Lab Project - Short Paper - Altendorfer & Eckmayr

Datum: 19.01.2024

Im Rahmen des Lab Projects haben wir einen Kaggle-Datensatz zu Testleistungen US-amerikanischer Schüler analysiert. Mit den Tools Spark und Plotly haben wir es uns zur Aufgabe gemacht, Muster und Korrelationen in den Daten zu entdecken und zu visualisieren. Unser Code, Ergebnisse und die Visualisierungen sind in unserem GitHub-Repository und im zugehörigen Jupyter Notebook, das über nbviewer zugänglich ist, vollständig dokumentiert. Für detaillierte Informationen möchten wir auf unser Jupyter Notebook verweisen, welches

Links zum Projekt:

Github Repository

Jupyter Notebook auf nbviewer

Datensatz auf Kaggle

Setup und Datenvorbereitung

Wir begannen mit dem Aufsetzen einer Spark-Session und bereinigten die Daten für die Analyse. Unsere erste Aufgabe war eine Überprüfung der Verteilung der Testvorbereitungskurse, gefolgt von einer tiefgehenden Untersuchung der Prüfungsergebnisse über verschiedene ethnische Gruppen hinweg.

Forschungsfragen und Analysen

Forschungsfrage 1: Unterschiede in Prüfungsergebnissen

Wir beobachteten signifikante Unterschiede in den Prüfungsergebnissen zwischen den ethnischen Gruppen. Gruppe E schnitt durchweg besser ab als die anderen Gruppen.

Forschungsfrage 2: Einfluss des elterlichen Bildungsniveaus

Interessanterweise zeigte sich, dass das Bildungsniveau der Eltern einen Einfluss darauf hat, ob ein Schüler einen Vorbereitungskurs absolviert oder nicht, jedoch nicht in dem Maße wie man das eventuell vermuten würde, nämlich das Eltern mit höheren Bildungsabschlüssen ihre Kinder eher in Vorbereitungskurse schicken. Dies ist nicht der Fall. Unsere Auswertung zeigt deutlich, dass Eltern mit „Some High School“ als höchsten Bildungsabschluss ihre Kinder mit 43% am häufigsten in Vorbereitungsangebote anmelden. Die Absolvierung eines Vorbereitungskurses wirkt sich positiv auf die Testergebnisse aus.

Forschungsfrage 3: Zusammenhang zwischen Mittagessen und Leistung

Unsere statistische Analyse zeigte einen signifikanten Zusammenhang zwischen der Art des Mittagessens und den Prüfungsergebnissen, wobei wir für den statistischen Test einen T-Test verwendet haben. Schüler mit Standardmittagessen erreichten im Durchschnitt höhere Werte als Schüler mit reduziertem Mittagessen.

Forschungsfrage 4: Vorhersage von Noten durch Machine Learning

Zuletzt setzten wir Machine Learning ein, um die Noten basierend auf verschiedenen Faktoren (Parental Level of Education, Lunch, Test Preparation Course) vorherzusagen. Wie zu erwarten war

die Note „Befriedigend“, die am häufigsten vergebene Note. Mit einer linearen Regression haben wir die Test-Scores separat vorhergesagt. Wir konnten Modelle entwickeln, die die Noten im österreichischen Schulsystem mit einer akzeptablen Genauigkeit vorhersagen, gemessen am RMSE (Root Mean Squared Error von 0.98). Es kann daher mit den genannten kategorialen Variablen eine Tendenz für die erzielten Testergebnisse vorhergesagt werden.

Visualisierungen

Mit Plotly erstellten wir interaktive Balkendiagramme, die die Durchschnittswerte und Unterschiede anschaulich darstellen. Histogramme ergänzten unsere Analyse, indem sie die Verteilung der Noten visualisierten. **Da wir zahlreiche Visualisierungen erstellt haben, möchten wir zur Ansicht auf unser Jupyter-Notebook verweisen.**

Fazit

Die Ergebnisse unseres Projekts zeigen, wie man mit der Kombination aus Big Data Technologien (Spark) und statistischen Methoden, Einblicke in Bildungsdaten gewinnen kann. Wir haben nicht nur wertvolle Praxiserfahrung im Umgang mit Big Data Tools gesammelt, sondern auch Einblicke in Erfolgsfaktoren für Schüler und Studierende erhalten – **Eat your Lunch!** 😉

Fabian Altendorfer und Andreas Eckmayr