FREIE UNIVERSITÄT BERLIN

FACHBEREICH WIRTSCHAFTSWISSENSCHAFT

Abschlussdokumentation

des Fachbereichs Wirtschaftswissenschaft

der Freien Universität Berlin

Animal Shelter

Berlin, den 12.08.2022

Inhaltsverzeichnis

Inhaltsverzeichnis 2

Abbildungsverzeichnis 3

1 Einleitung 4

2 CRISP-DM: Project Understanding 4

2.1 Content 4

2.2 Project Goal 5

2.3 Domain Knowledge 5

3 CRISP-DM: Data Understanding and Visualization 7

4 CRISP-DM: Data Preparation 14

4.1 Data Selection 14

4.2 Data Cleaning, Transformation und Integration 15

5 CRISP-DM: Modeling 17

5.1 Model Selection 17

5.2 Model Comparison 19

5.2.1 Model Ranking 19

5.2.2 Best Model - Confusion Matrix 20

5.3 Optimizing our Model 21

6 CRISP-DM: Evaluation 23

6.1 Comparison and Assessment 23

6.2 Opportunities for improvement 24

6.3 General recommendations for the shelter 24

7 Abschluss und Fazit 25

7.1 Reflexion des Vorgehens 25

7.2 Ausblick 25

Literaturverzeichnis 26

Appendix 26

Abbildungsverzeichnis

1 Einleitung

Die vorliegende Arbeit wird im Rahmen des Moduls Business Intelligence erstellt und zeigt die Dokumentation der Ergebnisse des Data-Mining Projektes „Animal Shelter“. Die Fragestellung dieser Ausarbeitung ist, eine Prognose des „Outcomes“ (beispielsweise Adoption oder Übergabe an den Besitzer) der Tiere beim Verlassen des Tierheims zu erstellen. Dies soll mit Hilfe des CRISP-DM Prozesses durchgeführt werden. Das Modell bestellt aus den sechs Phasen Project Understanding, Data Understanding, Data Preparation, Modeling und Evaluation, welche in den folgenden Kapiteln näher eingegangen wird. Zur Unterstützung der jeweiligen Schritte wurde die Programmiersprache Python herangezogen, zur Datenvisualisierung, Datenbereinigung, Datenkonvertierung, Modellierung und weitere wichtige Schritte.

2 CRISP-DM: Project Understanding

2.1 Content

Jedes Jahr werden 7,6 Millionen Haustiere in den USA in Tierheimen aufgenommen. Sie werden entweder von ihren Familien abgegeben, oder von Tierschützern gefunden, beziehungsweise in Gewahrsam genommen. Während manche Tiere neue Familien finden, werden 2,7 Millionen Hunde und Katzen jedes Jahr eingeschläfert (Kaggle, n.d.). Um das Outcome von Tierheimtieren vorherzusagen, hat das Austin Animal Center einen Datensatz zur Verfügung gestellt. Enthalten sind Merkmale wie Rasse, Farbe, Geschlecht und Alter. Die Wichtigkeit dieser Challenge wird durch die Auswirkungen der Corona-Pandemie noch deutlicher. Zu Beginn der Corona-Pandemie haben sich viele Familien, Paare und Singles einen Hund angeschafft, da sie durch die teils verkürzten Arbeitszeiten und das Home Office viel freie Zeit gewonnen haben. Mittlerweile haben einige Unternehmen die Home Office Arbeit wieder auf ein Minimum reduziert, weswegen Tierschützer davon ausgehen, dass in den kommenden Wochen und Monaten die Anzahl an abgegebenen Hunden in Tierheimen drastisch steigen wird (Thelen, 2022).

Eine verbesserte Vorhersage der Outcomes von Tierheimtieren kann auch dazu beitragen, Tiere aus dem Ausland zu retten. Vor allem in südlichen und östlichen Ländern Europas sind die Bedingungen in den Tierheimen deutlich schlechter, als in Deutschland (Deutscher

Tierschutzbund e.V., 2018). Gelingt es durch das Lösen der Challenge, dass mehr Tiere vermittelt werden, können die freien Plätze auch an besonders schwere Fälle aus dem Ausland vergeben werden.

2.2 Project Goal

Ziel des Projekts ist es, ein Verständnis dafür zu bekommen, welche Tiere welches Outcome haben. Tierheime können dadurch ihre Energie auf bestimmte Tiere fokussieren, die besonders schwer vermittelbar sind, um auch ihre Chancen zu erhöhen.
Da wir die Zielwerte mit den tatsächlichen Ergebnissen als Datensatz erhalten, muss die Aufgabe wie ein Supervised Learning Problem behandelt werden, was den Einsatz von Klassifikation und Regression ermöglicht. Außerdem haben wir es mit einem Offline-Lernsystem zu tun, das nicht in Echtzeit aktualisiert werden muss.
Das Maß für die Qualität des Modells ist durch die Spezifikationen des Kaggle-Projekts vorgegeben, nämlich der logarithmische Multiklassenverlust. Die erwartete Ausgabe soll eine Vorhersagerate für jede Ergebnisklassifizierung sein ("adoptiert", "eingeschläfert", "verlegt" und “Rückgabe an den Besitzer”)

2.3 Domain Knowledge

Ob ein Tierheimtier ein positives oder negatives Outcome aus dem Tierheim hat, kann von verschiedenen Faktoren beeinflusst werden. Die im Datensatz zur Verfügung stehenden Faktoren wie Rasse, Farbe, Alter oder Geschlecht können objektiv erfasst werden. In Deutschland sind einige Hunderassen gelistet, das heißt für die Führung eines solchen Hundes sind verschiedene Prüfungen abzulegen. Dies könnte die Vermittlung von verschiedenen Rassen beeinflussen. Außerdem suchen die meisten Menschen nach jungen Hunden, beziehungsweise Welpen, weswegen es ältere Hunde ebenso schwerer haben, ein zu Hause zu finden (Sparacino, 2021). Eine Vermutung über die verschiedenen Abhängigkeiten der Faktoren ist in Abbildung 1 zu sehen.

