Título del Proyecto	**Proyecto Final de Minería de Datos: *Walmart Trip Type Classification***
Equipo	René Rosado González 137085 — Arantza Ivonne Pineda Sandoval 141194 Ana Luisa Masetto Herrera 183203 — Ixchel Meza Chávez 172860
Profesor	Juan Salvador Mármol Yahya
Fecha de entrega	21 de diciembre de 2018

ANTECEDENTES Información general sobre la Empresa

Actualmente, la competencia existente en la industria de los supermercados es feroz gracias a nuevos actores que facilitan la obtención de información más detallada y profunda que provee ventaja competitiva a una empresa sobre otros competidores. De estos nuevos actores hay dos que han mostrado ser de gran utilidad en los últimos años: Análisis de Datos y Aprendizaje de Máquina, estos dos conceptos de ser aplicados propiamente pueden posicionar a una empresa muy por encima de las demás a pesar de que estas ya estén fuertemente consolidadas.

Es por eso que Walmart desde su comienzo fijó su cultura empresarial enfocada al entendimiento del comportamiento de sus clientes diseñando una experiencia de compra personalizada para que el cliente se sienta cómodo en sus instalaciones y por lo tanto quieran regresar; sin embargo, en los años más recientes Walmart se ha dado cuenta que para poder seguir como la gran e importante empresa que es, es necesario adaptar su cultura empresarial incorporando los actores previamente mencionados y así poder modernizar y enriquecer su sistema de segmentación personalizada. .

PROYECTO Comprensión del Problema y Plan de Trabajo

El proyecto seleccionado es Walmart Recruiting: Trip Type Classification, un proyecto presentado en la plataforma de Kaggle cuyo objetivo es realizar un análisis de la canasta de mercado para clasificar viajes de compras al supermercado. Con base en este problema a resolver, se presenta la propuesta del equipo de trabajo para generar modelos de manera comparativa que permitan tanto ajustar a los datos observados como predecir el comportamiento de nuevos datos. Para el desarrollo, se proporcionaron dos conjuntos de datos: el primero, un conjunto de entrenamiento (train.csv) que está organizado por 6 variables explicativas diferentes referentes al comportamiento de los clientes y sus compras en el supermercado (*ScanCount, VisitNumber, Weekday, Upc, DepartmentDescription y FinelineNumber*) y una variable que se quiere predecir *(TripType)* que hace referencia a 38 tipos de viajes identificados que realizan los compradores; el segundo, un conjunto de prueba (test.csv) con el cual se probará la eficiencia de los modelos propuestos.

El plan de trabajo consiste en cumplir una serie de hitos fijados por el equipo de trabajo conforme a las fechas y actividades indicados en el Anexo 1.

<u>Relevancia en la industria</u>

Este proyecto es de gran relevancia no solo por la robustez de los modelos que aquí se generan para el caso de Walmart, el gran gigante de los supermercados; sino que además, un modelo con todas las sutilezas y supuestos con el que fue construido puede ser implementado en muchas industrias BtoC, siendo el retail su punto fuerte. El elemento más impactante de este proyecto es la capacidad de predecir con alta precisión el tipo de viaje por cliente y de ofrecer una herramienta de web service que permita replicar el análisis para datos futuros. De este modo, los resultados contribuirán a incrementar el valor empresarial y obtener ventaja competitiva sobre la competencia de la industria al segmentar a sus clientes a través de los datos de consumo ya existentes.

OBJETIVOS Objetivos generales y específicos

Objetivo General

El propósito general de este proyecto es, siguiendo con la filosofía de Walmart y dándole al cliente un lugar especial dentro de su filosofía empresarial, adquirir información relevante proveniente de las transacciones y otras variables para poder clasificar a los clientes dependiendo de sus tipos de viaje a las instalaciones de la empresa y así mejorar en diversos aspectos.

