restantes.

Consideram-se os objetivos principais cumpridos, no entanto destacam-se alguns aspetos a melhorar, nomeadamente a introdução de uma terceira classe para o problema de classificação, referente a uma zona "mista"(Industrial e Residencial) para melhorar

o desempenho dos outros modelos; e adicionar mais dados representativos das classes sub-representadas no conjunto de dados utilizado para treinar o modelo regressor.

REFERÊNCIAS

[1] https://www.linkedin.com/pulse/comparing-machine-learning-models-find-best-fit-samad-esmaeilzadeh-dnzmf/
[2] https://neptune.ai/blog/performance-metrics-in-machine-learning-complete-guide
[3] https://spark.posit.co/
[4] https://spark.apache.org/