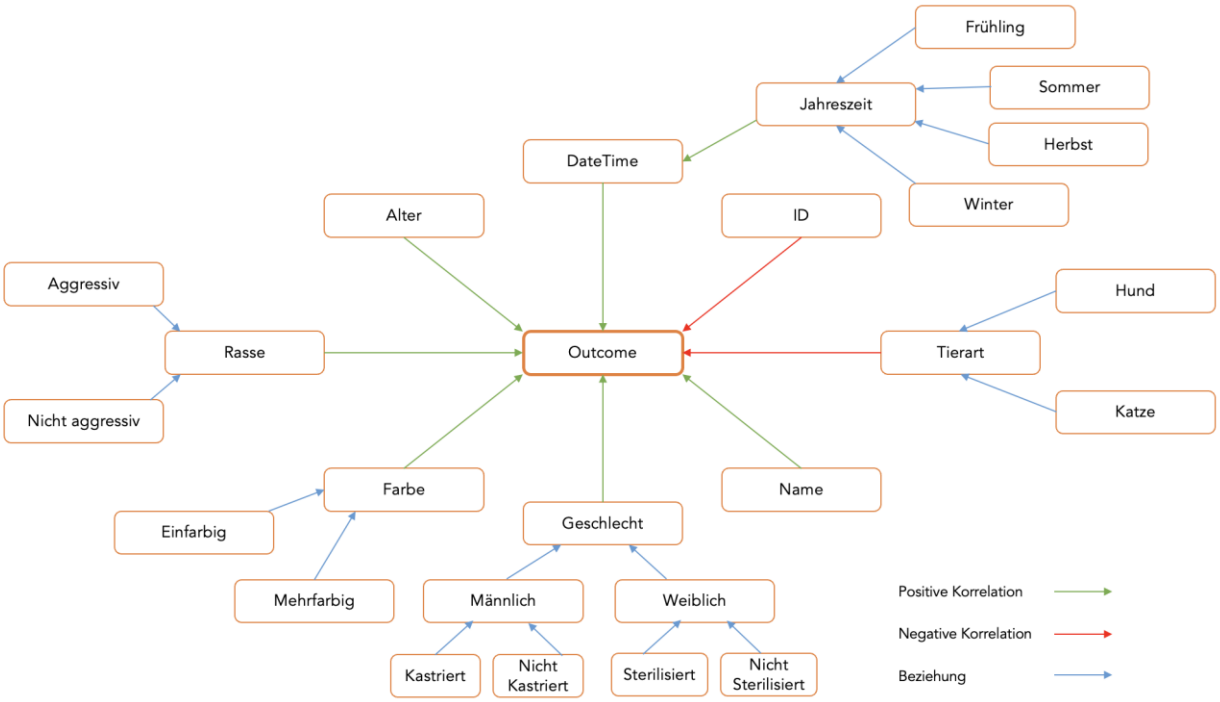


Abbildung 1: Abhängigkeiten

3 CRISP-DM: Data Understanding and Visualization

Um dem “Austin Animal Shelter” ein möglichst präzises Vorhersagemodell über die erwartbare Zukunft des Tieres ("adoptiert", "eingeschläfert", "verlegt" und “Rückgabe an den Besitzer”) zu entwickeln, werden umfangreiche relevante Daten benötigt, die einen Aufschluss über den Wirkungszusammenhang mit bestimmten Merkmalsausprägungen der Tiere geben. Kaggle stellt die Daten im Rahmen einer CSV-Datei zur Verfügung, welche vom “Austin Animal Shelter” in den Jahren 2013 bis 2016 gesammelt wurden. Der zentrale in der vorliegenden Arbeit verwendete Datensatz “train.csv” umfasst Einträge zu 26.729 Hunden und Katzen und beinhaltet jeweils Informationen zu zehn Attributen pro Tier: “AnimalID”, “Name”, “DateTime”, “OutcomeTime”, OutcomeSubtype”, “AnimalType”, “SexuponOutcome”, “AgeuponOutcome”, “Breed” und “Color”. Eine Übersicht über die ersten fünf Zeilen des verwendeten Datensatzes findet sich in Abbildung 2.

	AnimalID	Name	DateTime	OutcomeType	OutcomeSubtype	AnimalType	SexuponOutcome	AgeuponOutcome	Breed	Color
0	A671945	Hambone	2014-02-12 18:22:00	Return_to_owner	NaN	Dog	Neutered Male	1 year	Shetland Sheepdog Mix	Brown/White
1	A656520	Emily	2013-10-13 12:44:00	Euthanasia	Suffering	Cat	Spayed Female	1 year	Domestic Shorthair Mix	Cream Tabby
2	A686464	Pearce	2015-01-31 12:28:00	Adoption	Foster	Dog	Neutered Male	2 years	Pit Bull Mix	Blue/White
3	A683430	NaN	2014-07-11 19:09:00	Transfer	Partner	Cat	Intact Male	3 weeks	Domestic Shorthair Mix	Blue Cream
4	A667013	NaN	2013-11-15 12:52:00	Transfer	Partner	Dog	Neutered Male	2 years	Lhasa Apso/Miniature Poodle	Tan

Abbildung 2: Übersicht Datensatz “train.csv”

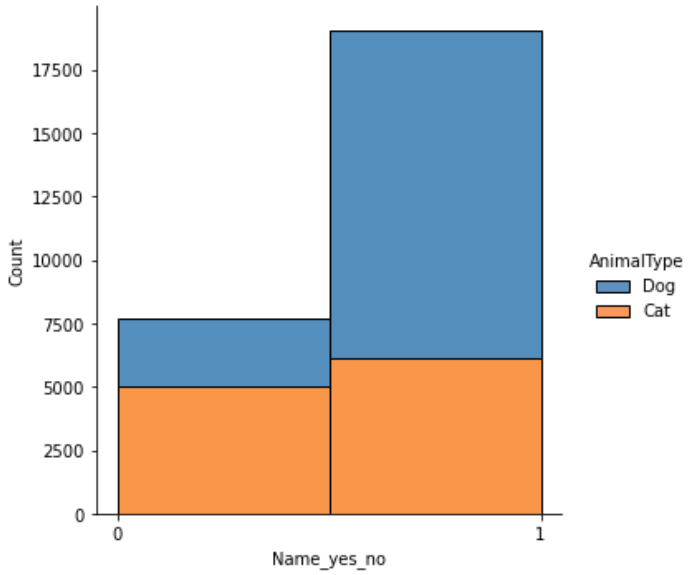
Im folgenden sollen die einzelnen Variablen beschrieben, visualisiert, auf Ihre Qualität hin überprüft und ihre Bedeutung für die Entwicklung des Vorhersagemodells erläutert werden.

Attribut 1: “AnimalID”

Die “AninmalID” weist jedem Tier eine spezifische Buchstaben und Nummer Kombination zu. Dieses Attribut liefert keinen Mehrwert in Bezug auf die Vorhersage über Zukunft (Outcome) eines Tieres und wird nicht weiter verwendet. Lediglich zur Bereinigung von möglichen Duplikaten wurde das Attribut verwendet - allerdings konnten keine doppelten Einträge festgestellt werden.

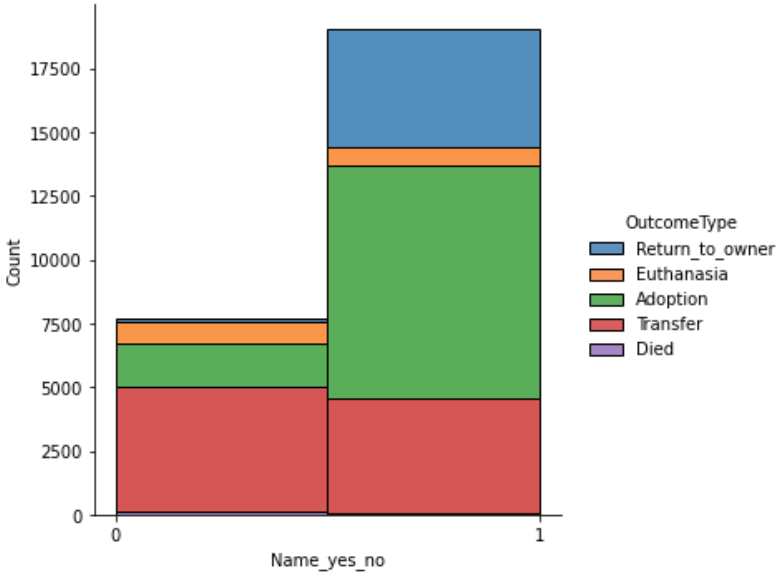
Attribut 2: “Name”

Bei dem Attribut “Name” könnte die grundsätzliche Unterscheidung zwischen Tieren mit Namen und ohne Namen relevant sein. In einem ersten Schritt soll daher überprüft werden, wie viele Tiere, unterteilt nach Hund und Katze, überhaupt einen Namen haben. Es zeigt sich, dass 71,226% der Tiere Namen besitzen, wobei Hunde im Vergleich zu Katzen deutlich öfters.



