Fecha	25/09/2015
Universidad	Tecnológico de Monterrey, Campus Santa Fe
Proyecto (Semana i)	"Business Analytics Microsoft"
Mentora	Profesora Lourdez Muñoz
Estudiantes	Gilberto Silva – Ing. en Tecnologías Computacionales Eric Zuchovicki - Ing. en Tecnologías Computacionales Alejandro Pineda – Ing. en Negocios y Tecnologías de la Información Damian Scarinci – Lic. en Administración de Empresas Alonso Alfaro – Ing. Industrial y de Sistemas

Reto: "Business Analytics Microsoft"
Reporte Ejecutivo

ÍNDICE

TABLA DE CONTENIDOS

1.- ANTECEDENTES

Como parte del nuevo modelo educativo del Tecnológico de Monterrey, la semana i representa un periodo en el que los alumnos de dicha institución tienen la oportunidad de acercarse de manera directa a las empresas para desarrollar de manera acelerada un proyecto relevante para su formación y crecimiento tanto personal como profesional.

En esta segunda edición de la semana i, nuestro equipo participa en el reto desarrollado por Microsoft. El equipo está formado por estudiantes de distintas carreras y con distintos enfoques con el fin de enriquecer el desarrollo del proyecto y alcanzar resultados óptimos.

Antes de comenzar el proyecto, los alumnos tuvieron la oportunidad de asistir a una serie de talleres impartidos por personal de Microsoft en los que se explicaron las herramientas tecnológicas necesarias para el desarrollo del proyecto, así como también se explicó de manera clara y puntual el reto a realizar, incluyendo posibles vertientes, enfoques y análisis del problema.

2. – OBJETIVO DEL PROYECTO

Realizar un proyecto multidisciplinario que permita responder a una pregunta real de negocios relacionada con la campaña llamada "Upgrade Your World México", la cual se encuentra vigente en México y otros países del mundo, por medio de la comprensión y aplicación de los conceptos de analítica de negocios al desarrollo de una solución basada en tecnología de la plataforma de datos Microsoft y que abarque de forma integral las etapas del ciclo de vida de ciencia de datos: entendimiento de negocio, entendimiento de datos, preparación de datos, modelado, evaluación y despliegue.

3. – PROBLEMA DE NEGOCIO

¿Cuáles son las organizaciones de la sociedad civil que tienen mayor cantidad de votos en *Twitter* en la campaña Upgrade Your World de México y qué podemos saber acerca de las personas que votaron por ellas o influenciaron en su elección?

4. – RESULTADOS ESPERADOS

Generación, a través de *Azure Machine Learning*, de dos modelos predictivos óptimos de clasificación de *tweets* hacia alguna fundación, tomando en cuenta las características del *tweet* y del usuario. Los modelos tomarán en cuenta las variables independientes del *tweet* que sean representativas y las características del usuario que lo publica. Posteriormente los modelos serán entrenados para poder clasificar futuros *tweets* hacía alguna fundación con cierta probabilidad.

Mediante el uso de la herramienta *Power BI* se hará un análisis de los datos recopilados para explicar y representar visualmente: el comportamiento del perfil de voto

para cada fundación, el comportamiento de los usuarios que participaron en el proyecto, el recuento de votos siguiendo criterios estrictos y el recuento de votos siguiendo criterios más holgados

5.- INTRODUCCIÓN

#UpgradeYourWorld es una iniciativa mediante la cual Microsoft busca apoyar, económica y tecnológicamente, a aquellas organizaciones no gubernamentales de 10 países diferentes que hacen de nuestro mundo un lugar mejor. Actualmente, son 5 las organizaciones que se han visto beneficiadas en México ("Fonabec", "Fundación Cinépolis", "Fundación Michou y Mau I.A.P", "Reforestamos México" y "Un Kilo de Ayuda") y ahora es turno del público en general elegir a las otras 5 fundaciones que recibirán los $50,000 dólares en efectivo y la tecnología correspondiente.

Con el objetivo de ayudar a Microsoft a encontrar las organizaciones que más votos recibieron por parte del público en twitter, así como saber qué llevó a los usuarios a votar por cierta fundación, decidimos realizar dos modelos con las siguientes variables regresoras. Uno incluyendo el comportamiento de los usuarios y uno no.

Modelo 1

Y1	**X1**	**X2**	**X3**	**X4**	**X5...X32**
Fundación	**Recuento de Fav's**	**Recuento de RT's**	**Fecha**	**Plataforma de Envío**	**Palabras Clave**

Modelo 2

Y1	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**	**X8...X35**
Fundación	**# Fav's**	**# RT's**	**Fecha**	**Plataforma**	**Seguidores del usuario**	**Amigos del usuario**	**Verificado**	**Palabras Clave**

Cabe destacar la suma importancia de las 27 variables binarias denominadas "palabras clave", dado que el modelo toma en cuenta la existencia, o no, de ciertas palabras clave en el tweet (como pobreza, hambre, niños) para poder así asociarlas a una u otra fundación. Esta es una característica muy importante del modelo pues ayuda a clasificar de manera más precisa un tweet a una fundación. Algunas de las palabras clave definidas a través de análisis del texto de una muestra son: niños, cancer, arte, pobreza, educación, seguridad, animal; etc. (El total de las palabras clave pueden ser vistas en el reporte técnico)

También, es importante mencionar que se utilizaron expresiones regulares para poder incluir los derivados de alguna palabra clave en el texto de un tweet. Por ejemplo, el código también detecta discapacitados, discapacidad, discapacidades, discapacitadas; etc.

