Wrangle Report

De maneira geral, a idéia da análise proposta foi verificar o impacto de cada estágio proposto para um animal (“doggo”, “puppo”, “floofer”, “pupper”) em relação ao número de tweets favoritos e retweets bem como a utilização da previsão de melhor qualidade realizada a cada um dos estágios.

Problemas de Qualidade

Os principais problemas de qualidade encontrados se referem a tipos de dados incorretos (como por exemplo das colunas de data/horário), dados faltantes (como por exemplo o fato de que nem todos os tweets tinham previsão relacionada) e dados incorretos (por exemplo tweets cujo denominador era ou menor que 10 ou maior que 10).

Vale lembrar que no caso de dados incorretos foram consultados os valores do texto original do tweet, que puderam ser utilizados como referência para determinar o valor correto da nota que deveria ser utilizado.

Em relação aos problemas de qualidade, a maior dificuldade encontrada se deu ao fato de verificar as diversas possibilidades de problemas e sua relação com a análise; existem por exemplo diversos tweets cujo nome do animal é inválido (como “None”, artigos e substantivos na língua inglesa, etc).

Problemas de Arrumação

Quanto aos problemas de arrumação, foram basicamente identificados dois principais: haviam nove colunas sendo utilizadas para as previsões (“p1”, “p1_conf”, “p1_dog”, “p2”, “p2_conf”, “p2_dog”), de forma que foram corretamente ajustadas para apenas três (“prediction”, “confidence_interval”, “is_dog_race”). Já quanto aos tweets, as informações do estágio do animal utilizavam quatro colunas (“puppo”, “doggo”, “floofer”, “pupper”) e foram normalizadas para duas: “phase” (indica o estágio do animal) e “is_in_phase” (indicando se o animal se encontra no determinado estágio).

O principal problema relacionado a este tipo de arrumação se deu em identificar quais as colunas relacionavam-se ao mesmo conceito e aplicar as transformações necessárias a estas colunas para simplificá-las.

Com base nesta estrutura foram aplicados alguns filtros para a análise em si.