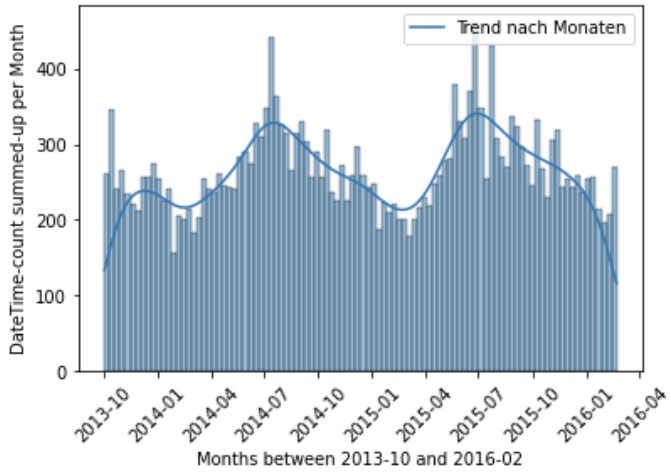
Um zu überprüfen, ob dieses Attribut in das zu modellierende Vorhersagemodell hat, wird der Einfluss auf die Variable “OutcomeType” untersucht. Die Ergebnisse deuten darauf hin, dass das Attribut insbesondere Auswirkungen darauf hat, ob ein Tier adoptiert oder an den Besitzer zurückgegeben wird. Dementsprechend wird das Attribut 2 zur Modellierung des Vorhersagemodells herangeszogen. Wie bedeutend das Attribut in Wirklichkeit ist, lässt sich anhand der Analyse allerdings nicht beurteilen. Hier ist zu erwähnen, dass Korrelation nicht gleichbedeutend mit Kausalität ist und entsprechend keine eindeutige Aussage über den Wirkungszusammenhang zwischen der der Benennung eines Tieres und der Zukunft getätigt werden kann. Dies gilt unter anderem auch, weil die Analyse der Auswirkungen des Attributs auf

andere Merkmalsausprägungen wie Verlegung oder Einschläferung keine relevanten Ergebnisse liefern.



Attribut 3: "DateTime"

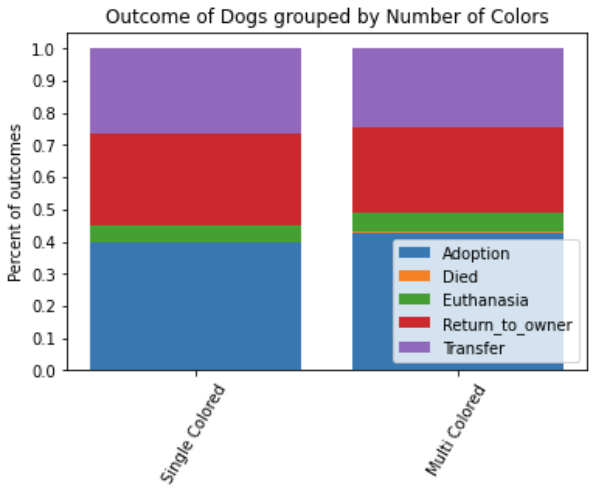
Das dritte zu betrachtende Attribut "DateTime" enthält Informationen über den Zeitpunkt beim Eintreten eines Outcome-Ereignisses. Hier ist insbesondere festzustellen, dass Tiere im Sommer um den Monat Juli öfters adoptiert, vermittelt oder euthanasiert werden. Dementsprechend wird das Attribut "DateTime" zur Modellierung des Vorhersage-Modells herangezogen.

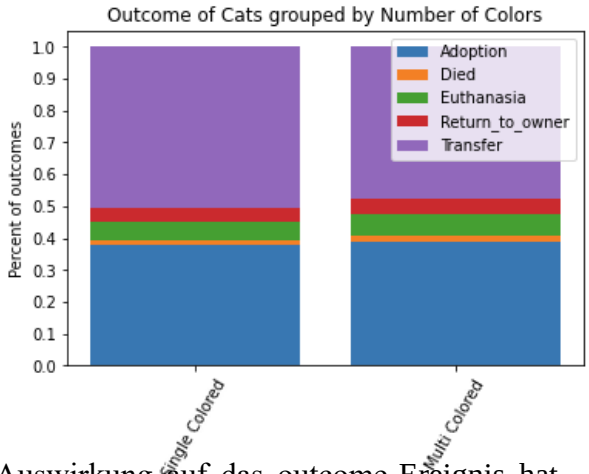


Attribut 4: Color

Das Attribut "Color" wird zur Analyse unterteilt in Tiere die eine einzelne Farbe besitzen und Tiere die mehrere Farben haben. Zusätzlich erfolgt eine separate Betrachtung von Hunden

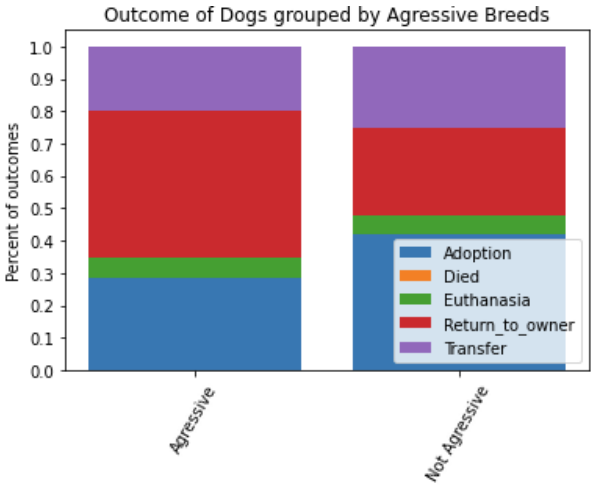
beziehungsweise Katzen. Es lässt sich erkennen, dass Farben, keine starke Auswirkung auf das Outcome-Ereignis besitzen.

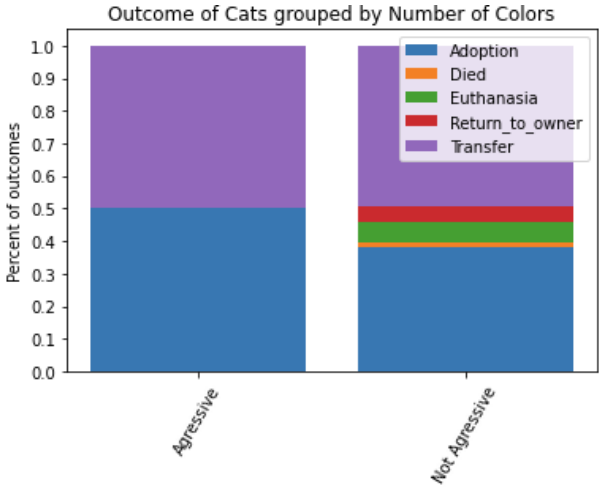




Attribut 5: Breed

Um zu Überprüfen, ob das Attribut “Breed” eine Auswirkung auf das outcome-Ereignis hat, werden Hunde als auch Katzen in aggressive bzw. nicht-aggressive Rassen eingeteilt. Problematisch ist hierbei, dass es zu aggressiven Hunden und Katzen nach der in der vorliegenden Arbeit verwendeten Einteilung nicht ausreichend viele Einträge im Datensatz enthalten sind - so sind 24 Katzen und 162 Hunde von 26.729 Tieren aggresiv.

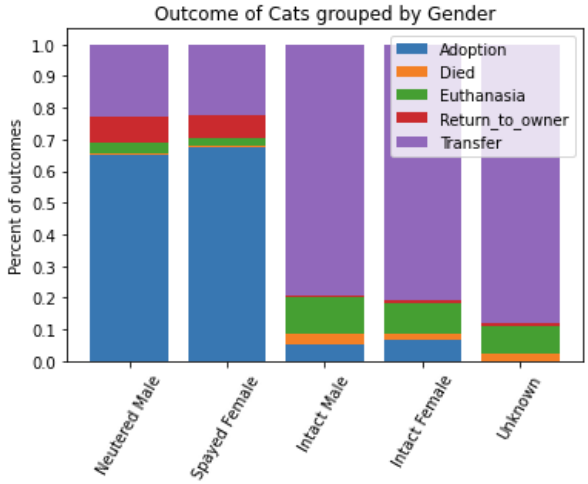




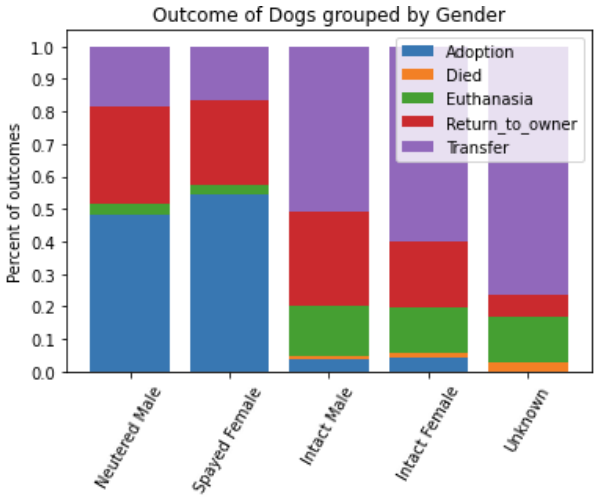
Attribut 6: SexuponOutcome

Das Geschlecht wird im ursprünglichen Datensatz in folgende Ausprägungen unterteilt:

