













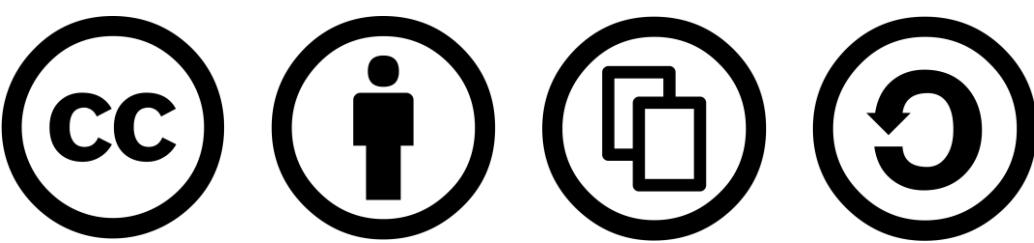


Data Science Canvas

Datum:	Teilnehmer:
Projekt:	Resultierende Aufgaben:

Problemstellung				Durchführung & Auswertung		Datenerhebung & -aufbereitung	
<div>Business Case & Mehrwert Welcher Business Case sollte analysiert werden und welchen Mehrwert erzeugt er? </div> <div>Business Case: Optimierung der Absatzprognosen durch die Berücksichtigung von unstrukturierten Daten zu Trends in Social Media Mehrwert:<ul style="list-style-type: none">Warendisposition verbessernRestbestände reduzierenEinkauf optimierenTransportlogistik optimieren</div>	<div>Modellauswahl Welche Analysemethoden kommen auf der Grundlage der spezifischen Datenlandschaft und des Business Case in Frage? </div> <div>Aufbereitung der Social Media Daten:<ul style="list-style-type: none">Sentimentanalyse mithilfe des Naive Bayes-Classifiers Absatzprognose<ul style="list-style-type: none">Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX)</div>	<div>Modellanforderungen Welche Modellanforderungen müssen erfüllt sein, um ein valides Modell zu erhalten? </div> <div>Naive Bayes Classifier:<ul style="list-style-type: none">Extraktion von Worthäufigkeiten mittels Bag-of-Words SARIMAX:<ul style="list-style-type: none">Sentimente als zeitbasierte TrendsStationäre Reihen zur Anwendung der Zeitreihenanalyse mittels SARIMAX</div>	<div>Skills Welche Fähigkeiten sind für die Bereitstellung der Daten und die Modellentwicklung erforderlich? </div> <div>Vorhanden:<ul style="list-style-type: none">Kenntnisse zur Modellierung von Zeitreihen Benötigt:<ul style="list-style-type: none">Kenntnisse im Bereich Machine Learning, speziell: Natural Language Processing, Bag-of-Words und SentimentanalyseData Engineer zur Datenintegration in die bestehende Architektur und Vorbereitung der Daten für das Machine Learning-Modell</div>	<div>Modellevaluation Welche Indikatoren erfordern Qualitätskontrolle und Validierung und wie sollten sie interpretiert werden? Ist eine Echtzeit-Überwachung notwendig? </div> <div>Naive Bayes Classifier: Überprüfung der Klassifikationsgüte mittels:<ul style="list-style-type: none">KonfusionsmatrixGenauigkeit (Precision)Trefferquote (Recall)Genauigkeit (Accuracy)F-Maß als Kombination von Trefferquote und Genauigkeiten SARIMAX: Überprüfung der Prognosegüte durch die Wurzel des mittleren quadratischen Prognosefehlers (RMSE). Dieser gibt an, wie stark die Vorhersage im Mittel von den historisch zugrundeliegenden Daten abweicht. Je höher der RMSE ausfällt, desto schlechter ist das Modell.</div>	<div>Data Storytelling Welche Anforderungen hat die Zielgruppe an die Präsentation der Ergebnisse und wie kommuniziere ich diese Daten effektiv? </div> <div>Management:<ul style="list-style-type: none">Visualisierung der Zeitreihenfortschreibung mit explizitem Aufzeigen von Strukturbrüchen und Trends aus Social MediaAngabe der Modellgüte Controlling:<ul style="list-style-type: none">Tiefere Darstellung über die Herleitung des Modells über Autokorrelations- (ACF) und Partielle Autokorrelationsfunktion (PACF)</div>	<div>Datenauswahl & -bereinigung Welche der verfügbaren Daten sind relevant? Müssen die Daten bereinigt werden? </div> <div>Historische Absatzdaten: Bereinigung von Ausreißern und Interpolation der Lücken Social Media-Daten: Bereinigung der Social Media-Daten von Stop-Words</div>	<div>Datenerhebung Wie und mit welchen Methoden sollen zusätzlich benötigte Daten erhoben werden? Welche Eigenschaften müssen diese Daten erfüllen? </div> <div>Abruf von Social Media-Daten (z.B. Kommentare) über die jeweiligen APIs</div>
<div>Datenlandschaft Welche Daten werden benötigt und welche sind bereits verfügbar? Welche zusätzlichen Daten müssen erhoben werden? </div> <div>Vorhanden:<ul style="list-style-type: none">Historische Absatz- und Kundendaten aus SAP BI-Systemhistorische Daten von Werbemaßnahmen und –kanäle vom Vertrieb Benötigt: Daten aus Social Media (z.B. Kommentare)</div>		<div>Software & Bibliotheken Welche Software sollte verwendet werden? Gibt es bereits eine Standardlösung? Welche Bibliotheken werden eingesetzt? </div> <div>Software: R Studio zur Aufbereitung, Analyse und Darstellung der Ergebnisse Bibliotheken:<ul style="list-style-type: none">e1071 package für den Naive Bayes Classifierforecast package für SARIMAXstopwords package zur Entfernung von Stoppwörtern</div>			<div>Datenintegration In welches System sollen die Daten aus verschiedenen Quellen migriert werden? </div> <div>Integration in einen Use Case Pool des Data Lakes</div>	<div>Explorative Datenanalyse Gibt es Ausreißer oder Strukturen, die zu berücksichtigen sind? Erstellung von beschreibenden Kennzahlen für die erste Beurteilung der Daten. </div> <ul style="list-style-type: none">Die historischen Absatzdaten weisen teils größere zeitliche Lücken aufEs lassen sich Trends und Saisonalitäten in den Sommermonaten erkennen	
<div>Kosten Was sind die Kostenkategorien? Wie hoch werden die Kosten sein? </div> <div>Zusätzliche Gehälter des Entwicklungsteams (Data Scientist, Data Engineer) für das Projekt</div>				<div>Einnahmen Wie kann das Modell Einnahmen generieren? Senkt das Projekt die Kosten? </div> <div>Verringerung der Lager- und Logistikkosten, Verbesserte Preispolitik durch Trenderkennung</div>			



Entwickelt von:
Thomas Neifer, Dennis Lawo, Margarita Esau, Paul Bossauer, Lukas Böhm, Gunnar Stevens

Zugehörige Publikation:
Neifer, T., Lawo, D., & Esau, M. (2021). Data Science Canvas: Evaluation of a Tool to Manage Data Science Projects. In *Proceedings of the 54rd Hawaii International Conference on System Sciences*.