Información del dataset

El dataset contiene información del perfil y reviews de 3197 cervezas correspondientes a 934 diferentes cervecerías, proveniente del sitio BeerAdvocate.com.

Las variables del dataset se pueden dividir en tres grandes grupos:

1) Variables que identifican y describen en general cada cerveza (nombre de la cerveza, nombre de la cervecería que la fabrica, estilo, contenido de alcohol);
2) Variables que trazan un perfil específico de cada cerveza (según sabor, gusto, aroma, etcétera);
3) Variables que brindan información respecto a las calificaciones otorgadas en las reviews de cada cerveza por parte de los consumidores (el score promedio general de calificaciones sobre las características, pero también el score promedio general de calificaciones y la cantidad de reviews de cada una de las cervezas).

Las variables descriptivas del perfil de las cervezas fueron construidas a partir de contar la cantidad de palabras encontradas en al menos 25 reviews de cada cerveza, sumando 1 punto por cada palabra asociada a la variable. Por su parte, las calificaciones otorgadas en las reviews de consumidores poseen una escala ordinal de 1 a 5 puntos, siendo 5 el puntaje más alto posible. En este sentido, las variables de review fueron construidas realizando los promedios de dicha escala.

Para más información, visitar las siguientes fuentes de datos:

* https://www.kaggle.com/datasets/ruthgn/beer-profile-and-ratings-data-set

* https://www.beeradvocate.com/community/threads/how-to-review-a-beer.241156/

Warning metodológico

El presente trabajo es el proyecto final de la carrera de Data Scientist en Coderhouse. Es decir, forma parte de un ambiente educativo, con propósitos de aprendizaje y desarrollo profesional. Realizado a conciencia y con la certeza de haber aplicado enfoques y herramientas ampliamente utilizadas en la industria, seguramente note alguna inconsistencia metodológica. De ellas, las dos más definitorias y de las cuales se debe dejar un warning son las siguientes:

1) La metodología con la cual se construyeron algunas variables del dataset no es la más recomendada: por ejemplo, la falta de estandarización en la construcción de

las variables descriptivas del perfil de las cervezas aporta una gran cantidad de ruido tanto en las relaciones entre variables como en los modelos implementados. La subjetividad que se manifiesta en la descripción de los consumidores, al menos en este caso, no es una herramienta adecuada para predecir la variable target. Es por ello que este grupo de variables, previo análisis exploratorio, se ha descartado al momento de entrenar los modelos.

2) Por otro lado, también es cuestionable la metodología de construcción del grupo de variables de review, las cuales surgen de la media de una escala de Likert. Dejando las discusiones teóricas para otra ocasión, y luego de identificar que este grupo de variables se correlaciona fuertemente con la variable target, aceptamos trabajar con ellas y refinar la interpretación y el análisis de los datos para compensar de alguna forma ese déficit de origen.

En resumen, esta advertencia no sólo se deja sentada para matizar las conclusiones de este proyecto, si no también para todas aquellos que puedan llegar a encontrarse con inconvenientes de factibilidad de los datos en el modelado similares a los aquí surgidos.

Descripción del problema de negocio

Definimos a continuación la variable objetivo y el problema de machine learning asociado a ella, a saber:

* Variable objetivo: review overall.

* Problema: predecir el review overall de las cervezas.

* Hipótesis: las características de las cervezas guardan relación con su review overall.

¿Hay relación entre las características de una cerveza y la calificación general que aquella logrará en las reviews de los consumidores?

Si es que existe esta relación, ¿Cuáles son esas variables?

¿Depende la calificación general de una cerveza de su cantidad de reviews?

¿Qué características de las cervezas son mejor valoradas? ¿Y las peor valoradas?

¿Es la descripción de las reviews una metodología adecuada para desarrollar un modelo de machine learning? ¿Y la media de una escala de Likert?