Intact female: weiblich, nicht kastriert; Intact male: männlich, nicht kastriert; Neutered male: männlich, kastriert; Spayed female: weiblich, kastriert und Unknown: unbekannt. Es soll nun überprüft werden, ob das Geschlecht in Verbindung mit der (nicht-)Kastration eines Tieres in einem Zusammenhang mit dem Outcome-Ereignis stehen. Dazu wird der Datensatz zunächst in Hunde und Katzen aufgeteilt und der Zusammenhang betrachtet. In Bezug auf Katzen zeigt sich folgender Zusammenhang:

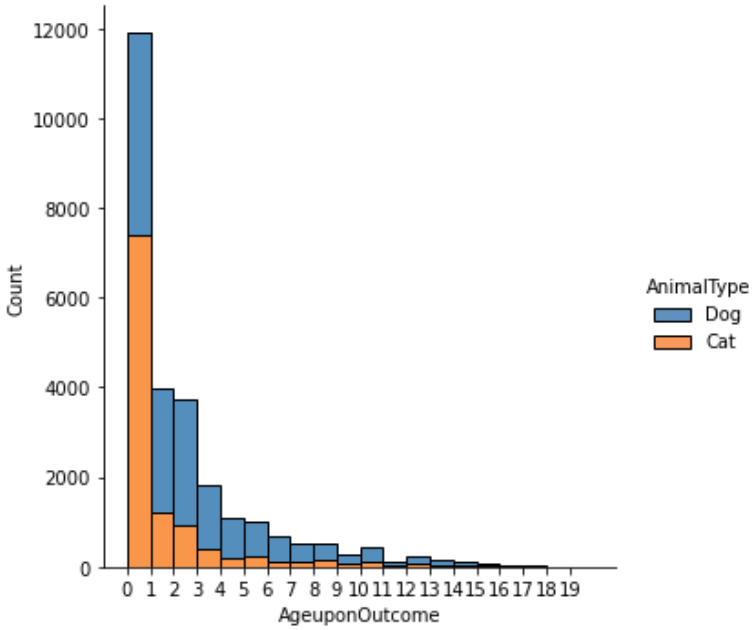


Zentral ist hierbei die Aussage, dass weibliche und männliche kastriere Katzen deutlich häufiger adoptiert werden, wohingegen männliche und weibliche nicht kastrierte Katzen häufiger transferiert werden. In Bezug auf Hunde zeigt sich bei der Analyse der Daten ein ähnliches Bild. Im Unterschied zu den Katzen, werden allerdings mehr Hunde an den Vorbesitzer zurückgegeben und das sowohl bei kastrierten als auch nicht kastrierten männlichen und weiblichen tieren.



Attribut 8: "AgeuponOutcome"

Das Attribut “AgeuponOutcome” liefert Daten für das Alter eines Tieres zum Zeitpunkt des Eintretens eines Outcome-Ereignisses. Hier wird argumentiert, dass das Alter eines Tieres einen Einfluss auf die einzelnen möglichen Ereignisse hat. Dazu wird in einem ersten Schritt das Alter von sowohl Katzen als auch Hunden zum Zeitpunkt des Eintretens eines Outcome-Ereignisses analysiert. Anhand der Nachstehenden Grafik wird erkenntlich, dass bei ungefähr 16.000 Tieren bereits nach einem Jahr in dem Tierheim ein Outcome-Ereignis stattfindet - das sind rund 45% der gesamten Tiere in dem Datensatz. Wichtig ist weiter, dass nur rund 1 % der Tiere über 14 Jahre alt ohne das Eintreten eines Outcome-Ereignisses im Tierheim verbleiben. Außerdem zeigt sich, dass eine Verteilung vorliegt.



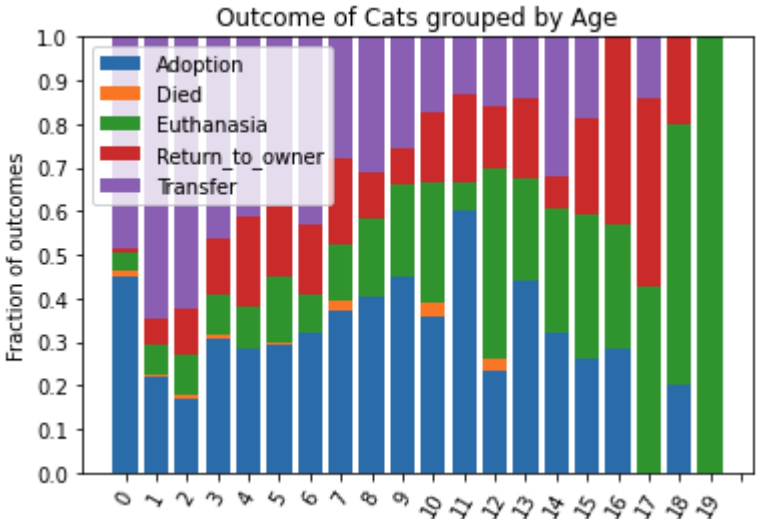
Im Rahmen weiterführender Analysen sollen die Wirkungsrichtung des Alters unterteilt nach Tierart und nach den einzelnen möglichen Outcome-Ereignissen untersucht werden.

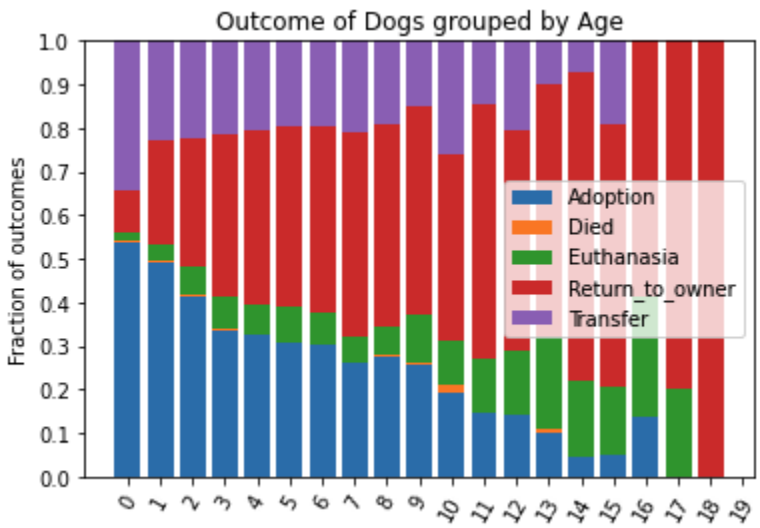
In Bezug auf Katzen lassen sich eine klare Korrelation zwischen dem Alter und Verlegung, Rückgabe an den Besitzer und insbesondere Euthanasie feststellen. Zu dem Outcome-Ereignis Adoption bzw. Einschläferungen können keine eindeutigen Schlussfolgerungen gezogen werden.

Für Hunde lässt sich grundsätzlich für die Outcome-Ereignisse Adoption, Verlegung, Rückgabe an den Besitzer und Euthanasie eine linerare Korrelation erkennen. Insbesondere in Hinblick auf die Adoptions- und Rückgabe an den Vorbesitzer Wahrscheinlichkeit lässt sich ein starker

Zusammenhang feststellen: So werden insbesondere junge Hunde viel wahrscheinlicher adoptiert als ältere und alte Hunde am häufigsten an den Besitzer zurückgegeben.

