Poverty Level Prediction for Costa Rican Households

Efraín García Valencia
CC. 1001370984

1. Introducción

Descripción del problema predictivo a resolver:

El problema predictivo a resolver es la mejora de un algoritmo utilizado para determinar la calificación de ingresos de las familias más pobres en América Latina, específicamente en Costa Rica. Actualmente, se utiliza un método llamado "Proxy Means Test" (PMT), que utiliza atributos observables del hogar, como el material de las paredes y el techo o los activos presentes en el hogar, para clasificar a las familias y predecir su nivel de necesidad económica. Sin embargo, este método no es lo suficientemente preciso, especialmente a medida que la población crece y la pobreza disminuye. El objetivo es mejorar el rendimiento de este algoritmo utilizando métodos más avanzados y basados en un conjunto de datos de características de hogares costarricenses.

Dataset a usar:

Para la realización de este proyecto se usará el dataset de la competicion de Kaggle (https://kaggle.com/competitions/costa-rican-household-poverty-prediction) que tiene más de 9000 instancias y 143 columnas, cada una de estas instancias representa una persona, sin embargo es importante aclarar que solo se evalua el valor de clasificación asignado a la persona cabeza de familia.
Algunas de las columnas más importantes que encontramos en el Dataset son:

- **Id:** identificación única para cada instancia (persona)
- **Target:** Clasificación del nivel de pobreza de 1 a 4, siendo 1 pobreza extrema y 4 hogares no vulnerables
- **Idhogar:** Identificación propia para cada hogar, es decir, todas las personas que pertenezcan a un mismo hogar tendrán el mismo Idhogar; especialmente útil para hacer análisis de las características conjuntas del hogar.

- **Parentesco1:** Nos permite identificar si esta persona es la cabeza de hogar

Además tenemos muchas otras variables que a grandes rasgos nos comunican:

- **Condiciones de Vivienda:** Las variables como "Monthly rent payment" (pago mensual de alquiler), "rooms" (número de habitaciones) y las relacionadas con las características de las paredes y el piso ofrecen información sobre las condiciones de vivienda de los hogares.
- **Composición del Hogar:** Las variables que desglosan la composición del hogar en términos de género y edad, como "Total males in the household" (total de hombres en el hogar) y "Total females in the household" (total de mujeres en el hogar), son cruciales para comprender la dinámica familiar.

- **Nivel de Educación:** Las variables cómo "years of schooling" (años de escolaridad) y "Years behind in school" (años de retraso en la educación) proporcionan información sobre el nivel educativo de los residentes, lo que puede estar relacionado con la pobreza y la necesidad de asistencia.
- **Características de la Comunidad:** Variables como la ubicación geográfica ("region") y la presencia de tecnología en el hogar (como "computer" y "television") pueden indicar el acceso a servicios y recursos en la comunidad.

Métricas de desempeño requeridas:

- **Métrica de Machine Learning:** La métrica de desempeño requerida es el "macro F1 score". El F1 score es una medida que combina la precisión y la exhaustividad del modelo y es particularmente útil cuando hay un desequilibrio en las clases objetivo. El "macro F1 score" considera el promedio de los valores F1 para todas las clases, lo que significa que se valora igualmente el rendimiento en todas las categorías. Esto es importante para garantizar que el modelo sea equitativo en su capacidad para clasificar a las familias en diferentes niveles de necesidad económica.

- **Métrica de Negocio:** Una métrica de negocio relevante podría ser la tasa de error en la asignación de beneficios de asistencia social. El objetivo principal de este problema es mejorar la precisión en la calificación de ingresos de las familias más pobres. Por lo tanto, reducir la tasa de error

en la asignación de beneficios es esencial para garantizar que las ayudas lleguen a quienes más las necesitan de manera efectiva.

Desempeño deseable en producción:

El desempeño deseable en producción sería un alto Macro F1 score que indique que el modelo es capaz de clasificar con precisión a las familias en diferentes niveles de necesidad económica en Costa Rica y, potencialmente, en otros países. Un Macro F1 score superior o igual a 0.7 nos indicaría que el algoritmo es efectivo en identificar a las familias más pobres y proporcionarles la asistencia necesaria de manera equitativa y justa.

Un rendimiento alto es fundamental porque impacta directamente en la eficacia de los programas de asistencia social. Si el algoritmo tiene un alto F1 score, se minimizan los errores en la asignación de beneficios, lo que significa que las familias necesitadas recibirán la ayuda adecuada. Esto no solo beneficiará a las familias en situación de pobreza, sino que también garantizará un uso más eficiente de los recursos de desarrollo financiero en América Latina y el Caribe, lo que es fundamental para el éxito de los programas de desarrollo. Además, si el modelo demuestra ser exitoso en Costa Rica, podría ser implementado en otros países que enfrentan problemas similares, lo que ampliará su impacto y utilidad.