Estos, entre otros interrogantes, son los que se abordarán a continuación. En última instancia, fabricantes, comercializadores y hasta el público consumidor podrán valerse de algunas de las conclusiones de este trabajo para mejorar sus decisiones de producción, venta y/o consumo.

Finalmente, intentaremos modelar una regresión, valiéndonos de las métricas comúnmente utilizadas para medir el desempeño de dichos modelos: R2, MAE, MSE, RMSE.

Etapas

El trabajo constó de una serie de etapas, las cuales se resumen brevemente.

Carga y transformación de los datos:

La acción más importante en este paso fue establecer un piso mínimo de reviews por el cual filtrar el dataset. Siendo que la construcción de algunas variables del dataset requirió de al menos 25 reviews, utilizamos este criterio.

Análisis exploratorio de los datos:

Durante el análisis univariado trabajamos con medidas de tendencia central (promedio, mediana, cuantiles, desviación estándar, coeficiente de variación). También profundizamos en los valores outliers calculando su proporción en porcentaje respecto al total de la muestra. Para el análisis gráfico nos valimos de histogramas y boxplots. Por último, operamos una reducción de dimensionalidad en cuanto a las variables Min/Max IBU, transformándolas en otra nueva, IBU Neto.

Para el análisis bivariado empleamos los coeficientes de correlación de Pearson y Spearman, así como los gráficos de heatmap y jointplot.

Mientras que en el análisis multivariado nos valimos del cálculo de las correlaciones parciales y de los gráficos relplot, tridimensional y distplot.

Implementación de modelos:

En general, hemos realizado selección de variables, tratamiento de outliers escalando los datos, validación cruzada y optimización de hiperparámetros, y empleamos las métricas

de regresión ya mencionadas, con el aditamento de recalcularlas sólo en el segmento de mayor varianza. Esto último con el fin de escoger los modelos que no sólo capturen la variabilidad de la variable target, si no que también reduzcan lo más posible el error, que por otra parte asociamos a la presencia de outliers inferiores.

Entrenamos los siguientes modelos: regresión lineal, Ridge, árbol de decisión, random forest, gradient boosting, k nearest neighbors y support vector regressor (SVR).

Resumen de resultados

La regresión lineal múltiple, junto al SVR, fueron los modelos con mejor performance, sea para explicar la variabilidad de la variable target, o bien para reducir las métricas de error, tanto en el total de las predicciones como en el segmento identificado con mayor varianza de las mismas respecto a los valores reales. Es importante mencionar que los resultados de las métricas de error mejoran notablemente previa normalización de los datos de entrada del algoritmo.

La regresión lineal mencionada alcanzó en la fase de test un R2 de 0.93 y redujo notablemente las métricas de error: MAE=0.0267, MSE=0.0011, RMSE=0.0338. En cuanto al SVR, obtuvo un R2 de 0.91, y sus métricas de error fueron: MAE=0.0306, MSE=0.0014, RMSE=0.0374.

Aislando el segmento de mayor varianza de las predicciones respecto a los valores reales, la regresión lineal múltiple con normalización de los datos nuevamente ha sido el modelo con mejor desempeño, capturando un 0.29 de variabilidad, con una desviación estándar del 0.07, y arrojando las siguientes métricas de error: MAE=0.047, MSE=0.0034, RMSE=0.0579. Por su parte, el SVR fue el otro modelo con mejor performance en este segmento, capturando una variabilidad incluso superior a la regresión lineal (0.39), con una desviación estándar parecida (0.08), aunque sus métricas de error quedaron por detrás de aquella: MAE=0.0442, MSE=0.0029, RMSE=0.0534.

Insights

De negocio:

Se puede decir que las cervezas con mayor cantidad de reviews son las más “populares”. Sin embargo, al momento de evaluar su relación con la variable review_overall, no se aprecia una diferencia considerable en los estadísticos de tendencia central respecto a las restantes cervezas, con lo cual el número de reviews no sería una variable con gran peso

para analizar los ratings de las cervezas. El cálculo de los coeficientes de correlación confirma esta apreciación. Pese a ello, descubrimos que las cervezas con un número de reviews mayor a 550 tienen más posibilidades de obtener un review overall que destaque del resto.

