Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Facultad de Ingenieria
Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Análisis de la Probabilidad y Factores de Evasión Estudiantil basado en el Historial Académico y datos de Perfil Educacional y Socioeconómico.

12 de noviembre de 2023

Maicon Araújo Fialho
Tomas Elsesser

Índice general

1. Resumen

La evasión de estudiantes en cursos de educación superior representa un desafío significativo tanto para las instituciones educativas como para los propios estudiantes. Identificar los factores que contribuyen a la evasión es crucial para desarrollar estrategias de intervención efectivas. En este estudio, el objetivo es evaluar la probabilidad de evasión de estudiantes en un curso, centrándose en dos conjuntos principales de variables: el historial de calificaciones en todas las materias cursadas y las respuestas a un cuestionario socioeconómico completado por cada estudiante. Comprender la relación entre el rendimiento académico y los factores socioeconómicos es fundamental para ofrecer un apoyo más personalizado a los estudiantes y, así, reducir las tasas de evasión.

2. Problema General

Esta investigación tiene como objetivo analizar la evasión estudiantil en cursos de educación superior, tomando en cuenta el impacto de las variables de rendimiento académico y factores socioeconómicos durante el periodo de 2013 a 2015. Utilizaremos datos recopilados a través de un cuestionario socioeconómico y registros de historiales académicos de estudiantes. Para abordar este problema, aplicaremos un modelo de Regresión Logística Generalizada (GLM), permitiéndonos evaluar la probabilidad de evasión estudiantil. La investigación se centrará en la relación entre las calificaciones en todas las materias cursadas y las respuestas al cuestionario, explorando cómo estas variables influyen en la probabilidad de evasión estudiantil. Además, buscaremos identificar patrones y tendencias en la evasión, considerando la interacción entre estas variables.

3. Objetivos de la Investigación

- Investigar la relación entre el historial de calificaciones académicas de los estudiantes y la probabilidad de deserción en el curso.
- Evaluar cómo los factores socioeconómicos, obtenidos a través de un cuestionario completado por los estudiantes, afectan la probabilidad de deserción.
- Desarrollar un modelo estadístico basado en la regresión logística generalizada (GLM) para predecir la probabilidad de deserción de los estudiantes basándose en sus historiales académicos e información socioeconómica.
- Utilizar el modelo para identificar los principales determinantes de la deserción y proporcionar información útil para la implementación de estrategias de prevención.

4. Investigaciones Previas

El abandono estudiantil ha sido objeto de numerosas investigaciones. Dada la complejidad de los factores involucrados, desde desafíos académicos hasta cuestiones socioeconómicas, estas investigaciones se han multiplicado en un intento por comprender este fenómeno. La necesidad de abordar el abandono estudiantil eficazmente hace que la revisión y síntesis de estas investigaciones sea esencial para establecer estrategias efectivas de retención. A continuación se listan trabajos e investigaciones que abordan esta temática:

1. **Student Dropout in Technical-Professional Higher Education: Exploring factors that influence freshmen students.**
   https://www.scielo.org.mx/scielo.php?pid=S0185-27602018000400109&script=sci_abstract&tlng=en

2. **Factors Influencing Dropout Students in Higher Education**
   https://www.hindawi.com/journals/edri/2023/7704142/#related-articles

3. **University student retention: Best time and data to identify undergraduate students at risk of dropout**
   https://www.tandfonline.com/doi/citedby/10.1080/14703297.2018.1502090?scroll=top&needAccess=true

5. Metodologia

5.1. Dataset

Estructura: El dataset cuenta con 285757 observaciones y 42 variables de estudiantes de una universidad brasileña que completaron o abandonaron el curso. cuenta con 4327 alumnos que finalizaron el curso y 1179 que abandonaron, así como sus calificaciones en todas las materias cursadas durante el periodo que permanecieron en el curso.

5.2. preprocesamiento

En el preprocesamiento, trabajaremos con estudiantes que hayan completado o abandonado sus cursos, cuya fecha de ingreso al curso fue de 2013 a 2015 y cuya forma de ingreso sea SISU[1].

Las siguientes variables se convertirán en factores, mientras que las demás mantendrán su tipo original, ya sea texto o numérico. La variable de matrícula del estudiante se cifrará para preservar el anonimato de los estudiantes:

[1]SISU: Sistema de Selección Unificada, un sistema brasileño que utiliza las notas del Enem.[2]
[2]Enem: Examen Nacional de Educación Media, una evaluación nacional brasileña utilizada como criterio de selección para el ingreso a instituciones de educación superior.

Las siguientes variables se convertirán en factores, mientras que las demás mantendrán su tipo original, ya sea texto o numérico. La variable de matrícula del estudiante se cifrará para preservar el anonimato de los estudiantes:

- Curso
- SituacaoAluno
- FormaIngresso
- IfesCodigo
- GrupoCotaConvocado
- TipoDisciplina
- EscolaEnsinoFundamental
- TurnoEnsinoFundamental
- EscolaEnsinoMedio
- TurnoEnsinoMedio
- QuandoConcluiuEnsinoMedio
- ParticipouEvento
- IngressouEnsinoSuperior
- QuantosCursosSuperiores
- Cor
- CausaCondicoesDiferenciadas
- CapacidadeEnxergar
- CapacidadeLocomover

Posteriormente, se realizarán análisis para verificar valores faltantes y valores atípicos que puedan influir en el resultado final del análisis. Luego, decidiremos si eliminamos o no los datos faltantes.