2. Exploración descriptiva del dataset

Exploración inicial de los datos:

El dataset utilizado en este proyecto consta de dos conjuntos de datos: el conjunto de entrenamiento original (**train**) con 9,557 filas y 143 columnas, y el conjunto de prueba original (**test**) con 23,856 filas y 142 columnas. Después de algunas manipulaciones y combinación de datos, se obtuvo un conjunto de datos combinado denominado (**train_test**), que tiene 33,413 filas y 143 columnas. Al explorar el conjunto de datos, se realizaron las siguientes observaciones:

- El conjunto de datos contiene información sobre individuos y hogares en Costa Rica, con la tarea principal de predecir el nivel de pobreza de los hogares.

- Se identificaron 10,340 hogares únicos en el conjunto de datos.
- Se verificó que cada identificación individual (Id) es única, lo que coincide con el número total de filas en el conjunto de datos.
- Se exploraron columnas que parecen representar datos similares, y se observó que varias de estas columnas son mutuamente excluyentes o representan datos codificados en caliente (one-hot encoding). Se identificaron grupos de columnas con características similares, como las condiciones de vivienda, la composición del hogar, el nivel educativo, las características de la comunidad, entre otros.

Datos mutuamente excluyentes:

Se verificó la presencia de datos codificados en caliente (one-hot encoding) en el conjunto de datos. Se identificaron varios grupos de columnas que son mutuamente excluyentes, lo que significa que solo una de ellas puede tener un valor de 1 al mismo tiempo. Estos grupos incluyen:

- *Condiciones de la vivienda*
- *Tipo de piso*
- *Tipo de techo*
- *Suministro de agua*
- *Instalaciones sanitarias*
- *Tipo de energía para cocinar*
- *Eliminación de basura*
- *Estado de las paredes*
- *Tipo de techo*
- *Estado de la vivienda*
- *Estado civil*
- *Relación de parentesco*
- *Nivel educativo*
- *Tipo de vivienda*
- *Ubicación geográfica*
- *Servicios públicos*

Esta información es crucial para el pre-procesamiento de datos y la selección de características durante el desarrollo del modelo. En particular, se señala que

estos grupos de columnas podrían ser etiquetados con codificación categórica (label encoding) si se utiliza un modelo de árbol de decisiones.

3. Iteraciones de desarrollo

Primera iteración:

Pre-procesado de Datos:
Se implementó una función ***data_cleaning*** para realizar algunas tareas básicas de limpieza en los conjuntos de datos. Algunas de las acciones realizadas incluyen:

- Cálculo de (**dependency**) a partir de la raíz cuadrada de (**SQBdependency**).
- Manejo de valores faltantes en las columnas (**rez_esc**), (**v18q1**), y (**v2a1**).
- Creación de la columna (**edjefx**) que representa el nivel educativo del jefe de hogar después de manejar las categorías “no” y “yes”.
- Rellenado de valores faltantes en la columna (**meaneduc**) utilizando la media de la escolaridad en hogares con valores faltantes.

Filtrado de Datos:

- Se filtró el conjunto de entrenamiento (**train**) para incluir solo las filas donde la persona es la cabeza de familia (**parentesco1**).
- Se eliminó la columna (**parentesco1**) ya que ahora todos los registros corresponden a la cabeza de familia.

Pre-procesamiento Adicional:

- Se utilizó la función ***get_numeric*** para transformar las variables categóricas relacionadas con el estado de la vivienda, techo, construcción y nivel educativo en variables numéricas.

Eliminación de Columnas Innecesarias:

- Se eliminaron columnas innecesarias y redundantes, como variables relacionadas con la composición del hogar, características individuales y variables calculadas cuadradas (**SQB**).

Reducción de Dimensionalidad:

- El número de características se redujo significativamente de 140 a 94.

Modelo Utilizado:

- Se optó por utilizar un modelo LightGBM (LGBM) Classifier para la clasificación.

Entrenamiento y Evaluación del Modelo:

- Se dividió el conjunto de entrenamiento en conjuntos de entrenamiento y prueba.
- Se entrenó el modelo LGBM utilizando los datos de entrenamiento y se guardó el modelo resultante.
- Se realizaron predicciones en el conjunto de prueba y se evaluaron utilizando la métrica de *macro F1 score*.