Las herramientas que se utilizaron para el desarrollo del reto y cumplir con los objetivos o resultados planteados son ***Power Bi, Azure Machine Learning*** *y* ***RStudio***.

6.- METODOLOGÍA EMPLEADA

Para el desarrollo adecuado del reto, formulamos cuatro etapas esenciales a seguir:

Investigación: Lo primero que se tuvo que hacer fue conocer más a fondo sobre la iniciativa de "Upgrade Your World", con el fin de comprender a detalle el problema de negocio, tener bien claro el camino a seguir y definir las tareas que se le iban a asignar a cada integrante del equipo.

Recopilación de datos: Una vez entendido el problema de negocio y planteados los objetivos, comenzamos con la recopilación de datos, es decir, se inició con la extracción de datos (tweets) para generar una base de datos que sirviera como input para entrenar y desarrollar el modelo deseado. Se obtuvo una muestra de 46,266 tweets en un periodo del 13 al 23 de septiembre, dadas las limitaciones de la API de Twitter.

Limpieza de datos: Con los datos registrados en nuestra base, se filtraron los tweets por día y se fueron eliminando/limpiando todos aquellos que contaban con algún error según las variables definidas:

- Se eliminaron las columnas innecesarias de la petición de búsqueda del twitter API
- Se hizo legible la plataforma
- Se consiguió el voto (la fundación)
- Se eliminaron los votos inválidos
- Los votos se pusieron en minúsculas
- Se sustituyeron los acentos por vocales
- Solo se contó un voto de aquellas personas que votaron dos veces en un día

Implementación: Finalmente, con la consolidación de los datos en un solo CSV, se sacaron los reportes utilizando Powe BI y se hizo el entrenamiento de dos modelos en Azure Machine Learning, un modelo se hizo con el dataset sin información del usuario, y el otro se hizo con información del usuario.

NOTA: Para mayor información, cada etapa se encuentra más detallada en el Reporte Técnico para su entendimiento.

7.- PRINCIPALES HALLAZGOS

- Generación de dos modelo predictivos tomando en cuenta a las características seleccionadas. Los modelos permiten, con un 99.75% de precisión (Modelo 1) y un 99.65% (Modelo2), predecir hacia qué fundación irá dirigido un voto.

- Generación de un comparativo del conteo de votos para las principales fundaciones en dos condiciones distintas. La primera, cuando se es estricto al contar los votos y son necesarios los dos "hashtags" escritos correctamente para el conteo del voto. La segunda, cuando se es más holgado y se permiten faltas de ortografía o hashtags similares a los originales.

- Generación de reportes de análisis del comportamiento de los votos participantes en el modelo. Por ejemplo, las mayores plataformas utilizadas para el concurso, la cantidad de participantes, y si éstos votaron una sóla vez o muchas, el top de fundaciones votadas, etc.

8.- CONCLUSIONES Y SIGUIENTES PASOS

Para concluir, los resultados fueron muy gratificantes dado que conseguimos generar dos modelos predictivos con un alto nivel de precisión. Creemos firmemente que estos resultados pueden ser utilizados en diferentes rubros y con resultados exitosos. Por ejemplo, se podría hacer un análisis predictivo de quien va a ganar el concurso a partir de una muestra; tomando en cuenta las plataformas más usadas, las palabras clave más repetidas, las inclinaciones de los tweets de los famosos; entre otras variables incluidas en el modelo. Tambíen, en el ámbito mercadológico, se podrían hacer planes estratégicos para las fundaciones para que éstas supieran las condiciones que más les favorecen y los tipos de usuarios que votan por ellos para que ellos desarrollen un plan de acción y así aumentar su probabilidad de éxito.

En cuanto a los siguientes pasos nos enfocamos en que existen dos vertientes importantes a desarrollar para la continuidad de nuestro proyecto:

1. Desarrollar el despliegue de los modelos. Es decir, desarrollar un dashboard, un formulario web o una aplicación móvil que pueda generar y mostrar de manera más visual y práctica los resultados del modelo. Facilitando así la lectura de los resultados.
2. Desarrollar una serie de estrategias completas y elaboradas para el uso de los resultados obtenidos. Generar planes estratégicos para las fundaciones, predicción de los resultados del concurso, análisis más profundo del comportamiento de los usuarios y de los votos así como las relaciones existentes entre estos; etc.

Creemos que el proyecto tuvo avances bastante importantes y conclusión de comportamientos con resultados interesantes, sin embargo, sabemos también que existen muchos posibles cursos de acción para el progreso del proyecto realizado.