Objetivos Específicos

Para poder estructurar un proyecto de minería de datos que permita alcanzar el nivel deseado en la competencia de Kaggle, se tienen que efectuar adecuadamente las siguientes acciones (objetivos):

- Descargar y leer los datos correspondientes a la base de datos a analizar.
- Limpiar la base de datos para que respete un formato tidy y no entorpecer el manejo de las variables y observaciones.
- Realizar un Análisis exploratorio de los Datos: Univariado,Bivariado y Multivariado, con el fin, de tener una visión general de la
  situación.
- Estructurar diversos modelos de aprendizaje de máquina que permitan clasificar a los visitantes en los tipos de viaje.
- Evaluar los modelos propuestos, mejorar y seleccionar el mejor que permita lograr el nivel deseado de la pérdida multiclase.
- Participar en la competencia de Kaggle y obtener el nivel deseado de la pérdida multiclase.

METODOLOGÍA Desarrollo de proyecto siguiendo CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM), es un modelo analítico que permite tener la base para desarrollar propiamente un proyecto de minería de datos. En este proyecto se sigue este modelo y a continuación se presentan algunas de las tareas que se llevaron a cabo en cada una de sus partes y los puntos más importantes que se obtuvieron como resultados de esta.

Gracias a esta parte del CRISP-DM se comprendió en su totalidad la filosofía empresarial de Walmart que consta en darle prioridad al cliente y para lograr esto, una estrategia de segmentación es de vital importancia. Además de esto, es en esta sección donde se establecen los objetivos y se establece el criterio de éxito, en este caso, lo que se buscaba era participar en la competencia en Kaggle de Walmart y obtener la menor pérdida multiclase posible.

Comprensión de los datos

Es en esta sección donde se tiene el primer acercamiento con la base de datos a utilizar en el proyecto. La bases de datos que proporcionó Walmart son 2, clasificadas como entrenamiento y prueba, utilizando la primera para hacer el análisis exploratorio de los datos. El análisis exploratorio de los datos arrojó información muy importante como: las correlaciones entre departamentos son muy importantes de analizar porque se puede observar cómo es que se comportan las variables en conjunto; también se pudo observar que existen variables que parecieran no aportar mucha información ya que el comportamiento de estas pareciera ser homogéneo. Para ilustrar mejor los puntos anteriores, hay una gráfica que muestra que el total de elementos comprados depende mucho del departamento del que se hable, también que el comportamiento de los diferentes tipos de viaje de cliente no varía mucho en relación con el día de la semana, es decir, hay clases que simplemente compran más que otras, pero en contraste, el consumo total de los productos si varia mucho en relación con el día de la semana.

También se pudo observar en esta parte los errores de registro en la base de datos, errores como dobles espacios, diagonales y otros caracteres extraños que entorpecen el manejo de la base.

Preparación de los datos

Esta fue de las partes a la que se le dedicó gran parte del tiempo, en esta sección se llevó a cabo la limpieza profunda de los datos y la ingeniería de características y son dos los dos puntos más importantes que se pueden resaltar. El primero es que realizar limpieza de datos no es una tarea fácil, por ejemplo, en la base de datos se podían observar celdas con diversos caracteres extraños, también había categorías en la variable de departamento que no están

correctas; y el segundo es que en el caso de un proyecto cuyo objetivo es hacer una segmentación lo mejor que se puede hacer es utilizar variables con más categorías, ya que no se generaliza y permite al modelo aprender comportamiento de los datos que en otro formato no se pueden apreciar o al menor no con tanta facilidad.

Modelado

En esta sección se proponen modelos que puedan servir para resolver el problema en cuestión, para el caso de este proyecto se propusieron varios modelos de clasificación, sin embargo, los tres que tuvieron mejor desempeño fueron Regresión Logística, GBoost y Bosques Aleatorios. Además de proponer los modelos, estos se codificaron en python. Es de suma importancia mencionar que los datos en esta sección construyen tres bases de datos, ya que esto nos permite entrenar el modelo, validarlo y evaluarlo.

Evaluación

Es en esta sección donde se selecciona el mejor modelo y se ajusta con los parámetros estimados para proceder con la competencia de Kaggle. Los puntos más importantes a considerar en este punto es que poder computacional es necesario para poder ejecutar el mayor número posible de códigos. El mejor rankeado y también el de mayor precisión fue el GBoost.