Im direkten Vergleich zwischen der Gruppe der Katzen und Hunde zeigt sich, dass Katzen eine höhere Verlegungsrate haben und im steigenden Alter häufiger euthanasiert werden. Hunde werden im Vergleich zu Katzen grundsätzlich häufiger an die Besitzer zurückgegeben und seltener euthanasiert.





4 CRISP-DM: Data Preparation

In dieser Projektphase Data Preparation werden die zuvor identifizierten Probleme nun im Rahmen eines Bereinigungsprozesses behoben werden. Zunächst soll eine Selektion der Daten durchgeführt werden, um ein Fundament für einen validen Data Mining Prozess zu schaffen.

4.1 Data Selection

Die folgende Tabelle stellt den Prozess der Selektion dar, welcher zur Identifizierung der relevanten und irrelevanten Daten durchgeführt wurde. Grundlage hierfür sind die Ergebnisse aus den vorherigen Plots und Hypothesen, welche in Zusammenhang mit der Fragestellung überlegt wurde (siehe Spalte “Kommentar”).

Attribut	**Selektion**	**Kommentar**
Animal ID	nein	Irrelevant
Name	ja	Katzen mit Namen werden eher adoptiert als ohne
Date Time	ja	Im Sommer werden eher Tiere adoptiert, als im Winter (Jahreszeiten aus Datumswerten extrahieren)
Outcome Type	ja	
Outcome Subtype	nein	Viele Missing Values, irrelevant
Animal Type	ja	Mehr Hunde werden zu Besitzer zurück gegeben
Gender (aus Attribut SexuponOutcome gesplittet)	ja	Könnte interessant sein, ob mehr weibliche oder männliche Tiere adoptiert werden
Neutered (aus Attribut SexuponOutcome gesplittet)	ja	Könnte interessant sein, ob kastrierte Tiere mehr adoptiert werden (actionable step: mehr Kastrationen durchführen um mehr Adoptionen)
AgeUponOutcome	ja	Junge Tiere werden eher adoptiert
Breed	nein	Unterteilung in aggressiv/nicht aggressiv hat zu zu wenigen Datenpunkten geführt.
Color	nein	Keine Auswirkung in den Plots zu erkennen.

4.2 Data Cleaning, Transformation und Integration

Im nächsten Schritt „Data Cleaning“ des Projektes werden die (fehlenden) Daten näher betrachtet und gesäubert. Zunächst wurde aus dem Attribut „Date Time“ die Jahreszeiten ermittelt und somit ein neues, separates Attribut Season erstellt, um zu analysieren, ob die Tiere saisonbedingt adoptiert werden.

Für die fehlenden Werte aus den Attributen „Gender“ und „Neutered“, welche aus dem Attribut „SexuponOutcome“ extrahiert wurden, wurde eine Hypothese aufgestellt mit Hilfe dessen das Geschlecht und der Status abgeleitet wird. Es wurde davon ausgegangen, dass die Pfleger im Tierheim bei der Erkennung des Geschlechts bei weiblichen Tieren Schwierigkeiten hatten, ob ein Tier kastriert war oder nicht, da die Operationen im Körper des Tieres stattgefunden haben und somit nicht offensichtlich sind. Daraus wurde die logische Schlussfolgerung gezogen, dass die Daten mit keinem Wert oder „unknown“ des Attributes „SexuponOutcome“ auf weiblich kastrierte Tiere hinweisen könnten, welche dementsprechend in den Datensatz übernommen wurden.

Aus dem Attribut „Breed“ wurden ebenfalls weitere Attribute, wie „BreedMix“, welche beschreiben soll ob die Tiere Reinrassig sind oder nicht, und „Aggressive“, welches zeigen soll, ob die Tiere aggressiv sind oder nicht. Dies wurde anhand von einer Internetrecherche festgelegt (Petsdeli & Htgetrid). Jedoch hat sich im Verlauf gezeigt, dass dieses Attribut nicht relevant für die weiteren Schritte ist, da der Anteil der Werte aggressiver Tiere nur sehr gering ist und somit für eine Analyse wenig aussagekräftig wären.

Bei dem Attribut „AgeuponOutcome“ wurden die fehlenden Werte mit dem Durchschnitt von 1 Jahr belegt. Alle Werte wurden in Tage umgewandelt und normalisiert, indem der größte Wert (MaxAge) in diesem Attribut bestimmt wurde und die einzelnen Daten durch diesen Wert geteilt wird.

Um die Daten noch besser und genauer analysieren zu können, wurde der Datensatz nochmal in Katzen und Hunde aufgeteilt. In der nachfolgenden Tabelle ist zu erkennen, wie die Daten numerisch oder binär-numerisch, umgewandelt wurden, mit Ausnahme des Attributes „OutcomeType“, welches den Datentyp String behält.

Attribut	**Normalisierung**
Name	Tiere mit Namen = 1, Tiere ohne Namen = 0
Date Time	Keine Missing Values, Monat als separates Attribut (Jahr
Outcome Type	Keine Missing Values
Season	Winter 1, spring 2, summer 3, autumn 4
Animal Type	Hund = 0, Katze = 1
Gender	Male = 0, Female = 1
Neutered	Kastriert und Sterilisiert = 1, Intakt = 0
AgeUponOutcome	Missing Values auf 1 Jahr setzen (Durchschnittsalter), W
Breed	Attribut Mix = 1, Reinrassig = 0; Attribut aggressiv = 1,

5 CRISP-DM: Modeling

5.1 Model Selection

Das Ziel der vorliegenden Arbeit ist es, mithilfe einer Vielzahl von Attributen eine Vorhersage darüber zu treffen, ob ein Tier adoptiert, eingeschläfert, verlegt oder an an den Besitzer zurückgegeben wird. Es sind entsprechend geeignete mathematische Modelle im Bereich des "Supervised learning" heranzuziehen, welche ein höchstmaß an vorhersagegenauigkeit ermöglichen und das vorliegende Klassifikationsproblem lösen. In der Literatur werden zahlreiche Modelle vorgestellt und diskutiert. Zu nennen sind zum Beispiel die logistischen Regression, Entscheidungsbäume, Random Forest, Support Vector Maschines, k-nearest-neighbour, neuronale Netze oder der naive bayes Schätzer. Aufgrund des Umfangs der Seminararbeit und der Vorteile in Bezug auf die vorhandene Datenstruktur werden einerseits die logistische regression und andererseits das random Forest Verfahren herangezogen und auf ihre "Performance" in Bezug auf die Vorhersagegenauigkeit verglichen. Beide Modelle werden zusätzlich mit einem simplen Basismodell ("baseline performance") gegenübergestellt, welche anhand eines "DummyClassifiers" (hier: "most-frequent") berechnet wird. Um ein Overfitting zu vermeiden, muss der Datensatz in einen Trainings- und einen Testdatensatz aufgeteilt werden. Ein indikator für ein Overfitting wäre mitunter, wenn die sich Performance des Modells durch Kreuzvalidierung mit dem Trainingsdatensatz signifikant von der gemessenen performance beim Testdatensatz mit demselben Modell unterscheidet. Als Methode zur Aufteilung des Datensatz wurde die geschichtete Zufallsstichprobenziehung ("stratified sampling") mit einer Trainingsmenge von 80% verwendet. Damit wird eine gleichmäßige Aufteilung der Zielklassen in beiden Datensätzen erreicht. Zusätzlich wurde der verwendete Trainingsdatensatz aufgrund der strukturellen gegebenheit in einen Datensatz für Hunde und einen Datensatz für Katzen aufgeteilt.

