**Erfahrungsbericht**

**Innovationsprojekt - Daten aus dem "Motoren-Linearitätstest" eines Diagnostischen Instruments**

Bevor ein diagnostisches Instrument für den Verkauf freigegeben wird, müssen diese etliche Tests in der Produktion durchlaufen. Diese Tests stellen sicher, dass die Instrumente innerhalb ihrer definierten Spezifikationen liegen.

Im Rahmen dieses Innovationsprojektes wurde die Daten aus dem cobas® LIAT "Motoren Linearitätstests" analysiert. Es sollte untersucht werden, ob die Konfigurationsdaten mit den Daten zur Motorenlinearität korrelieren und ob ggf. anhand der vorhanden Daten eine Voraussage gemacht werden kann, ob der Lin0-Test "Passen" oder "Failen" wird.

Anfänglich wurden Daten mit 3800 Beispielen und 67 Features geliefert. Das Hauptproblem bestand jedoch darin, dass Daten sehr „Imabalanced“ waren. Es waren lediglich 60 von 3800 Datensätzen als „Failed“ gelabled. Die Daten wurden bereinigt und es wurden die Features, welche stark korrelierten von der weiteren Betrachtung ausgeschlossen. Wegen dem „Imabalanced Dataset“ wurde die „Anomaly Detection“ auf die Daten angewendet. Kurz nachdem die „Anomaly Detection“ durchgeführt war wurden komplett neue Daten geliefert, welche sich strukturell stark vom ersten Datenshet unterschieden haben. Die Originaldaten hatten ein Shape von (15630, 1036) welches sich jedoch nach der Bereinigung auf ein Shape von (8719, 574) reduzierte. Der neue Datensatz ist zwar auch „Imablanced“ jedoch nicht so stark wie der erste Datensatz. Von insgesamt 8719 Beispielen sind 1630 Failed und 7098 Passed.

Nach Rücksprache mit unserem Betreuer wurde die Entscheidung getroffen, dass wir uns nur noch auf die neuen Daten konzentrieren und dass die Untersuchungen basierend auf dem ersten Datensatz auch nicht Teil des einzureichenden Innovationsprojektes sein sollen.

Die Bereinigung des neuen Datensatzes erforderte das Entfernen von Duplikaten, Nullwerten und Features, welche mehr als 95% miteinander korrelierten. Des Weiteren wurden 44 Statuswerte (Kategorische Features) entfernt, da diese sehr stark mit Labelwert „lin0\_bonsai\_state“ korrelierten (war einer dieser Statuswerte „Failed“ war in fast allen Fällen auch der Labelwert „Failed“).

Für den weiteren Verlauf haben wir uns dazu entschieden eine „Random Forest“ Analyse auf die Daten anzuwenden. Das Random Forest Modell performte mit einem Recall von ca. 98% und einer Precision von ca. 93% was aus unserer Sicht für ein erstes Modell zufriedenstellend ist und als Referenz für weitere Modelle dient.

Im nächsten Schritt wurde untersucht, ob sich Cluster finden in denen besonders viele Failed Tests wiederfinden würden. Dabei bestand die Hoffnung, dass bei so einem Cluster ggf. Muster für Failed Tests zu finden sein könnten. Die Anwendung von KMeans erzeugte jedoch kein Cluster mit mehr als 60% Failed Tests. Die Cluster wurden daher nicht weiter berücksichtigt.

Als nächstes wurden mehrere „Deep Neuronal Network“ Modelle erstellt. Ein erstes Modell ohne L2 Regulierung zeigte starkes „Overfitting“. Ein weiteres Modell mit L2 Regulierung zeigte zwar ein wenig bessere Werte hinsichtlich dem „Overfitting“ Verhalten approximierte aber wie auch schon das erste Modell schlecht. In einem dritten Modell wurden dann noch die Gewichte für „Passed“ und „Failed“ angepasst um ggf. die „Imabalance“ zwischen den „Passed“ und „Failed“ Beispielen auszugleichen. Dieses dritte Modell approximierte nun um einiges besser und es zeigte auch akzeptable Werte hinsichtlich „Overfitting“.

Bei den neuronalen Netzen wurde noch versucht mit „Dropout Layer“ und weniger tiefen Netzen die Performance zu verbessern. Hiermit konnten allerdings keine verwertbaren Ergebnisse erzielt werden. Es wurde daher nicht weiter in dieser Richtung evaluiert.

Nun wurden von allen vier Modellen mit dem Test-Datenset eine Prediction erstellt und die Ergebnisse evaluiert. Da die Daten „Imbalanced“ sind wurde die Accuracy nicht als Metrik für die Bewertung verwendet, sondern Recall und Precision.

In dem Innovationsprojekt hat sich gezeigt, dass wir mit dem bisher angeeigneten Wissen in der Lage sind die gelieferten Daten angemessen zu beurteilen und Modelle zu finden die weitaus besser sind als „Random“. In dem Projekt konnte auch gezeigt werden, dass durch die Anwendung von Regulierung, und das Ausgleichen von „Imbalanced“ Daten durch Anpassen der Gewichte in einem Neuronalen Netzwerk die Performance solcher Netzwerke gesteigert werden kann.

Im Rahmen des Innovationsprojektes wurde die Modelle zu „Anomly Detection“, „Random Forest“, „KMeans“ und „Deep Neuronal Networks“ auf bereitgestellte Daten angewendet. So konnte das Wissen über diese Modelle auch vertieft werden.

Aus unserer Sicht ist mit den Vorlesungen und den Übungen zum Kurs Machine Learning eine gute Basis für die Vertiefung der Kenntnisse geschaffen. Unsicherheiten bestehen im Wesentlichen noch bei der systematischen Anwendung von Hyperparametern bei neuronalen Netzen. Unsicherheiten bestehen auch bei Modellen, welche im Innovationsprojekt nicht berücksichtigt wurden. Wir gehen allerdings davon aus, dass mit Hilfe der Ressourcen aus den Vorlesungen, den Übungen und den Informationen im Internet auch diese Modelle auf konkrete Probleme angewendet werden können.