Analysedokument für Automatidata

Projekt:

- **Titel:** Automatidata - Analyse und Prognose von Taxi-Fahrpreisen in New York City
- **Projektleiter:** Dominik Vogel
- **Datum:** [Aktuelles Datum]
- **Version:** 1.0

Bird Analytics

Inhaltsverzeichnis

1. Projektübersicht

- **Projektname**: Automatidata - NYC Taxi Preis Vorhersage
- **Projektziel**: Entwicklung eines präzisen Modells zur Vorhersage von Taxi-Fahrpreisen basierend auf historischen Daten und relevanten Variablen.

2. Ziele der Analysephase

1. **Datenqualität evaluieren**: Identifikation von Nullwerten, Anomalien und anderen Datenproblemen.
2. **Datenstruktur verstehen**: Analyse der Verteilung wichtiger Variablen, z. B. Fahrstrecke und Fahrpreis.
3. **Schlüsselvariablen bestimmen**: Identifikation der Variablen mit dem höchsten Einfluss auf die Fahrpreisprognose.
4. **Hypothesen entwickeln**: Formulierung von Hypothesen basierend auf explorativen Erkenntnissen (z. B. Zusammenhang zwischen Fahrstrecke und Fahrpreis).
5. **Erkenntnisse für Modellbildung dokumentieren**: Alle wesentlichen Punkte der Datenstruktur für die nächsten Schritte festhalten.

3. Methodik der Explorativen Datenanalyse (EDA)

1. **Datenüberprüfung und -vorbereitung**:
    - Fehlende Werte identifizieren und Strategien zur Behandlung formulieren.
    - Datentypen und Konsistenz der Werte validieren.
    - Dokumentation von unplausiblen Datenpunkten (z. B. "0"-Distanz bei hohen Kosten).
2. **Untersuchung der Verteilung und Schiefe der Variablen**:
    - Erstellung von Histogrammen und Boxplots, um die Verteilung der Variablen zu analysieren.
    - Analyse der Schiefe und Auswirkungen auf die Modellierung.
3. **Zusammenhang zwischen Variablen analysieren**:
    - Scatterplots und Korrelationsanalysen für numerische Variablen.
    - Ermittlung von Zusammenhängen, die für die Modellierung relevant sind.
4. **Ermittlung von Anomalien und Ausreißern**:
    - Identifikation von unplausiblen Werten (z. B. kurze Strecken mit hohen Kosten).
    - Strategien zur Behandlung solcher Anomalien entwickeln.
5. **Visualisierung**:
    - Erstellung von Diagrammen (z. B. Histogramme, Heatmaps, Scatterplots), um Erkenntnisse zu präsentieren.

4. Risiken und Abhängigkeiten in der Analysephase

Risiko	Auswirkung	Maßnahmen
Fehlende Werte in Schlüsselvariablen	Ungenauigkeit in Modellen	Nullwerte identifizieren und Imputation planen
Unplausible Ausreißer	Verfälschte Ergebnisse	Anomalien erkennen und Strategien entwickeln
Ungleichgewicht in Datenkategorien	Verzerrte Modellvorhersagen	Sampling- oder Gewichtungsmethoden anwenden

5. Deliverables

1. **EDA-Bericht**:
    - Zusammenfassung der Datenstruktur, Verteilungen und Schieflagen.
    - Dokumentation von Nullwerten und Anomalien.
2. **Visualisierungen**:
    - Histogramme, Scatterplots und Heatmaps, um Zusammenhänge und Verteilungen darzustellen.
3. **Empfehlungen für die Modellphase**:
    - Liste der wichtigsten Variablen.
    - Vorschläge zur Handhabung von Nullwerten und Ungleichgewichten.
    - Dokumentation der Anomalien und ihrer potenziellen Auswirkungen.

6. Nächste Schritte nach der Analysephase

1. Datenbereinigung basierend auf den EDA-Ergebnissen.
2. Erstellung einer Liste validierter Variablen für die Modellentwicklungsphase.
3. Hypothesenüberprüfung und statistische Tests zur Verifizierung der Beziehungen zwischen Variablen.
4. Erstellung des finalen Datensatzes für die Modellierung.