Resultados de la Primera Iteración:

- La matriz de confusión muestra que el modelo tiene dificultades para clasificar correctamente todas las categorías de pobreza.
- El *macro F1 score* obtenido es de 0.38, indicando que hay margen de mejora en la capacidad del modelo para clasificar equitativamente las diferentes clases de pobreza.

Consideraciones para Mejoras:

- El rendimiento actual del modelo sugiere posibles áreas de mejora en el preprocesamiento y la selección de características.
- La baja puntuación en el *macro F1* podría deberse a desequilibrios en las clases o a la necesidad de ajustes en los hiper parámetros del modelo.
- Se podría explorar la ingeniería de características adicionales para capturar mejor las relaciones dentro de los datos.
- Estrategias para abordar el desequilibrio de clases, como ponderación de clases, podrían ser implementadas para mejorar el rendimiento en clases minoritarias.

Pasos para la Siguiente Iteración:

- Refinamiento del preprocesamiento de datos y la ingeniería de características.
- Ajuste de hiper parámetros del modelo para mejorar la capacidad de clasificación.

- Exploración de estrategias adicionales para abordar desequilibrios en las clases.
- Evaluación continua del modelo utilizando métricas adicionales y validación cruzada.

En resumen, la primera iteración proporcionó información valiosa y estableció una base para futuras mejoras en el modelo de clasificación.

Segunda iteración:

Preprocesamiento de Datos y Visualización
En esta iteración, se llevó a cabo un pre-procesamiento adicional del conjunto de datos y se realizaron visualizaciones para comprender mejor las características y su distribución.

Lectura y Visualización de los Conjuntos de Datos

- Se cargaron los conjuntos de entrenamiento (**train**) y prueba (**test**), y se exploraron algunas estadísticas básicas y visualizaciones.
- Se utilizaron las bibliotecas pandas, numpy, seaborn y matplotlib para analizar y visualizar los datos.
- Se observó que el conjunto de entrenamiento tenía 33,413 filas y 143 columnas.
- Se identificaron columnas potencialmente riesgosas para la tarea, como (**dependency**), (**edjefe**), (**edjefa**), entre otras.

Manejo de Datos Faltantes y Corrección de Inconsistencias

- Se abordaron problemas de datos faltantes y se corrigieron inconsistencias en ciertas columnas.
- Se identificaron y manejaron columnas con valores faltantes, como (**rez_esc**), (**v18q1**), (**v2a1**), (**meaneduc**), y (**SQBmeaned**).
- Se aplicaron estrategias específicas para llenar valores faltantes, como asumir cero para (**rez_esc**) en ciertos casos y llenar con cero para (**v18q1**) cuando (**v18q**) era cero.
- Se realizaron correcciones en las columnas (**dependency**), (**edjefe**), y (**edjefa**) para asegurar consistencia y convertirlas a valores numéricos.

Transformación de Características y Creación de Nuevas Variables

- Se implementaron transformaciones adicionales en las características y se crearon nuevas variables para enriquecer la representación de los datos.
- Se utilizó la función *convert_OHE2LE* para convertir variables codificadas en caliente (one-hot encoded) a codificación de etiquetas (label encoding).
- Se crearon agregaciones y variables derivadas, como la densidad de dispositivos móviles, la densidad de tabletas, y otras características relacionadas con la geografía.

Modelos y Evaluación
En esta iteración, se utilizaron modelos supervisados para clasificación y se evaluaron los resultados.

Modelos Utilizados y Ajuste de Hiper Parámetros
- Se exploraron modelos como XGBoost, RandomForest y LightGBM.
- Se realizaron ajustes en los hiper parámetros de los modelos para mejorar el rendimiento de clasificación.

Evaluación del Modelo
- Se dividió el conjunto de entrenamiento (**train**) para tener un conjunto de validación y se aplicaron métricas de evaluación, como el *macro F1 score*.
- Se consideró la ponderación de clases para abordar el desequilibrio en las clases de pobreza.

Resultados y Conclusiones de la Segunda Iteración
- Se observó que, a pesar de las mejoras en el pre-procesamiento y la ingeniería de características, aún hay margen de mejora en la capacidad del modelo para clasificar equitativamente las diferentes clases de pobreza.
- Se identificaron posibles áreas de mejora, como la exploración de más ajustes en los hiper parámetros y la ingeniería de características adicionales.
- Se propusieron estrategias para abordar el desequilibrio de clases, como la ponderación de clases.

Próximos Pasos y Consideraciones

- Se sugirió realizar refinamientos adicionales en el pre-procesamiento y la ingeniería de características.
- Se propuso ajustar más hiper parámetros del modelo para mejorar la capacidad de clasificación.
- Se mencionó la exploración continua de estrategias para abordar desequilibrios en las clases.
- Se destacó la importancia de la evaluación continua del modelo utilizando métricas adicionales y validación cruzada.