5.3. Variables

Las variables incluidas son: Variables originadas del historial escolar:

- MatriculaC: Identificador único de la matrícula del estudiante.
- Curso: Nombre del curso.
- AnoAdmissao(tipo numérica): Año de admisión.

- SemAdmissao(tipo numérica): Semestre de admisión.
- AnoConclusao(tipo numérica): Año de conclusión.
- SemConclusao(tipo numérica): Semestre de conclusión.
- SituacaoAluno(tipo factor): Situación del estudiante.
    - Posibles Valores: Conclusão; Abandono
- FormaIngresso(tipo factor): Forma de ingreso del estudiante.
    - Posibles Valores: SiSU - Sistema de Seleção Unificada; Mudança de Curso; Vestibular; Transferência; Portador de Diploma; Processo Seletivo Alternativo - SiSU; Processo Seletivo Alternativo - Vestibular; Sequência; Convênio - PEC/G; Programa Especial
- IfesCodigo(tipo factor): Código de la institución.
    - Posibles Valores: SISU2013; VRUFV13-1; UFV2013; SISU2014; UFV2014; UFV2015; SISU2015; LEC2014; LEC2015
- GrupoCotaConvocado(tipo factor): Grupo de cuotas convocado.
    - Posibles Valores: Ampla concorrência; Estudante de escola pública, Renda mensal menor ou igual a 1,5 salários mínimos(por pessoa da família); Estudante de escola pública, Renda mensal maior que 1,5 salários mínimos(por pessoa da família), Etnia: Autodeclarado Preto, Pardo ou Índio; Estudante de escola pública, Renda mensal maior que 1,5 salários mínimos(por pessoa da família), Outro; Estudante de escola pública, Renda mensal menor ou igual a 1,5 salários mínimos(por pessoa da família), Outro; Estudante de escola pública, Renda familiar mensal menor ou igual a 1,5 salários mínimos, Etnia (Autodeclarado Preto, Pardo ou Índio); Estudante de escola pública, Renda familiar mensal maior que 1,5 salários mínimos, Etnia (Autodeclarado Preto, Pardo ou Índio); Não informado ou não se aplica; Estudante de escola pública, Renda familiar mensal menor ou igual a 1,5 salários mínimos, Outro; Estudante de escola pública, Renda familiar mensal maior que 1,5 salários mínimos, Outro; Docentes que atuam ou já atuaram em escolas do campo; Trabalhadores do campo ; Sujeitos com vínculos aos movimentos sociais do campo; Egressos de escolas do campo; Educadores populares ou monitores vinculados à educação do campo ; Índios e quilombolas ; Demanda social
- CRA(tipo numérica): Coeficiente de Rendimiento Acumulado en el curso.
- CienciasNatureza(tipo numérica): Nota en Ciencias de la Naturaleza en el Enem.
- CienciasHumanas(tipo numérica): Nota en Ciencias Humanas en el Enem.

- LinguagensCodigos(tipo numérica): Nota en Lenguajes y Códigos en el Enem.
- Matematica(tipo numérica): Nota en Matemáticas en el Enem.
- Redacao(tipo numérica): Nota en Redacción en el Enem.
- ENEM(tipo numérica): Nota Média el Enem.
- NotaIngressoUFV(tipo numérica): Nota de ingreso a la UFV.
- CodDisciplina: Código de la disciplina.
- TipoDisciplina(tipo factor): Tipo de la disciplina.
    - Posibles Valores: Obrigatória; Optativa; Facultativa; Reconhecimento de Optativa
- Disciplina: Nombre de la disciplina.
- AnoNota(tipo numérica): Año de la nota.
- SemestreNota(tipo numérica): Semestre de la nota.
- Conceito: Concepto de la disciplina.
- Nota(tipo numérica): Nota de la disciplina.
- FaltasPratica(tipo numérica): Faltas en la parte práctica.
- FaltasTeorica(tipo numérica): Faltas en la parte teórica.
- ExameSuficiencia(tipo binário): Realizó del Examen de Suficiencia.
- ExameComplementar(tipo binário): Realizó del Examen Complementario.

Variables originadas de un cuestionario que el estudiante llena al ingresar en la facultad:

- EscolaEnsinoFundamental(tipo factor): Escuela de Educación Primaria.
    - Posibles Valores: Todo em escola particular; Todo em escola pública; A maior parte em escola pública; A maior parte em escola particular; Todo em Curso Supletivo; A maior parte em Curso Supletivo
- TurnoEnsinoFundamental(tipo factor): Turno de la Educación Primaria.
    - Posibles Valores: Todo em horário diurno; A maior parte em horário diurno; A maior parte em horário noturno; Todo em horário noturno
- EscolaEnsinoMedio(tipo factor): Escuela de Educación Secundaria.

