American Porter & American Stout

Bayesian Recommendation

Daniel Robles Leong

Motivación Mi parte favorita de estudiar Ciencia de Datos, es la capacidad de los modelos estadísticos para resolver dudas, basado en observaciones se intenta inferir y predecir casos generales. Este proyecto mas que ser ambicioso, busca contestar una pregunta que me surgio cuando recién me introducia en estos temas, el contexto es en el siguiente:

Explorando locales por la ciudad, encuentras un establecimiento que ofrece a sus clientes cervezas importadas de todo el mundo, por curiosidad decides entrar y el mesero te pregunta ¿Que desea ordenar?, es entonces que surge la pregunta ¿Que me recomiendas?

La pregunta es muy simple y diversas respuestas son aceptables, el mesero en su experiencia, puede recomendar su favorita (Experiencia Personal), la más popular (Por la que más preguntan), la mas vendida (Una opción popular y económica) o la mejor puntuada por otros usuarios (reseñas en linea).

Fue mi primer acercamiento al problema, sin embargo, investigando, descubrí el sub-mundo de variedades que existen, por poner un ejemplo las cervezas pueden variar por acidez, por la técnica de fermentación, por la levadura, ingredientes extras añadidos, hasta porcentajes de alcohol para diferentes propósitos. En total, según el Beer Judge Certification Program institución encargada de juzgar competencias cerveceras alrededor del mundo existen 115 estilos clasificados en 32 categorías.

El objetivo principal es cuantificar la relación entre dos estilos dentro de la misma categoría basados en la experiencia de los usuarios, obtenida de reseñas perceptivas. Ambiciosamente se intenta predecir la recepción de una cerveza basado en la experiencia previa de otra asociada al mismo estilo o la misma categoría

Esta propuesta surge a partir de datos recopilados de libre acceso por parte de la comunidad cervecera internacional, compuestos por información sobre aproximadamente 350,000 cervezas y 17 millones de reseñas sobre estas (Recopiladas de ratebeer.com[1]). Además de contar con clasificaciones sobre familias de cervezas según la *Beer Judge Certification Program Guidelines 2021* [2].

American Porter and Stout Se decidío utilizar esta categoria por fines prácticos además de su popularidad, nos centraremos en 5 cervezas de cada estilo sindo las mas informativas respectivamente como se muestra en la Figura 5, procederemos a explicar su relación citando a la guía:

"These beers all evolved from their English namesakes to be wholly transformed by American craft brewers. Generally, these styles are bigger, stronger, more roast-forward, and more hop-centric than their traditional Anglo cousins. These styles are grouped together due to a similar shared history and flavor profile."

Id	Porter Beer	# reseñas
P1	Anchor Porter	1330
P2	Great Lakes Edmund Fitzgerald Porter	1665
P3	Mocha Porter	1292
P4	Sierra Nevada Porter	1253
P5	Stone Smoked Porter	1643

Id	Stout Beer	# reseñas
S1	Bell's Kalamazoo Stout	1287
S2	Bell's Special Double Cream Stout	1205
S3	Chicory Stout	1240
S4	Chocolate Stout	1810
S5	Sierra Nevada Stout	1230

Figura 1: Cantidad de Reseñas por Estilo y por Cerveza

Como medida para diferencias ambos estilos tenemos las llamadas estadísticas vitales, estos son valores números sobre elementos de fabricación, de esta forma podemos observar que a pesar de ser miembros de la misma categoría y estar relacionados por sabor, mantienen cierta diferencia relevante.

- IBU: *International Bitterness Unit*, es la unidad utilizada para medir el amargor de la cerveza.
- SRM: *Standard Reference Method* es un sistema de referencia para medie el color de las cervezas.
- OG: *Original gravity* mide la cantidad de azúcar presente en el mosto antes de que se fermente. La gravedad final
- FG: *Final Gravity* es la cantidad de azúcar que queda cuando se realiza la fermentación.