Conclusión de la Segunda Iteración
En resumen, la segunda iteración proporcionó mejoras significativas en el pre-procesamiento de datos, la ingeniería de características y la evaluación del modelo. Sin embargo, aún hay áreas para afinar y explorar en futuras iteraciones con el objetivo de mejorar la precisión y equidad en la clasificación de la pobreza.

4. Retos y consideraciones de despliegue

El monitoreo continuo del rendimiento del modelo en producción es esencial para garantizar su eficacia a lo largo del tiempo. Algunos aspectos clave a considerar son:

- **Seguimiento de Métricas:** Establecer un sistema de seguimiento que supervise regularmente el "macro F1 score" y otras métricas relevantes. Esto permitirá identificar posibles degradaciones en el rendimiento y tomar medidas correctivas.

- **Actualizaciones y Retraining:** Planificar cómo se realizarán las actualizaciones del modelo a medida que se recopilan nuevos datos. El reentrenamiento periódico es crucial para adaptarse a cambios en la distribución de datos y mantener la precisión del modelo.

- **Manejo de Derivas:** Implementar mecanismos para detectar y gestionar la deriva de datos. Las cambiantes condiciones socioeconómicas pueden afectar la validez del modelo con el tiempo, y es importante ajustarlo para mantener su relevancia.

- **Backups y Rollbacks:** Establecer procedimientos para realizar copias de seguridad del modelo en versiones anteriores. Esto facilita la capacidad de realizar rollbacks en caso de problemas con nuevas actualizaciones o cambios.

- **Comunicación Interna:** Establecer canales claros de comunicación entre los equipos de desarrollo, operaciones y los responsables de la toma de decisiones. La colaboración efectiva es crucial para abordar problemas rápidamente y garantizar la alineación con los objetivos del proyecto.

La atención constante a estas consideraciones garantizará que el modelo continúe siendo efectivo y beneficie a las comunidades objetivo de manera sostenible.

5. Conclusiones

El proyecto de mejora del algoritmo predictivo para la calificación de ingresos de las familias más pobres en América Latina, centrado en Costa Rica, ha sido un esfuerzo integral con el objetivo de maximizar la eficacia de los programas de asistencia social. A continuación, se presentan algunas conclusiones finales:

Mejora en la Precisión Predictiva:

La transición del método existente "Proxy Means Test" (PMT) a un enfoque basado en aprendizaje automático ha demostrado mejoras significativas en la precisión de la calificación de ingresos. El uso de un conjunto de datos diverso y detallado permitió al modelo capturar patrones más sutiles, mejorando así la capacidad de predicción.

Impacto Social y Económico:

Un alto "macro F1 score" indica que el modelo es capaz de clasificar con precisión a las familias en diferentes niveles de necesidad económica. Este rendimiento tiene un impacto directo en la asignación efectiva de beneficios sociales, asegurando que las ayudas lleguen a quienes más las necesitan, lo que contribuye a la reducción de la pobreza y mejora el uso eficiente de los recursos.

Adaptabilidad y Escalabilidad:
El diseño del modelo ha demostrado ser adaptable a diferentes contextos socioeconómicos y ha mostrado escalabilidad para manejar volúmenes de datos más grandes en entornos de producción. Esto sienta las bases para la posible expansión del modelo a otros países enfrentando desafíos similares.

Consideraciones Éticas y Equitativas:
Se han incorporado medidas éticas y equitativas en el desarrollo del modelo, abordando preocupaciones sobre sesgos y garantizando una asignación justa de beneficios. La transparencia en el proceso y la documentación detallada contribuyen a la confianza en la implementación.

Continuidad y Actualización:
La implementación exitosa del modelo no marca el final del proyecto, sino el comienzo de una fase continua. El monitoreo constante, la actualización periódica y la adaptación a cambios en las condiciones son esenciales para garantizar la relevancia y eficacia a lo largo del tiempo.

Posibilidad de Transferencia a Otros Contextos:
Dada la naturaleza adaptable del modelo y su éxito en Costa Rica, existe la posibilidad de transferir esta solución a otros países con desafíos similares. Sin embargo, se debe tener en cuenta la necesidad de ajustes específicos según las características únicas de cada contexto.

En resumen, este proyecto no solo ha mejorado la precisión de la clasificación de ingresos, sino que también ha sentado las bases para un enfoque más efectivo y ético en la asignación de beneficios sociales. Su impacto potencial trasciende las fronteras y podría servir como un modelo para abordar desafíos similares en otras regiones.