    - Posibles Valores: Todo em escola particular; Todo em escola pública; A maior parte em escola pública; A maior parte em escola particular; Todo em Curso Supletivo; A maior parte em Curso Supletivo

- TurnoEnsinoMedio(tipo factor): Turno de la Educación Secundaria.
    - Posibles Valores: Todo em horário diurno; A maior parte em horário diurno ;A maior parte em horário noturno; Todo em horário noturno

- QuandoConcluiuEnsinoMedio(tipo factor): Cuando completó la Educación Secundaria.
    - Posibles Valores: Nos últimos 5 anos; Nos últimos 15 años; No último año; Há mais de 20 anos; Nos últimos 20 anos

- ParticipouEvento(tipo binario): Participó en el evento "La graduación en la UFV".
    - Posibles Valores: Sim; Não

- CursoPreparatorio(tipo factor): Asistió a un Curso Preparatorio para el Ingreso a la Educación Superior.
    - Posibles Valores: Sim; Não

- IngressouEnsinoSuperior(tipo factor): Después de cuántos intentos ingresó la educación superior.
    - Posibles Valores: Na segunda tentativa; Após a terceira tentativa; Na primeira tentativa; Na terceira tentativa

- QuantosCursosSuperiores(tipo factor): Cantidad de cursos superiores ya cursó.
    - Posibles Valores: Está cursando o seu primeiro Curso Superior; Está cursando outro Curso Superior; Já iniciou, mas abandonou; Não iniciou nenhum, até o momento; Já concluiu um Curso Superior

- Cor(tipo factor): Raza o color.
    - Posibles Valores: Pardo(a); Branco(a); Preto(a); Amarelo(a); Indígena

- CausaCondicoesDiferenciadas(tipo factor): Razón para condiciones de aprendizaje diferenciadas.
    - Posibles Valores: Nenhuma diferenciação; Distúrbio de Atenção e Hiperatividade (TDAH); Autismo (Síndrome de Asperger); Distúrbios de leitura (dislexia), ou na escrita (disgrafia, disortografia), ou na matemática (discalculia); Alta capacidade de raciocício/desempenho; Problemas crônicos de saúde física (renal, cardíaca, etc); Deficiência de visão, audição ou física; Mais de uma condição presente entre as mencionadas acima

- CapacidadeEnxergar(tipo factor): Capacidad para ver.
    - Posibles Valores: Sem dificuldade visual; Com resíduo visual; Sem visão
- CapacidadeLocomover(tipo factor): Capacidad para moverse.
    - Posibles Valores: Sem dificuldade de caminhar; Com pouca dificuldade de caminhar; Sem condição de caminhar; Com grande dificuldade de caminhar

5.4. Modelo Estadístico

La investigación utilizará un modelo estadístico basado en la regresión logística generalizada (GLM). Este modelo permitirá relacionar las variables independientes (calificaciones académicas y factores socioeconómicos) con la variable dependiente (probabilidad de deserción). El uso de la regresión logística generalizada es apropiado para modelar problemas de clasificación binaria y calcular la probabilidad de un evento, en este caso, la probabilidad de deserción.

5.5. Validación del Modelo

Se implementará una estrategia de validación cruzada para evaluar el rendimiento del modelo, garantizando así una evaluación robusta y una generalización adecuada. La división entre conjuntos de entrenamiento y prueba se realizará de manera cuidadosa para asegurar la robustez del modelo ante nuevos datos.

5.6. Métricas de Evaluación

La evaluación del desempeño de nuestro modelo de Regresión Logística Generalizada (GLM) se centrará en métricas fundamentales, específicamente la Curva ROC (Receiver Operating Characteristic) y el área bajo la curva (AUC). Estas métricas son esenciales para comprender la capacidad del modelo de distinguir entre las clases positivas y negativas.

5.7. Desbalance de Clases

Dada la posible desigualdad en la distribución de clases, se aplicarán técnicas para abordar el desbalance, tales como el ajuste de pesos de clase o el empleo de técnicas específicas diseñadas para manejar conjuntos de datos desequilibrados.

6. Comentarios Adicionales

6.1. Privacidad de los Estudiantes

Para salvaguardar la privacidad de los estudiantes, se seguirán estrictos protocolos éticos en el manejo de datos sensibles. Se emplearán técnicas de anonimización y

cifrado para proteger la confidencialidad de la información, asegurando así el respeto de la privacidad de los involucrados en la investigación.

Bibliografía

Ignacio, G. C. F. (s.f.). *Student dropout in technical-professional higher education: Exploring factors that influence freshmen students.* Descargado de https://www.scielo.org.mx/scielo.php?pid=S0185-27602018000400109&script=sci_abstract&tlng=en

Long, N., A., Z., y Noor, M. F. M. (2023, February 8). Factors influencing dropout students in higher education. *Education Research International*, *2023*, 7704142. Descargado de https://www.hindawi.com/journals/edri/2023/7704142/ doi: 10.1155/2023/7704142