- AVG *Alcohol By Volume* Se utiliza para medir el contenido de alcohol de la cerveza, el vino, los licores destilados y otras bebidas alcohólicas

Porter	
IBU	25 - 50
SRM	22 - 40
OG	1.050 - 1.070
FG	1.012 - 1.018
ABV	4.8 % - 6.5 %

Stout	
IBU	50 - 90
SRM	30 - 40
OG	1.075 - 1.115
FG	1.018 - 1.030
ABV	8 % - 12 %

Figura 2: Estadisticas Vitales por Estilo de Cerveza

Best Models Selection Cada una de las reseñas anteriormente mencionadas contiene 5 variables cuantitativas relacionadas a una reseña perceptiva de la cerveza que toma valores entre 0 y 5, especificamente tenemos:

- Reseña General (RG)
- Reseña Aroma (RA)
- Reseña Visual (RV)
- Reseña Sabor (RS)
- Reseña Sensación en Boca (RB)

Nuestro primer paso será ajustar un modelo donde **PG** sea nuestra variable respuesta y el error sea minimo, esto para cada uno de los estilos de cerveza, visualizaemos como se relacionan nuestros datos independientemente con la variable respuesta, aplicamos el método jitter para visualizar, esto al sumar un ruido $\sim N(0, 0.15^2)$

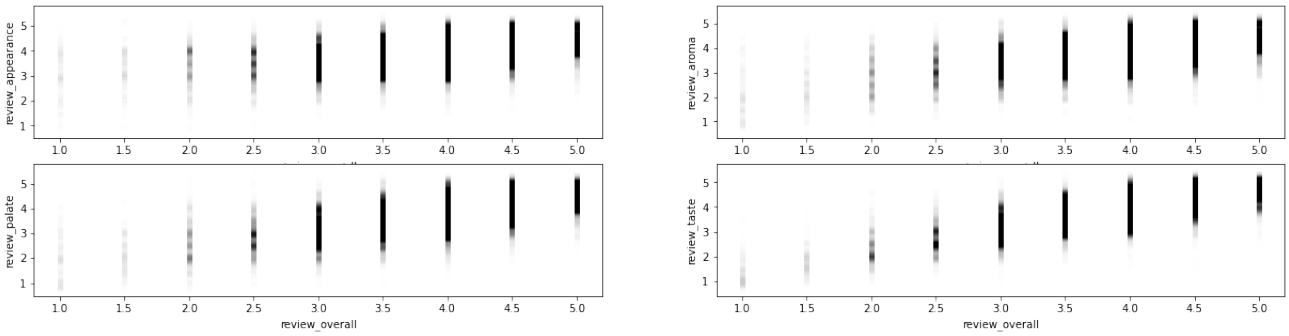

Figura 3: Porter Scatterplot

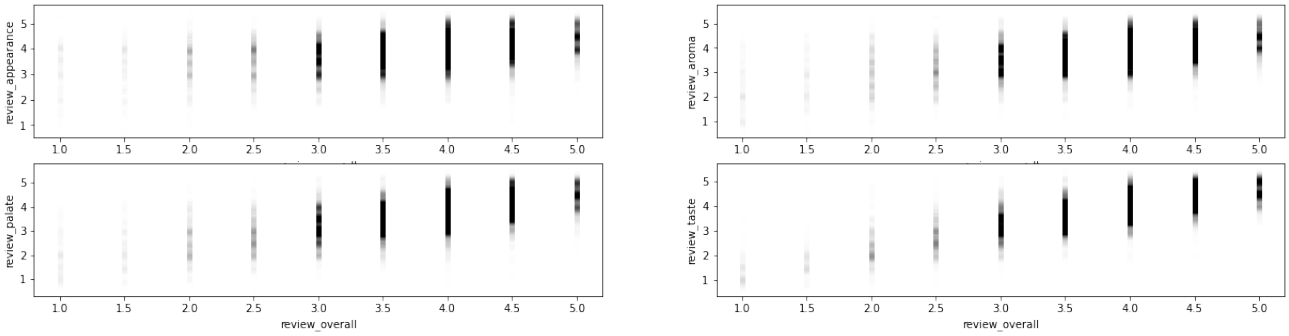

Figura 4: Stout Scatterplot

A simple vista, se puede observar una correlación positiva por parte de todas las variables, ajustando cada uno de los modelos obtenemos:

Porter Model	ECM (10-fold Cross Validation)
Básico	0.14466
Completo	0.14136
Reducido	0.14138

Stout Model	ECM (10-fold Cross Validation)
Básico	0.15266
Completo	0.14961
Reducido	0.14961

Figura 5: Error Cuadrático Medio por Modelo

Se consideraron 3 modelos para cada estilo de cerveza, un modelo básico considerando todas las variables de la base de datos, un modelo completo que incluye todas las interacciones entre variables y un modelo reducido bajo el criterio AIC, en cada uno de ellos se realizo un 10-fold Cross Validation, calculando el ECM, definido como:

$$ECM = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}$$

En ambos casos las variables básicas resultaron ser estadísticamente significantes, además el considerar interacciones entre variables no reduce significativamente el ECM, debido a esto procederemos a calcular los intervalos de confianza con el modelo básico, es decir:

$$RG \sim RA + RV + RS + RB$$

Prior Assumptions Nuestra información a priori estará dada por los valores obtenidos de nuestros datos, observemos nuestras distribuciones por cada una de las cervezas, podemos asumir que pertencecen a la familia normal y tomar la media y desviación estandar respectiva.

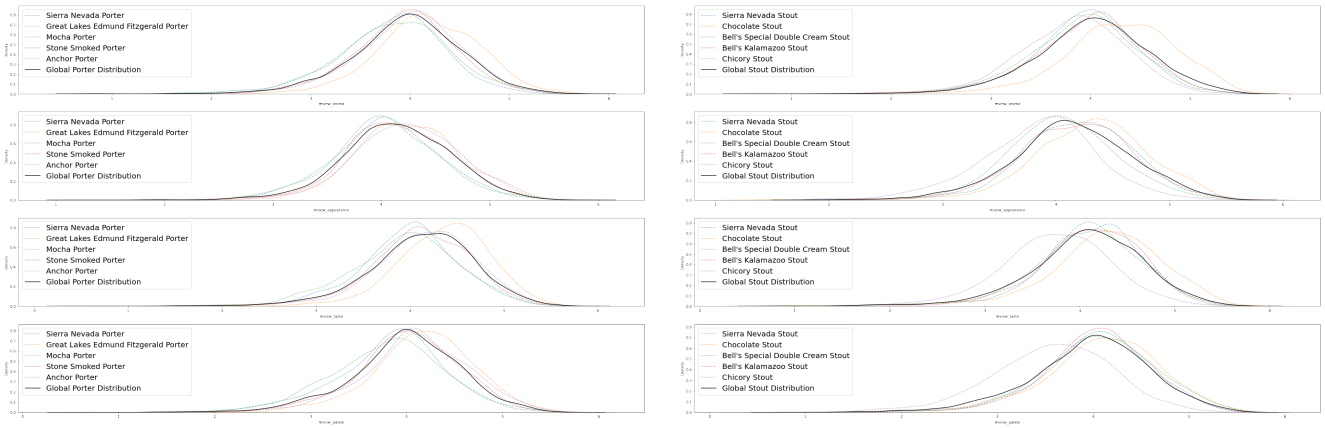

Figura 6: Distribuciónes por cervezas Porter y Stout

RJAGS simulation for multivariate regression Proceremos ajustar nuestro modelo de regresión lineal multivariable bayesiano con las variables obtenidas en nuestra selección de modelo, el cual es el siguiente para ambos casos:

$$RG_{pred} = a + b \cdot RA + c \cdot RV + d \cdot RS + d \cdot RB$$

Calculamos las estimaciones de nuestras variables mutilizando JAGS, con una cadena MCMC con 10,000 iteraciones, como observación en ambos casos podemos confirmar visualmente la convergencia además de una baja incertidumbre en las distribuciones de nuestras variables.

