Verslag Distributed Databases

Dylan Cluyse, Laura Renders, Liam Goethals

14 december 2022

Inleiding

Wij maakten tijdens het opleidingsonderdeel 'Distributed Databases' kennis met Apache Spark. Spark biedt data-analyse gericht grootschalige gegevensverwerking. Deze technologie wordt aangeboden voor Java, Python, R en Scala. Voor dit opleidingsonderdeel werd Java gekozen als programmeertaal. Spark bevat enkele zijtakken dat zich richt op andere aspecten binnen dataverwerking, één daarvan is MLLib. MLLib is een pakket gericht op machine learning met Spark. Om meer kennis te vergaren kregen wij de groepsopdracht om via het platform Kaggle deel te nemen aan machine learning (ML) gerichte competities.

In dit verslag nemen wij u mee in de wereld van ML gedreven door Spark. Wij willen u met volle plezier enkele concrete casussen tonen die gebruik maken van de verschillende regressie- en classificatiemethoden.

Allereerst willen wij de score bij het bordspel Scrabble achterhalen met regressie. Vervolgens detecteren wij kredietkaartfraude met behulp van binaire classificatie. Als derde oefening willen wij op basis van tekstinhoud achterhalen of een Tweet gerelateerd is aan een ramp. Tot slot geven wij u onze bevindingen mee van het MLLib pakket. Hier leggen we de aanpak van Spark en SKLearn, een pakket dat we in het opleidingsonderdeel Machine Learning zagen, parallel tegenover elkaar.

1. Tijd van een taxi-verplaatsing voorspellen met regressie.

Bron: https://www.kaggle.com/competitions/nyc-taxi-trip-duration

Aanpak

* Flag is categorische variabele –> indexen

De pickup en dropoff datetime wordt meegegeven als datetime-object. Dit moeten we veranderen naar een bruikbaar formaat voor het regressiemodel.

Evaluatie

2. Credit Card Classificatie

Bron: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Aanpak

We hebben hier twee klassen: correcte tweets en fraudulente tweets. We moeten een aanpak vinden voor een binaire classificatiemodel.

Evaluatie

Bij het evalueren van het classificatiemodel ging onze voorkeur uit naar de confusionmatrix. Zo hebben wij een zicht op hoe goed ons model de klassen kan voorspellen. Bij de logistische regressie merken wij een goed evenwicht op tussen valse positieven en valse negatieven. De classificatiemodellen met Random Forest en Decision Trees voorspellen daarentegen weinig valse negatieven, maar wel uitbundig veel valse positieven. Als extra hebben wij ook de nauwkeurigheid van het model berekent.

3. NLP.

Aanpak

Gebruikte regressiemodellen: * Lineaire regressie * Random forest regressie

De resultaten van de metrieken waren zwak. Onze presumptie lag bij de volatiele waarden. Sommige waarden waren tussen het bereik van 0 - 10. De ... feature daarentegen had een range van 0 - 1000. Dit hebben we opgelost door een MinMaxScaler toe te voegen aan de pipeline.

Evaluatie

4. Verschillen highlighten met andere machine learning pakketten

Sklearn

...

Conclusie

Uit document: