Insurance for Stallions at Coolmore Stud

Business goals

Background

Despite the fact that our dataset is a training one and its main goal is to give you the opportunity to practice your skills in Data Sense, a similar business task could be found in real life. For example, Coolmore Stud is one of the largest breeders of thoroughbred racehorses. In addition to breeding thoroughbreds for competition, they also maintain farms around the world and dozens of riding schools: the farm in Australia alone has 1000 horses. Of course, both from the humane side and from the interest of maximising business profits, Coolmore Stud insures all its stallions, which is quite expensive. Insurance companies, in turn, are interested in the health status of each stallion during insurance to understand whether to provide insurance services for it and the probability of the animal's survival.

Business goals

In this case, the business goals would be:

Goal 1- Enhance Precision in Premium Pricing: Base the decision to provide insurance to a new stallion on predictions of the animal's mortality risk. Avoid insuring animals that already have a high mortality risk at the time of the insurance process. Thus, maximise the profit received from Coolmore Stud services.

Goal 2: Use the results of an implemented clustering algorithm to categorise stallions into health-based groups, allowing for more precise pricing and tailored insurance coverage and enabling the insurance company to set differentiated premiums based on the health conditions within each cluster.

Business success criteria

In this case, our business criterion would be to increase the profit generated by servicing this client. If the cost of implementing this solution was sufficiently lower than the resulting profit, then this project could be implemented.

Assessing your situation

Main dataset for the project: https://www.kaggle.com/datasets/yasserh/horse-survival-dataset

Deadline: 11.12.2023.

Risks and contingencies
We assume that the features in the dataset may not be sufficient to obtain a clear separation into non-overlapping clusters. At this stage, it can be assumed that only clusters characterised by digestive system problems can be deduced and possibly viral diseases.

Terminology

Premium Pricing
The process of determining the cost of insurance premiums is based on factors such as the health condition, age, and other risk-related characteristics of the insured stallions.

Differentiated Insurance Coverage:
Providing varied insurance coverage options and terms based on the specific risk profiles and health conditions of individual stallions or groups of stallions.
*Example**:** Stallions in a lower-risk health cluster may receive more comprehensive coverage compared to those in higher-risk clusters.*

Silhouette Score:
A metric used to assess the goodness of a clustering algorithm by measuring the separation between clusters. A higher Silhouette Score indicates better-defined and well-separated clusters.

Davies-Bouldin Index:
A metric that measures the compactness and separation of clusters in a clustering algorithm. A lower Davies-Bouldin Index indicates better clustering.

Data-mining goals

Goal 1: Develop a forecasting model to estimate the probability of stallion survival at the time of insurance.

Goal 2: To develop a model for classifying the state of health of a horse depending on the type (group) of disease.

Data-mining success criteria

- Develop a model that will correctly determine the probability of horse survival at least 75%.
- Achieve clear separation between health within at least 2 clusters.

Data Understanding

Data gathering

For our project, we will use a dataset generated by a machine-learning model trained on the “Horse Survival Dataset”. The dataset was placed on the Kaggle platform, where it is available and can be used by participants of the “Predict Health Outcomes of Horses” competition.

The dataset consists of horses’ health measurements (pulse, temperature, blood protein, etc.) and some other important descriptive aspects of horses' health that help to identify the wellness of an animal (e.g., the colour of mucous membranes, abdominocentesis appearance). The last ones are either written in words or numeric variables where each value has some corresponding descriptions. There is surely an attribute that doesn't add any value to achieving the aim of our project – hospital number. It is quite hard to say what other attributes won't have much impact on a horse's outcome prediction, as the importance of some of them is not clear for us as non-veterinarians; we plan to find out less important variables during prediction modelling.

The dataset is created on a base survival dataset from colic, so predicting a horse's outcome is relevant to be applied only to horses with gastric-related diseases. There could be some problems with processing the dataset due to many blank values for some attributes; we also must find a way to make predictions with three possible results. Some troubles could occur since many important attributes are categorical with many possible values (rate of pain, kind of surgical lesion) – processing them as dummies will increase the dimensionality of the dataset with what we will need to deal with while building a prediction model so that it won't work for too long.