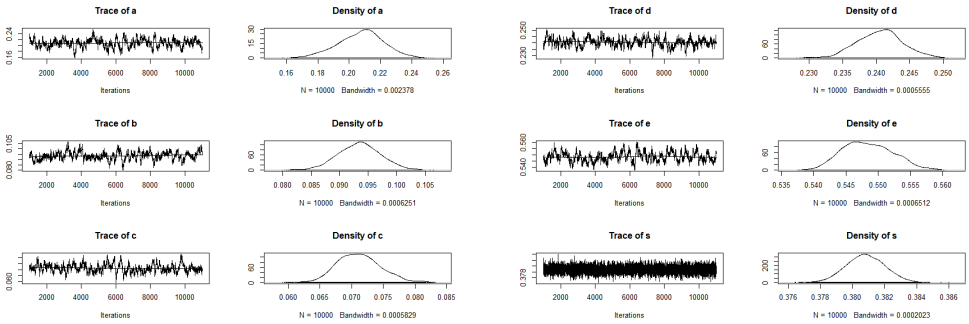

Figura 7: Simulación RJAGS para Porter

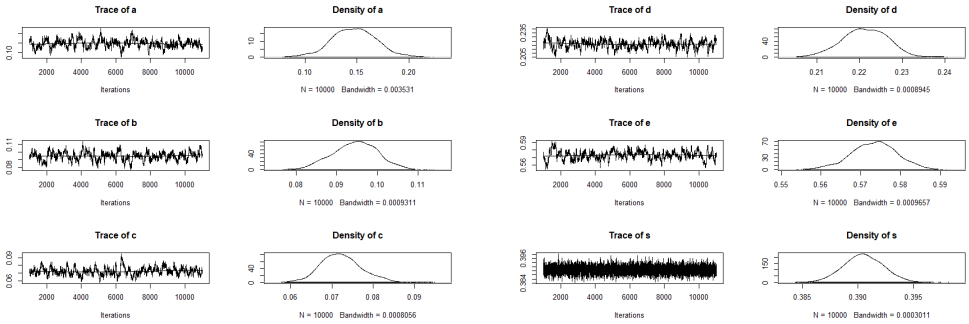

Figura 8: Simulación RJAGS para Stout

Estimaciones Sobre Usuarios Una vez obteniendo nuestras cadenas de markov, lo cual nos permitirá crear intervalos de confianza para nuestras predicciones, sin embargo, nuestra predicción depende de la recepción perceptiva de la cerveza, recordemos que nuestro objetivo es intentar predecir esta recepción basada en un elemento similar dentro de la misma categoría.

Para esto observaremos la recepción de los 2 estilos, utilizando reseñas de 72 usuarios sobre el promedio de las 5 cervezas por categoría mencionadas con anteioridad.