Implantación

Se realizó un servicio web para la implementación del modelo GBoost seleccionado, utilizando lenguaje python y flask

RECURSOS Requerimientos para Ejecución del Proyecto

Para poder cumplir con los objetivos del proyecto es necesario contar con ciertos recursos que faciliten el desarrollo de éste. Para comenzar, y como recurso más importante a considerar, se requiere de un equipo de personal con conocimientos variados en áreas como: programación y estadística, en el caso de este proyecto el recurso humano consta de un equipo de 4 personas que se van a encargar de llevar a cabo cada una de las actividades enlistadas previamente. El siguiente recurso necesario es poseer las herramientas computacionales que permitan procesar los datos y hacer el análisis necesario, para esto se requiere de un programa llamado R; similar a lo anterior, se requiere un programa que permita hacer el modelado más eficiente, por lo tanto, el programa Python también va a ser utilizado. Para finalizar, es requerida una cuenta de Azure, una herramienta computacional que permite crear, administrar e implementar aplicaciones en la nube.

RESULTADOS Resumen de productos obtenidos

Se obtuvo un modelo GBoost con 65.88 de precisión.

El análisis realizado sugiere que se cuenta con suficiente información para determinar patrones de consumo y correlación entre varios productos. Sin embargo, para atender la problemática planteada se sugiere recopilar datos que generen una mejor personificación de los consumidores como individuos. Por ejemplo, recopilar la hora de entrada y salida, el método de pago, y el número de personas que le acompañan.

EVALUACIÓN Contraste con el Criterio de Éxito

Modelo	Precisión	Kaggle Score
GBOOST	65.88	1.1616
Logistic Reg	61.38	1.2937
RandomForest	42.48	17.10

CONCLUSIONES Comparación de modelos

VENTAJAS	LIMITACIONES
La **regresión logística** puede ser interpretable si así se desea.	Su capacidad predictiva es menor al Gboost.
El **Gboost** tiene una mejor capacidad predictiva.	La interpretación de los procesos resulta compleja.
Ambos modelos son perfectibles.	Se requeriría re-evaluar la data para obtener características adicionales.

Anexo 1

no.	Actividades	Subtareas	Periodo	
1	Definir equipos	-	01-dic-18	01-dic-18
2	Definir base de datos a utilizar	Indagar en las reglas de las competencias de cada una de las posibles bases de datos Descargar las bases de datos Definir el objetivo de cada base de datos	01-dic	07-dic
3	Estudiar la base de datos a detalle	Determinar los errores de registro para limpiar la base de datos Determinar si se cuenta con el poder de computo necesario para procesar los datos	07-dic	08-dic
		CRISP-DM		
4	Comprender el Negocio	Determinar los antecedentes de Walmart, es decir, el ambiente en el que se desarrolla y comprender su funcionamiento y cultura. Establecer el objetivo general y los objetivos específicos Determinar el criterio de éxito del proyecto (análisis de la competencia de kaggle para determinar medida de comparación) Establecer el plan del proyecto a detalle. Generar un documento en formato Rmarkdown con los elementos estudiados	08-dic	10-dic
5	Comprender los datos	Lectura de datos Evaluar con detalle los aspectos que deben limpiarse de la base de datos Generar un reporte reproduble que pueda ejecutar la lectura de datos y reporte los aspectos a limpia y considerar valores faltantes.	10-dic	11-dic
5	Preparar los datos	Seleccionar e integrar los datos Realizar la limpieza de datos (Quitar símbolos que entorpezcan el manejo de la base de datos, imputar datos, etc.) Realizar ingeniería de características Generar archivos reproducibles que faciliten la limpieza de datos y la ingeniería de características	11-dic	15-dic
6	Analizar los datos	Realizar el análisis univariado de los datos Realizar el análisis bivariado de los datos Realizar el análisis multivariado de los datos	15-dic	16-dic
7	Modelar en python	Proponer modelos a utilizar para cumplir con los objetivos del proyecto Modelar los datos utilizando python Correr modelos con datos de entrenamiento y prueba Comprar desempeño de modelos y ajustar	16-dic	19-dic
8	Evaluar el modelo	Seleccional el mejor modelo Comparar su desempeño en la competencia de Kaggle Mostrar lugar obtenido en la competencia	20-dic	20-dic
9	Generar un reporte final	Reportar todo lo enlistado con anterioridad Obtener conclusiones del proyecto	20-dic	20-dic
10	Preparar entregables	Elaborar una presentación que dure aproximadamente 15 minutos Acomodar archivos entregables Entregar todo lo necesario para cumplir con el objetivo del proyecto	20-dic	20-dic