APRENDIZAJE DE MÁQUINA - PARCIAL 3

Fecha: noviembre 15 de 2024. Duración: 3 horas

REGLAMENTO DE ESTUDIANTES de la PONTIFICIA UNIVERSIDAD JAVERIANA

Faltas disciplinarias graves y gravísimas:

Artículo 123 inciso (d): El fraude en actividades, trabajos y evaluaciones académicos y la posesión o utilización de material no autorizado en los mismos. (Falta disciplinaria grave)

Artículo 124 inciso (b): Todas las modalidades de plagio. (Falta disciplinaria gravísima)

Artículo 124 inciso (e): La suplantación en una evaluación académica, en exámenes preparatorios, en trabajos de grado y tesis. (Falta disciplinaria gravísima)

Sanciones Disciplinarias:

Artículo 127: Las faltas graves serán sancionadas con amonestación escrita con cargo a la hoja de vida del estudiante y la imposición de matrícula condicional durante el tiempo necesario para cumplir la condición.

Artículo 128: Adicional a la sanción disciplinaria, el fraude en actividades, trabajos y evaluaciones académicos se sancionará académicamente con la pérdida de la asignatura, la cual será calificada con nota definitiva de cero punto cero (0.0).

NO SE PERMITE INTERCAMBIO DE INFORMACIÓN CON PERSONA ALGUNA DURANTE EL DESARROLLO DEL PARCIAL. SOLO PUEDE INTERACTUAR CON EL PROFESOR. PARA COMUNICARSE CON EL PROFESOR, ENVÍELE UN MENSAJE PRIVADO POR TEAMS.

IMPORTANTE

- Debe generar 2 Notebooks: uno para el PUNTO 1 y otro para el PUNTO 2, responda las preguntas planteadas dentro de una celda de texto en el notebook. El PUNTO 3 debe ser contestado dentro de este documento Word. Nombre cada archivo así: "**Punto X – nombre del estudiante**". Antes de las 10:10 am, debe cargar los 3 archivos en el buzón "Parcial 3" configurado en la plataforma de BS. Si un archivo se carga entre las 10:10 am y las 10:40 am, se califica sobre 4.0. Después de las 10:40 am no se recibirán parciales. Se recomienda cargar cada archivo apenas lo termine.
- Para el entrenamiento y prueba de los modelos, **solo puede utilizar los comandos vistos en clase**.
- Para todos los comandos que involucren una selección aleatoria, por ejemplo, dividir el dataset entre entrenamiento y prueba, o seleccionar registros de un dataset, fije siempre la semilla con los **últimos 4 dígitos de su cédula** de ciudadanía.
- Para evaluar el desempeño de sus modelos, utilice**: ACCURACY, RECALL y PRECISION**.
- Cuando se pida regularización, **NO aplique Ridge, Lasso o Elastic Nets**. Simplemente haga un cambio en los parámetros del modelo que conlleve a que el modelo disminuya su ACCURACY en entrenamiento, pero que la aumente en prueba

DATASET

Se tiene una base de datos (**Datos_hospital**) con la información disponible sobre 3698 pacientes de un hospital en la India. La información está compuesta por las siguientes variables las cuales están relacionadas con la estadía del paciente en el hospital, y con su percepción sobre los servicios del hospital, incluido el NET PROMOTER SCORE (NPS), el cual será nuestra variable objetivo. Las columnas de este dataset son:

1. **SN**: Número de identificación del paciente
2. **AgeYrs**: Edad del paciente
3. **Estimatedcost:** Costo total de la estadía del paciente en el hospital.
4. **LengthofStay:** Número de días que el paciente permaneció en el hospital.
5. **InsPayorcategory:** Categoría del paciente para efectos del pago de los servicios: 1: CORPORATIVO, 2: EXEMPTO, 3: CON POLIZA DE SEGURO, 4: INTERNACIONAL, 5: REGULAR.
6. **CE_VALUEFORMONEY:** Opinión del paciente sobre la pregunta: *¿Ha recibido una buena relación calidad-precio?* Valor numérico entero entre 1 y 4, donde 1 es el más bajo y 4 el más alto.
7. **CE_CSAT:** Opinión del paciente sobre la pregunta: En general, *¿Está usted satisfecho con los servicios que ha recibido?* Valor numérico entero entre 1 y 4, donde 1 es el más bajo y 4 el más alto.