Als Bewertungskriterium dient neben einer einfachen Genauigkeitsbetrachtung ("Accuracy") der Kreuz-Entropie-Verlust oder "logistic loss" ("log-Verlust"), welcher sich aus der logarithmischen Verlustfunktion ergibt und im Gegensatz zur weitaus verbreiteten einfachen Genauigkeitsbetrachtung (Anzahl richtiger Entscheidungen/Anzahl aller Entscheidungen) eine bessere Entscheidungsgrundlage beim Vergleich der Modelle liefert. Der Log-Verlust ist ein Wahrscheinlichkeitsmaß für die Genauigkeit und kann dabei Werte zwischen 0 und unendlich

annehmen, wobei kleinere Werte eine bessere Vorhersagegenauigkeit bedeuten. Einfache Genauigkeitsbetrachtung, die die Anzahl richtiger Klassifizierungen zählen, werden regelmäßig durch zum Beispiel ungleichmäßige Verteilungen beeinflusst.

Zusätzlich wurden k-fache Kreuzvalidierungen durchgeführt. Mithilfe des Verfahren kann die Güte eines Modells gemessen werden ohne es mit dem Testdatensatz vergleichen zu müssen und erlaubt es ein mögliches Overfitting zu erkennen. Dafür werden die Trainingsdaten in k zufällig gleich große Blöcke (oder “Folds”) aufgeteilt. Dabei wird ein Datenblock zum Testen der Daten verwendet, während die restlichen Blöcke zum Trainieren des Modells herangezogen werden. Dieser Vorgang wird k mal wiederholt bis jeder Block zum Testen verwendet wurde. In der vorliegenden Arbeit beträgt die Anzahl der Blöcke k=5.

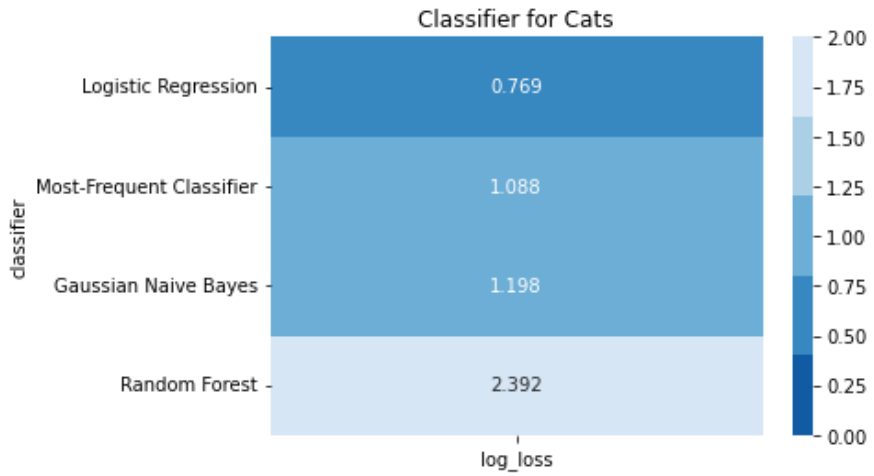
Die logistische Regression ist ein in der Fachliteratur weit verbreitetes Modell zur Klassifikation von Datensätzen. Dies gilt insbesondere deshalb, weil es im Vergleich zu anderen Modellen einfach implementiert und interpretiert werden kann und eine hohe Recheneffizienz bietet. Zu beachten ist allerdings, dass die logistische Regression Linearität annimmt und anfällig für Ausreißer ist.

Als eine Weiterentwicklung zum einfachen Entscheidungsbaum, wird das Random Forest Verfahren betrachtet. Das Random Forest Verfahren besteht grundsätzlich aus einer Vielzahl von Entscheidungsbäumen, die individuell ein Klassifizierungsergebnis berechnen. Die insgesamt am häufigsten berechnete Klassifizierung der einzelnen Entscheidungsbäume bildet die Vorhersage. Das Random Forest Verfahren verspricht im Vergleich zum einfachen Entscheidungsbaum eine bessere Vorhersagegenauigkeit. Nachteil des Verfahrens ist allerdings, dass hohe Rechenressourcen benötigt werden und das finale Modell nicht visualisiert werden kann. Für die vorliegende Untersuchung eignet sich das Modell insbesondere aufgrund der hohen Anzahl an Attributen und der geringeren Anfälligkeit für Ausreißer.

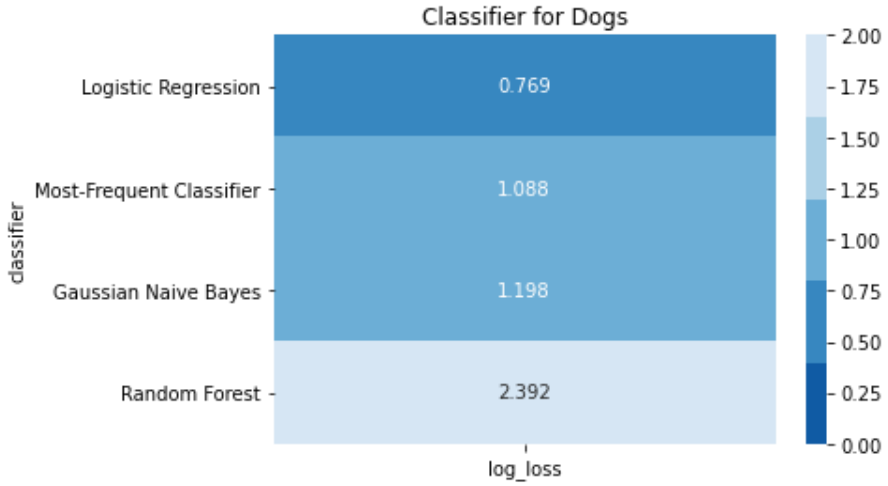
5.2 Model Comparison

5.2.1 Model Ranking

In diesem Unterkapitel sollen die der erreichten Log-Loss Scores für die betrachteten Modelle miteinander verglichen werden und eine Auswahl für die weiteren Modellierungsschritte (z.B. Optimierung) getroffen werden getroffen. Zunächst werden dafür die erreichten Scores für Katzen miteinander verglichen. Die nachstehende Grafik zeigt dabei, dass die logistische Regression mit einem erreichten Log-Loss von 0,769 die beste Modellvariante darstellt. Das schlechteste Modell ist das Random Forest Verfahren mit einem Wert von 2,392. Überraschend ist hierbei, dass das Random Forest Verfahren schlechter als unser Benchmark der Most Frequent Classifier abschneidet. Beim Vergleich der Werte der k-fachen Kreuzvalidierung kann außerdem kein signifikantes Overfitting festgestellt werden.



Als nächstes werden die erreichten Log-Loss Scores für Hunde verglichen. Zur Überraschung der Autoren gleichen sich die erreichten Log-Loss Scores mit denen für Hunde. Auch bei Katzen ist die Logistische Regression mit einem Wert von 0,769 die beste Modellvariante, wohingegen das Random-Forest verfahren am schlechtesten abschneidet. Die erreichten Werte für Hunde können der nachstehenden Grafik entnommen werden:

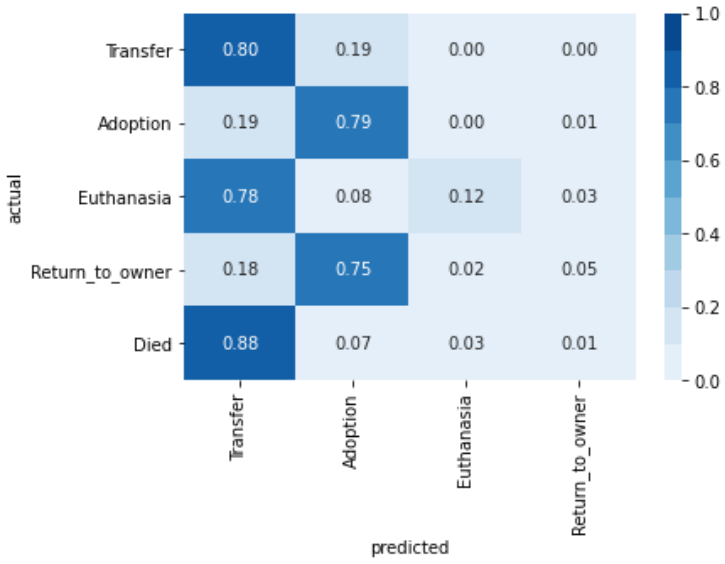


Das beste Modell für beide Datensätze ist also Logistic Regression in unserem Fall. Es lohnt sich daher die Confusion Matrix anzusehen.

