Rapport de modélisation sur les données météorologiques Australiennes

Formation : Data Scientist (05/2023 – 04/2024)

Encadrant : Francesco MADRISOTTI

Réalisateurs : Sophie BERTHIER

Luciano LANGHI

Quyen THIEU MARCAUD

Le ../../2023 – work in progress

Table des matières

[1 Introduction 3](#_Toc150092872)

[2 RainTomorrow 3](#_Toc150092873)

[2.1 Rappel sur déséquilibre 3](#_Toc150092874)

[2.2 Métriques 3](#_Toc150092875)

[2.3 Résultats de la classification par approches « classiques » via scikit-learn 3](#_Toc150092876)

[2.3.1 Modèles étudiés 3](#_Toc150092877)

[2.3.2 Impact du feature enginerring 3](#_Toc150092878)

[2.3.3 Conlusions 3](#_Toc150092879)

[2.4 Deep Learning avec Keras et TensorFlow 3](#_Toc150092880)

[2.4.1 DNN 3](#_Toc150092881)

[2.4.2 RNN 4](#_Toc150092882)

[2.5 Interprétabilité des modèles 4](#_Toc150092883)

[2.6 Exploitabilité et limites 4](#_Toc150092884)

[3 Extension à Rain+J 4](#_Toc150092885)

[4 MaxTemp 4](#_Toc150092886)

[4.1 Métriques 4](#_Toc150092887)

[4.2 Résultats de la régression par approches « classiques » via scikit-learn 4](#_Toc150092888)

[4.3 Séries Temporelles 4](#_Toc150092889)

[4.4 Deep Learning 4](#_Toc150092890)

[4.4.1 DNN 4](#_Toc150092891)

[4.4.2 RNN 4](#_Toc150092892)

[4.5 Interprétabilité des modèles 4](#_Toc150092893)

[4.6 Exploitabilité et limites 4](#_Toc150092894)

[5 Extension à MaxTemp+J 4](#_Toc150092895)

[6 Autres variables cibles 5](#_Toc150092896)

[7 Conclusion 5](#_Toc150092897)

# Introduction

Synthèse du feature enginnering effectué (reprendre conclusions du 1er rapport sur variables ajoutées et gestion des NA)

Trois approches de modélisation:

* Macro, sur toute l'Australie
* Micro, par station météo
* Intermédiaire, par zone climatique

Variables cibles étudiées, approches

# RainTomorrow

## Rappel sur déséquilibre

## Métriques

* Accuracy est non pertinente pour analyse fine
* Recall faible à optimiser, mais choix de la métrique dépend de l’objectif du modèle, des besoins du client, du coût des FP et des FN
* AUC comme critère pertinent de classification binaire de classe déséquilibrée
* Ajustement du seuil de probabilité pour améliorer le recall, au prix d’une baisse de l’accuracy et de la précision
* Tests X² pour vérifier que les résultats soient significatifs (utile en particulier pour analyses micro et/ou accuracy faible)

## Résultats de la classification par approches « classiques » via scikit-learn

### Modèles étudiés

* GridSearch, liste des modèles testés (KNN, Random Forest, GradientBoosting, XGBoost, MLP)

### Impact du feature enginerring

* Peu d’impact du travail de feature engineering et du nettoyage préalable effectué pour le 1er rapport
* Ajout variables explicatives (Luciano : J-1)

### Conlusions

* Peu d’impact du travail de feature engineering et du nettoyage préalable
* XGBoost best

## Deep Learning avec Keras et TensorFlow

### DNN

Modèle optimal (2 couches, pas trop de neurones, tanh, sigmoid en sortie)

Pas meilleur que XGBoost, mais beaucoup plus long => pas intéressant

### RNN

## Interprétabilité des modèles

* SHAP
* Variables explicatives
* Différences de l’importances variables selon type d’approche

## Exploitabilité et limites

* Précision restant faible

# Extension à Rain+J

* Tentative de prédire non seulement RainTomorrow (=Rain+1) mais également Rain+J, avec un modèle XGBoost par J
* Résultats sur une année, en global, par zone climatique, par location
* Comparaison avec RNN
* Conclusions

# MaxTemp

## Métriques

## Résultats de la régression par approches « classiques » via scikit-learn

Prédiction pour le lendemain uniquement

## Séries Temporelles

Reprendre partie écrite dans 1er rapport. Constat de la saisonnalité. Résidu restant important.

Détermination des valeurs max (p,d,q)(P,D,Q)I

Impossibilité de modéliser via SARIMAX sur une saisonnalité de 365 jours (trop de temps de calcul)

## Deep Learning

### DNN

### RNN

## Interprétabilité des modèles

## Exploitabilité et limites

# Extension à MaxTemp+J

* Tentative de prédire non seulement MaxTemps pour le lendemain, mais également pour plusieurs jours, avec un modèle XGBoost par J
* Résultats sur une année, en global, par zone climatique, par location
* Comparaison avec RNN
* Conclusions

# Autres variables cibles

Rainfall => résultats très mauvais pour le moment

# Conclusion

Difficultés rencontrées, échecs

Aspects positifs

Pistes de travaux d’amélioration (sunshine à récupérer sur site du BOM si on avait + de temps car souvent utilisé comme critère important par SHAP)