8. **AE_PATIENTSTATUSINFO:** Opinión del paciente sobre "*Orientación e información sobre el estado de salud del paciente*". Valor numérico entero entre 1 y 4, donde 1 es el más bajo y 4 el más alto.
9. **CE_ACCESSIBILITY:** Opinión del paciente sobre la pregunta: *¿Nos encontró cuando nos necesitaba?* Valor numérico entero entre 1 y 4, donde 1 es el más bajo y 4 el más alto.
10. **CE_NPS:** Opinión del paciente sobre la pregunta: *¿Qué tan probable es que nos recomiende a un amigo o familiar?* Valor numérico entero entre 0 y 10, donde 0 es que no es nada probable que lo recomiende, y 10 es que es extremadamente probable que si lo recomiende.
11. **NPS_Status:** Valor de CE_NPS clasificado en 2 categorías: **DETRACTOR** (CE_NPS <= 8) y **PROMOTOR** (CE_NPS >=9).

La base de datos original tiene muchas más variables, aunque la variable objetivo es la misma. El dataset está ligeramente desbalanceado, pero no se preocupe por este tema. No balancee el dataset.

Para desarrollar los modelos solicitados, Usted deberá tomar un sub-conjunto de entrenamiento del 60% de los datos, y un sub-conjunto de prueba con el 40% restante. **Para el desarrollo de los modelos, utilice las variables apropiadas. Elimine las variables que no necesite**

PUNTO 1 (24). RANDOM FOREST

Desarrolle un modelo de RANDOM FOREST, para detectar DETRACTORES/PROMOTORES utilizando los parámetros por defecto del comando randomForest de R. Llame este modelo el "modelo RF base". **(6)**

1. Determine si es mejor utilizar un umbral (recuerde que el dataset está desbalanceado) para clasificar un individuo en DETRACTOR/PROMOTOR con base en la probabilidad generada por el modelo (pruebe un valor bajo y un valor alto del umbral), o utilizar la clasificación directa entregada por el modelo (umbral = 0.5). Tenga en cuenta está decisión para responder los items 2 y 3 de este punto. **(6)**

   Para calcular probabilidades con randomForest en R puede usar el siguiente código:

```
# Train a random forest model
rf_model <- randomForest(Species ~ ., data = train_data)
# Predict probabilities on test data
probabilities <- predict(rf_model, newdata = test_data, type = "prob")
head(probabilities)
```

2. ¿Cómo se podría aplicar una regularización en el modelo RF base? Si es posible, aplíquelo y demuestre que el modelo fue regularizado. Si no es posible, explique brevemente porqué. Recuerde que regularización es cualquier cambio en los parámetros del modelo que disminuya la exactitud en entrenamiento, y que la aumente en prueba. **(6)**
3. Utilizando el modelo RF base, determine la variable que más influye en determinar si alguien es DETRACTOR o PROMOTOR. Teniendo en cuenta esta variable, ¿cuál podría ser una estrategia para aumentar el NPS? **(6)**

APRENDIZAJE DE MÁQUINA - PARCIAL 3

Fecha: noviembre 15 de 2024. Duración: 3 horas

Pontificia Universidad JAVERIANA Colombia

PUNTO 2 (18). BOOSTING

Desarrolle un modelo de BOOSTING para detectar DETRACTORES/ PROMOTORES utilizando alguna de las librerías de BOOSTING en R vistas en clase (ADABOOST, GBM, XGBOOST). Llame este modelo el "modelo GB base". **(6)**

1. ¿Cómo se podría obtener un sobre-ajuste en el modelo GB base? Si es posible, aplíquelo y demuestre que hay un sobre-ajuste. Si no es posible, explique brevemente porqué. **(6)**
2. ¿Cómo se podría obtener un sub-ajuste en el modelo GB base? Si es posible, aplíquelo y demuestre que hay un sub-ajuste. Si no es posible, explique brevemente porqué. **(6)**

PUNTO 3 (8). CONCLUSIÓN

Considerando todos los modelos desarrollados en los puntos 1 y 2, responda las siguientes preguntas (debe proveer una respuesta para cada punto, no se admite una sola respuesta general):

1. ¿Cuál es el mejor modelo para detectar PROMOTORES? Explique brevemente porqué. **(2)**

2. ¿Cuál es el mejor modelo para detectar DETRACTORES? Explique brevemente porqué. **(2)**

3. ¿Cuál es el modelo más confiable al predecir un PROMOTOR? Explique brevemente por qué. **(2)**

4. ¿Cuál es el modelo más confiable al predecir un DETRACTOR? Explique brevemente por qué. **(2)**