5.2.2 Best Model - Confusion Matrix

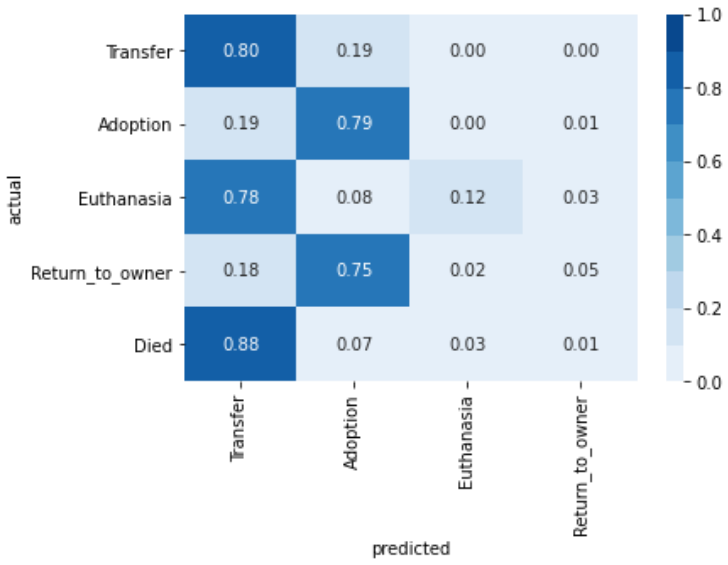
Wir benötigen eine Confusion Matrix, im Prinzip ist dass eine graphische Veranschaulichung der True Positiv, True Negative, False Positive, False Negativ Aufstellung. Da wir nicht nur 2 Klassen haben, sondern 5, haben wir uns dafür entschieden, den Logloss, also die Abweichung der korrekten Vorhersagen der Klassen, gegenüberzustellen.

Für den Katzen-Datensatz sieht die Matrix wie folgt aus:



Man kann erkennen, dass einige Klassen gut und andere weniger gut vorhergesagt werden können. Z.B. Transfer-Vorhersage ist oft falsch und hat die Möglichkeit, in Wahrheit für das Tier als Euthanasia oder als Died zu enden. Bei der Adoption wird ebenfalls oft falsch vorhergesagt, jedoch kann es auch eine Klasse Return-to-owner sein, also für das Tier ein Happy End haben.

Seltsamerweise ist die Confusion-Matrix für Hunde ebenfalls mit den gleichen Werten ausgestattet:



Die Klasse “Died” wird nicht vorhergesagt, daher fehlt in der Prediction-Seite eine Spalte.

5.3 Optimizing our Model

Mit der in diesem Modul vorgestellten Library "sklearn" gibt es zwei Möglichkeiten, unser bestes Model - die logistische Regression - zu optimieren. Zum einen die Grid-Search und andererseits die Grid-Random-Search. Beide Suchen haben zur Aufgabe, die optimalen Hyperparameter im Wertebereich des Models zu finden.

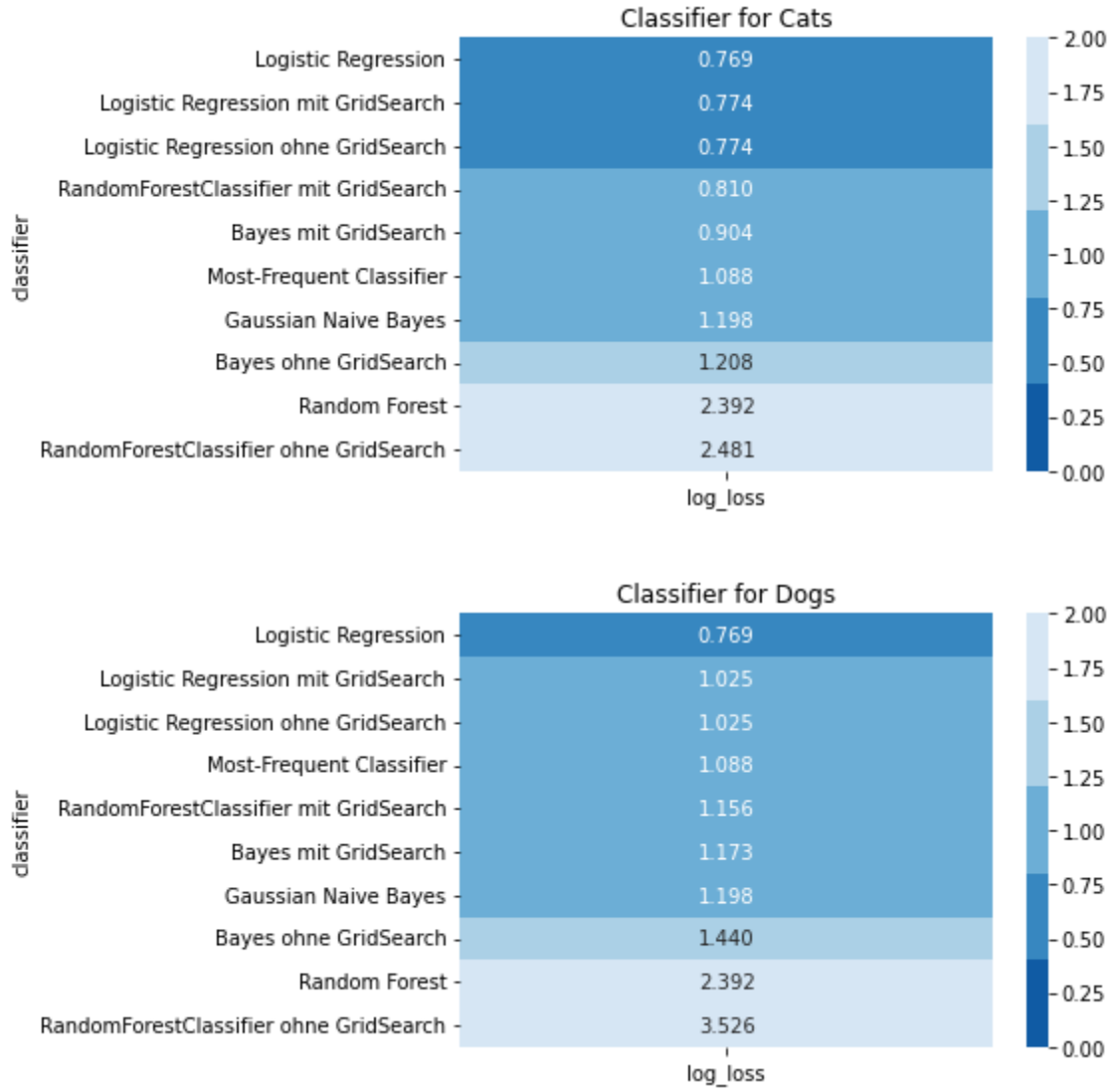
Ein Hyperparameter ist ein Parameter, der zur Steuerung des Algorithmus verwendet wird und dessen Wert im Gegensatz zu anderen Parametern vor dem eigentlichen Training des Models festgelegt werden muss.

Wir haben in unserem Projekt die Algorithmen: Logistic Regression, MultinomialNB (für Gaussian Bayes) und Random Forest mit Hilfe der Grid Search analysiert.

Wir haben nicht nur die logistische Regression optimiert, da es möglich sein kann, dass durch eine Optimierung ein anderer Algorithmus besser perfomiert.