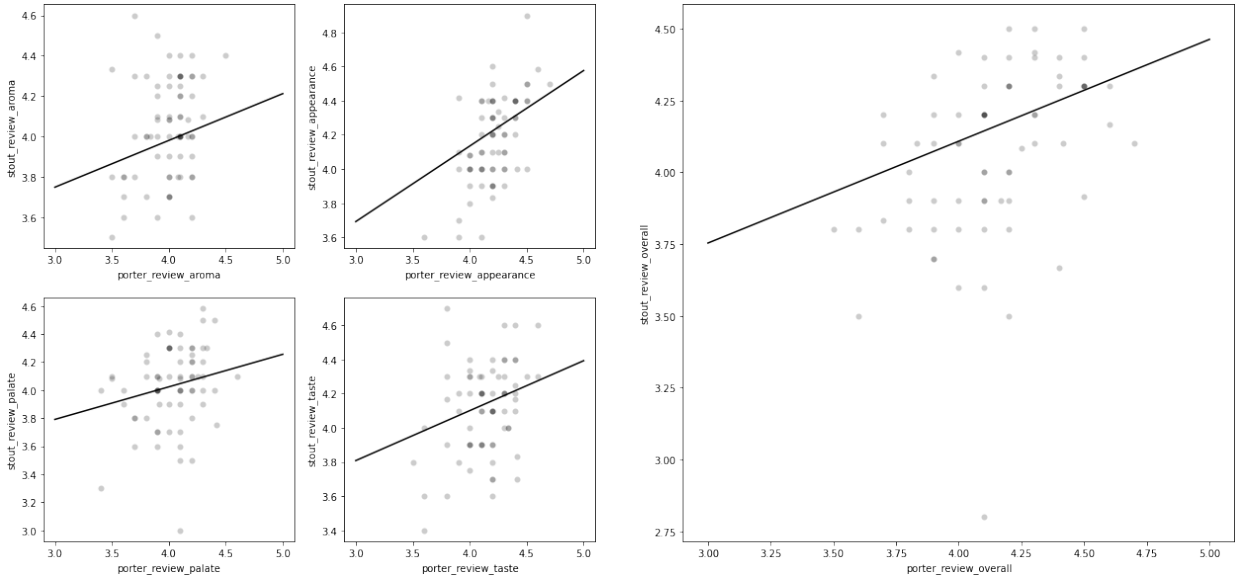

Figura 9: Relación entre variables Porter y Stout

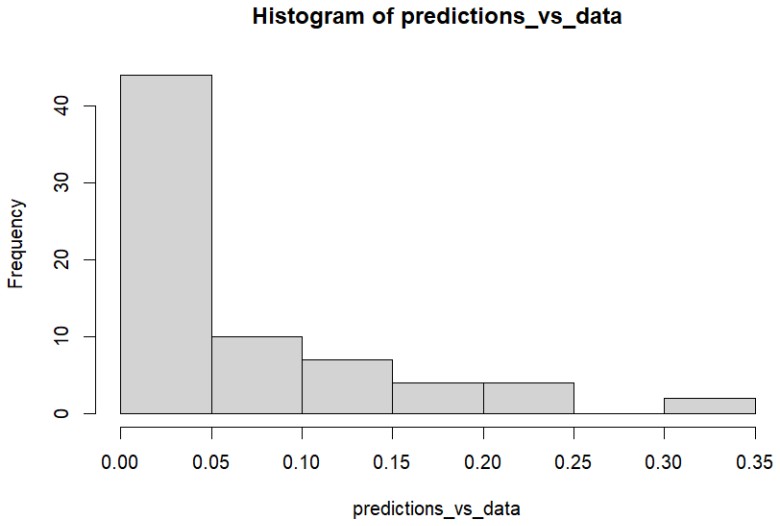

Figura 10: Error Cuadrático (Predicciones VS Datos)

Implementación, Resultados y Comentarios Finales Para este resultado se considero los datos aportados por 72 usuarios, sobre los cuales calculamos su percepción de la categoría como la media de las 5 cervezas por grupo, sobre estas medias, realizamos las estimaciones puntuales de la percepción porter sobre la stout, sobre estas predicciónes calculamos la predicción general con las cadenas de markov que ajustamos considerando todos los datos, sobre estos calculamos la media de los valores obtenidos y calculamos el error cuadrático medio con los datos sobre la recepción, esto son explorados en la figura 10.

Como comentario final, el resultado si bien es prometedor, hay muchos aspectos que podrían mejorar este modelo, mismos de lo que me di cuenta desarrollando esta idea:

- **Cuantificación de la incertidumbre:** Se consideraron dos modelos, el primero es una predicción general sobre las variables perceptivas, el modelo demostro tener poca incetidumbre, esto lo sabemos por los intervalos de confianza de nuestras predicciones, sin embargo el problema radica en el segundo modelo, una aproximación puntal no es suficiente para capturar la variabilidad entre la relación de un estilo con otra.
- **El resultado es demasiado optimista**, esto por la forma del entrenamiento esto se planteo asi como un primer acercamiento, sin embargo el error de predicción se hace sobre el mismo conjunto de entrenamiento, lo cual puede causar un overfitting.
- **Precaución con la interpretación de relación**, queda claro que el que dos variables aletorias se distribuyan similarmente, no necesariamente implica que esten relacionadas, el uso de las variables perceptivas y el basarse en una guia es una respuesta a este acermaiento. en caso de explorar esté metodo sobre otros estilos, es necesario tener certeza que existe una relación perceptiva.

Bibliografía

[1] E. Hallmark, "Beers, breweries, and beer reviews ("https://www.kaggle.com/ehallmar/beers-breweries-and-beer-reviews"),"
2018.

[2] B. J. C. Program, "Bjcp guidelines 2015 (https://www.bjcp.org/bjcp-style-guidelines/),"
2015.