Data description

The data set consists of 1235 instances and has 29 attributes. Most of the attributes are string ones, 8 attributes are decimal health measurements (pulse, temperature), and 3 are integer (categorical ones related to surgical lesions' types) attributes. The dataset has many empty values in some attributes.

Some attributes are binary values since they only have two possible string values (surgery, age, surgical lesion, capillary refill time); there are also linear decimal values, which represent some sort of health measurements (temperature, respiratory rate, PH, packed cell volume, protein). The rest of the attributes are categorical and represent some descriptive aspects of horse wellness (colours, subjective pain rates, etc.). Again, some of the attributes have many possible categories and dividing them into dummies will lead to high data set dimensionality that can slow down our prediction model.

One of our data-mining goals is to build a model that will predict the outcomes of horses with an accuracy of 75% based on the provided test dataset attributes. The dataset suits that since we found people who achieved this accuracy on the Kaggle competition board. We also aimed to try clusterisation on a given dataset. At the current moment, it's hard to say whether we will be able to cluster the data meaningfully based on some particular health aspects.

Data Exploration

There are a few numeric attributes that can be summarised. The temperature distribution is, for example, pretty balanced –50% of a dataset is made up of horses with normal or lower temperatures, and the other part is normal and higher. Some attributes, however, have more values beyond the normal; for example, the lower percentile of 'pulse' already exceeds the normal pulse rate for adults, the same as the respiratory rate. The distributions of the attributes are not similar to any typical distributions, but many of them are right-skewed.

Most of the instances are notes about adult horses. Many horses in the dataset are mentioned to be depressed (35% of the dataset) or have mild pain (33%), which may indicate that many of the health analyses from the dataset occurred because the owners noticed that something was wrong with their horses. That factor could bias future predictions as horses in the dataset may be more prone to being ill.

Some of the attributes have a lot of null values in them. For example, for attributes ‘nasogastric_tube’ and ‘nasogastric_reflux’, almost one-third of the values are skipped. Up to 20% of the ‘rectal_exam_feces’, ‘abdomen’, and ‘abdominal_distention’ columns are null as well. 5 attributes have a low percentage of null values (up to 5%); most of them are subjective criteria (colour of a mucous membrane, pain rates) or more complex health testes (peripheral pulse, capillary refill time, etc.)

Verifying data quality

It is hard to conclude the data quality clearly since a machine learning model generated it. Even though it is based on the real dataset, it still has differences in the data distributions that could influence the prediction on real-life datasets. More in-depth discussions about some aspects of dataset could be read above.

Project plan

1. Data Cleaning and Preprocessing(1 hour Nadiia & 2 hour Yana):
    a. Clean and preprocess the data to handle missing values, outliers, and ensure data quality (Nadiia)
    b. Analyze the data obtained, visualize, describe the insights found and the assumptions made that can be useful for building a prediction and clustering models (Yana)

Inputs: Raw data
Outputs: Cleaned data, preprocessing report(including insights and important remarks)

2. Predictive Modeling(12 hours Nadiia & 13 hours Yana)
    a. Analyze recommendations for using certain algorithms to solve similar problems. Choose 2 models that seem to be the most relevant. (1 hour Nadiia & 1 hour Yana)
    b. Prepare the data for the model (balancing, converting data categories, etc.) (1-2 hours Yana)
    c. Implement Model 1(based on point a, turn to random forest or SVM), Cross-validation(10 hours Nadiia).
    d. Implement Model 2, Cross-validation(based on point a, turn to XGB or LGBM) (10 hours Yana).
    e. Try to prepare the data for the model differently (1 hour Yana)
    f. Validate models on data again.
    g. Choose the best model and run it on test data.
    h. Write a short report on the process and the selected model(1 hour Nadiia)

3. Health-Based Clustering(10 hours Nadiia & 10 hours Yana)
    a. Analyze recommendations for using certain algorithms to solve similar problems. Choose 2 models for clustering that seem most relevant.
    b. Implement PCA or K-means based on a (Nadiia).
    c. Implement DBSCAN or Agglomerative Hierarchical Clustering (Yana).
    d. Write a short report on the process and the chosen model.

4. Document the entire process, results, and recommendations(2 hours Nadiia & 1.5 hours Yana)

5. Prepare a poster(0.5 hours Nadiia & 0.5 hours Yana)