Die Grid-Search geht alle möglichen Hyperparameter-Kombinationen systematisch durch und gibt innerhalb der von uns vorgegebenen Grenzen ein lokales Optimum.

Wir haben die Optimierung mit der Grid Search wieder auf den getrennten Datensätzen für Hunde und Katzen laufen lassen:



Wenn an der Bezeichnung keine genaue Angabe zu Grid Search ist, bitte davon ausgehen, dass Grid Search nicht verwendet wurde. Hier wurden 2 unabhängige 5 Fold-Validierungen mit Logloss betrachtet, daher können die Werte von dem einen Durchlauf (ohne Bezeichnung mit Grid Search) und die gegenübergestellten Modelle (ohne Grid Search und mit Grid Search) abweichen.

Dennoch bietet dieses Ranking der Modelle eine gute Übersicht. Man kann generell sagen, dass die Algorithmen bei dem Hunde-Datensatz einen leicht größeren Log-loss haben. Ebenfalls ist erstaunlich, dass Random-Forest bei beiden Datensätzen der schlechteste Algorithmus ist - noch schlechter als unser Benchmark der Most-Frequent Classifier.
Hervorzuheben ist auch der Unterschied bei dem optimierten Random-Forest, hier wurde durch die Optimierung erreicht, dass der Random-Forest besser als der Benchmark wurde, allerdings nur bei Katzen, bei Hunden hingegen ist er schlechter als der Benchmark.

Auch ist interessant zu beobachten, dass die Logloss Werte von der Grid Search Optimierung bei der logistischen Regression nicht wirklich einen Unterschied machen, siehe die Werte “Logistic Regression mit Grid Search” und “Logistic Regression ohne Grid Search”.

Wichtig zu erwähnen ist, dass Grid Search enorm viel Rechenleistung beansprucht. Daher sind die Grenzen der zu betrachtenden Hyperparameter von uns so reduziert worden, sodass unsere Rechner mit der Aufgabe zurechtkommen. Das bedeutet, es ist möglich, dass keine optimalen Hyperparameter gewählt wurden. Damit ist die obere Darstellung einer Rangordnung der "optimierten" Algorithmen mit sehr großer Vorsicht zu betrachten.

6 CRISP-DM: Evaluation

6.1 Comparison and Assessment

Durch unsere Analyse der zwei Datensätze und Modele konnte festgestellt werden, dass die logistische Regression der beste Klassifizierer für das Tierheim ist. Sowohl für Hunde als auch für Katzen ist unter den betrachteten Klassifizierungs-Algorithmen die logistische Regression der Beste. Eine Optimierung mit Grid-Search, so wie wir sie durchgeführt haben, hat nichts daran geändert, dass die logistische Regression der beste Algorithmus für diesen Datensatz ist.

Durch die Konfusionsmatrix in 5.2.2 Best Model - Confusion Matrix der Logistischen Regression lässt sich sagen, dass die Klassifizierung eines Tieres nach “Euthanasia” und “Return-to-Owner” besonders gut funktioniert. Bei der Klasse “Adoption” lässt sich immerhin sagen, dass dem Tier

ein Happy End zu Gute kommt, also eine tatsächliche “Adoption” oder ein “Return-To-Owner” stattfinden wird. Auf das Label “Transfer” sollte das Tierheim nicht zu sehr achten, hier ist die logistische Regression nicht gut im Vorhersagen des Schicksals.

6.2 Opportunities for improvement

Wir haben in unserem CRISP DM Prozess 3 Iterationen geschafft. In der letzten Iteration haben wir einige Attribute erkannt, die manuell aus dem Model genommen werden können, dazu zählen Color, Fellfarbe und Breed, also Rasse des Tieres. Es müsste nun verglichen werden, ob die Modele besser werden, wenn wir manuell diese 2 Attribute rausnehmen oder nicht.
Gleichzeitig muss man sagen, dass die aktuell verwendeten 9 Attribute jetzt schon keine große Anzahl für ein Model sind.

Ein weiterer Verbesserungsvorschlag wäre, weitere Algorithmen einzubinden wie Neuronale Netze, Support Vector Machines, K-nearest Neighbors oder einen Deeplearning Algorithmus. Mit ihnen könnte man ebenfalls das Ranking wie in Kapitel 5.3 aufstellen.

Eine naheliegende Möglichkeit, die aktuellen Rankings der Modelle zu verbessern, wäre mehr Rechenkapazität zu verwenden, um eine vollständige Grid-Search durchführen zu können. Damit könnte man die tatsächlich optimalen Kalibrierungen der Algorithmen vergleichen.

6.3 General recommendations for the shelter

Durch die Datenvisualisierung hat sich gezeigt, dass einige Merkmale der Tiere häufiger zur Adoption führen können, z. B. sollten Tierheime:

- ihren Tieren Namen geben,
- die Tiere kastrieren.

Diese Erkenntnisse gelten sowohl für Hunde als auch für Katzen.

7 Abschluss und Fazit

7.1 Reflexion des Vorgehens

Um die Aufgabe von Kaggle zum Outcome von Tierheimtieren zu bearbeiten, wurde CRISP-DM als Prozessmodel angewendet. Das Ziel dabei war es, ein Vorhersagemodel für Tierheime zu entwickeln, welches sie dabei unterstützt, bessere Outcomes für die Tiere zu erzielen. Dabei wurde in Kapitel 2 ein Überblick über das Projekt verschafft. Im darauffolgenden Kapitel wurde der von Kaggle gestellte Datensatz näher betrachtet und analysiert, wobei ein Verständnis für die verschiedenen Attribute, die einen Einfluss auf den Outcome der Tiere haben können, geschaffen wurde. Anschließend wurden die Daten für die Modellierung aufbereitet. Für die Modellierung wurden zwei verschiedene Modele herangezogen, welche verglichen und optimiert wurden. Zum Abschluss wurden die Ergebnisse evaluiert. Insgesamt ist das Vorgehen als sehr gut zu bewerten, da sich strikt an das Framework des CRISP-DM gehalten wurde.
Es muss jedoch beachtet werden, dass das Projekt Limitationen hat.

7.2 Ausblick

Für die inhaltliche Aufbereitung: Bitte an der generellen Aufgabenstellung orientieren (Foliensatz "Projektaufgabe", Folie 3). Konkret bedeutet dies

- Verdeutlichung der durchdachten Durchführung der einzelnen Schritte des CRISP-DM
- Gegenüberstellung der Ergebnisse von (verschiedenartig) erstellten Modellen inkl. Vergleich mit einem einfachen alternativen Ansatz, der eine Baseline-Performance vorgibt
- Dabei klare Erläuterung der Modellerstellung (z.B. Wie wurde versucht Overfitting zu vermeiden?) sowie der Evaluationskriterien (siehe z.B. Vorlesung 13, 14)
- Reflexion des gesamten Projektes (z.B. Grenzen des eigenen Vorgehens, aufgetretene/gelöste Probleme, weiterführende Möglichkeiten bzw. potenzieller Ausblick)

Literaturverzeichnis

Deutscher Tierschutzbund e.V. (2018). Ansätze zur Lösung des Straßenhundeproblems. https://www.tierschutzbund.de/fileadmin/user_upload/Downloads/Hintergrundinformationen/Ausland/Strassentierproblematik_Loesungen.pdf.

Kaggle (n.d.). Shelter Animal Outcomes. Abgerufen am 04.08.2022, https://www.kaggle.com/competitions/shelter-animal-outcomes/overview.

Sparacino, A. (2021). Thanks to Animal Shelters and Rescues, Nearly 1 Million Pets Found Homes in 2021. Abgerufen am 04.08.2022, https://be.chewy.com/thanks-to-animal-shelters-and-rescues-nearly-1-million-pets-found-homes-in-2021/.

Thelen, N. (2020). Corona-Hunde werden zum Problem. Abgerufen am 04.08.2022, https://www.tagesschau.de/inland/corona-tierheime-103.html.