En el análisis gráfico de la etapa exploratoria se pueden apreciar 3 grupos de características según las reviews de los consumidores: las que mayores puntajes poseen ('Malty', 'Sweet'); las de menores valores ('Astringency', 'Alcohol', 'Spices'); y un grupo del medio ('Body', 'Fruits', 'Hoppy', 'Bitter', 'Sour').

El grupo de variables de review mantiene masas centrales de datos y distribuciones similares, entre 3.5 y 4.25 puntos, donde ubicaríamos a las cervezas “promedio”.

Si se pretendiera modelar el problema de negocio como un problema de clasificación, las review_overall entre 3 y 3.5 puntos constituirían el segmento de cervezas “malas”. Como vimos, los valores por debajo de los 3 puntos son infrecuentes, aunque cuando ocurren, estamos en presencia de cervezas “muy malas”. En contrapartida, por arriba de los 4.25 puntos se ubicarían las cervezas “muy buenas”.

Metodológicos:

Salty es una variable con bajo poder de predicción dada la gran cantidad de valores 0 presentes en las reviews (más de la mitad). Cualquier intento de incluirla en el entrenamiento de los modelos introducirá ruido.

No es una buena idea filtrar la muestra excluyendo las cervezas con mayor número de reviews. Los coeficientes de correlación crecen (2 puntos aproximadamente) pero quedan fuera más de la mitad de las cervezas con mayor review_overall, modificando la distribución de la variable target y afectando potencialmente la capacidad predictiva de los modelos.

Incluir todas las variables descriptivas en una única variable, aprovechando que se encuentran en la misma escala, no eleva el coeficiente de correlación con la variable review_overall (se queda en 0.45). Esto vale tanto teniendo un criterio amplio (incluyendo variables descriptivas con coeficientes iguales o mayores a 0.2 con la variable target) como teniendo un criterio más selectivo (incluyendo sólo variables descriptivas con coeficientes cercanos o mayores a 0.3 con la variable target). Sea calculando el coeficiente de correlación de Pearson o el de Spearman.

Agrupar las variables descriptivas en los grupos sugeridos por el contexto de negocio (Mouthfeel, Taste, Flavor&Aroma) no alcanza para que los coeficientes de correlación con la variable target alcancen relaciones fuertes de tipo lineal y no lineal.

Consideraciones Finales

impacto en el negocio

El review_overall de un producto/servicio se puede predecir a partir de desagregar las reviews en dimensiones que lo describen más específicamente. Esto ayuda tanto a concentarse en evaluar aspectos más concretos de los mismos (en nuestro caso, aroma, gusto, paladar, etcétera, de las cervezas) como a que el puntaje global guarde relación directa con aquellos. Desde una mirada metodológica colabora en que el reviewer focalice su atención en la experiencia sensorial del consumo, ya que realizar el ejercicio de evaluación general de cualquier cosa requiere un nivel de abstracción que no tiene que ver con el sentido inmediato que tienen las reviews.

Por ejemplo, los modelos de árbol/es aportaron gráficamente una interesante conclusión: en la experiencia de una cerveza es más importante el paladar que el aroma o la apariencia (variable ésta última eliminada de los modelos). Esto también se afirma calculando la importancia de los predictores en los modelos en los cuales se encuentra disponible este atributo.

En ese sentido demostramos que, con algunas limitaciones y adecuando la interpretación de negocio, sí se puede trabajar con los promedios de las escalas de Likert en machine learning: esto es especialmente útil para todos los modelos de negocio basados en reviews (desde deliveries hasta plataformas on demand, por mencionar los casos más paradigmáticos).

Agradecimientos

Al profesor Iair Moisés Linker San Juan y al tutor Leyton Jean Pierre Castro Clavijo por el feedback y las sugerencias realizadas.