OCEAN.5q Predictive Test

Juan Eloy Suarez - PH125 Data Science - Harvard

11/June/2021

1 Introduction

Executive Summary

In this project, ... * Prestigio del test BigFive/OCEAN * Valor de los resultados aproximados VS "pesadez" de 50 preguntas = interés en lograr resultados aprox(*) con sólo 5 preguntas* (*): defino una función de Accuracy HighLow por trait acertando >=3 de 5 preguntas

OBJETIVOS: * 1) minimizar test * 2) hacerlo móvil * 3) preparar para vincular con otros datos; 4) medir y mantener alta accuracy

ORIGINALIDAD: * Adaptar el modelo de Recomendación (aunque la matriz no tenga sparsity) * Relleno de IBCF (si lo hago) * Versión "disclosed" para móvil * 1) usar recommender; 2) servir de "módulo" para futuras integraciones

PREMISAS: * no usar test para parámetros (uso sub-partición); 2) computabilidad local; 3) respeto estricto al test

2 Understanding the data

Exploration of the ...

- explicar prescriptores descartados (tiempos de carga, ubicación, etc)
- criterio país
- criterios año y mes
- correlaciones
- gráficos de ratings

3 Methods and train

3.1 Strategy, process

1) simplify prescriptors
2) use questions as items for recommendation
3) calculate OCEAN results
4) measure accuracy
5) optimize questions to show
6) pre-load free fields
7) publish shiny
8) store results

-PARTITIONS -FUNCIÓN ACCURACY (deciles) -INICIAL: comprobar si “reverse questions” afectan al recommender (estudiar impacto en: 1)recommenderPrincipal; 2)recommenderShiny; 3)cálculoOCEANinicial; 3) cálculoOCEANshiny) || fórmula: abs(6-x) -TRAIN: criterio selección preguntas (random? cor?) -TRAIN: criterio # preguntas (5?) -TRAIN: criterio algoritmo a usar (UIBF?) -TRAIN: tamaño óptimo dataset

3.2 The shiny application

-SHINY: registrar usuario, país y timestamp + vbles1:3 -SHINY: almacenar resultados -SHINY: añadir traducción -SHINY: crear un “submit” para queda el resultado sea “inamovible” (o explicar lo dejó así “interactivo” y mostrando las predicciones de preguntas con fines académicos +- “disclosed mode”)

4 Results

We have got aa appretiable result just with Phase 1 Model, and in case we have event time data available for cases to predict, we will obtain even more improvement on RMSE. As a results sumamry, running the final algorithm of our modelling on the –validation– set yields the following rating:

5 Conclusion

We have implemented a model that ...

Finally, we can mention some potential improvements for future versions: -Optimizar tuning Recommender (evaluationScheme, k-fold, bootstrap, no-sparse,) usando sub-particiones -Probar aumentando el tamaño del dataset (usando sub-participes de training) -Ensemble de mejor algoritmo/trait -Negociar con ADE para crear modelo de rendimiento académico “5q” -Mantener random de preguntas (para diversificar datos obtenidos) pero mejorar selección de otras preguntas por «correlación ó CART -Posibilidad de >5 preguntas -Bias Month, Country (mostrar en EDA) -Mejoras SHINY: multialgoritmo_ensemble, botónSubmit, añadirPreguntasControl, almacenarResultados_ojoGDPR, ... -Vincular las 5 preguntas a otros juegos de preguntas/prescriptores para obtener in modelo único con outcome OCEAN

As a final conclusion, we can state we have reached challenge goal in terms of project requirements and accuracy for the validation set provided.