

Exploración de datos y modelo predictivo de canales
Sistema Único de Atención Ciudadana



SECCIONES

● INTRODUCCIÓN

RESUMEN DEL PROYECTO

DESCRIPCIÓN DEL DATASET

PREGUNTAS Y MOTIVACIONES

● EXPLORACIÓN DE DATOS

PRIMERAS VISUALIZACIONES

RANKING DE TICKETS

DETALLE POR CONCEPTO

MAPAS Y ACTION ITEMS

● MODELO PREDICTIVO

PREPROCESSING Y MODELO

BACKGROUND TEÓRICO

DESARROLLO DEL MODELO

RESULTADOS

● CONCLUSIONES Y REFERENCIAS

CONCLUSIONES

REFERENCIAS UTILIZADAS

SECCIONES

● INTRODUCCIÓN

RESUMEN DEL PROYECTO

DESCRIPCIÓN DEL DATASET

PREGUNTAS Y MOTIVACIONES

● EXPLORACIÓN DE DATOS

PRIMERAS VISUALIZACIONES

RANKING DE TICKETS

DETALLE POR CONCEPTO

MAPAS Y ACTION ITEMS

● MODELO PREDICTIVO

PREPROCESSING Y MODELO

BACKGROUND TEÓRICO

DESARROLLO DEL MODELO

RESULTADOS

● CONCLUSIONES Y REFERENCIAS

CONCLUSIONES

REFERENCIAS UTILIZADAS

- Resumen del proyecto

Nuestro proyecto se basó en dos pilares fundamentales: La **exploración de datos** y la creación de un **modelo predictivo** sobre el dataset del Sistema Único de Atención Ciudadana (SUACI), alojado en el sitio de Buenos Aires Data.

El SUACI es el sistema oficial del Gobierno de la Ciudad para que el ciudadano pueda ingresar reclamos, denuncias, quejas y solicitudes provenientes de los diferentes barrios porteños de la ciudad.

A través del proceso de exploración de datos buscamos encontrar patrones o tendencias que nos ayuden a entender mejor la naturaleza de los reclamos y las características principales de los mismos. A su vez, el desarrollo del modelo predictivo nos permitió disponibilizar una nueva herramienta para el Gobierno de la Ciudad que sirve para predecir futuros canales de comunicación, a partir de diferentes características de las solicitudes.

Como visión, a lo largo del proyecto siempre se buscó la generación de accionables claros y la disponibilización de recursos que puedan ayudar a mejorar la calidad de vida de las personas en la Ciudad de Buenos Aires.

Descripción del Dataset

Es una clasificación de los contactos realizados al Sistema Único de Atención Ciudadana (SUACI), con fechas y horarios.

Sistema único de atención ciudadana 2017

Campo	Tipo	Descripción	Ejemplo
CONTACTO	texto	Registro único generado en la base SUACI a partir de la interacción con habitante o visitante (en adelante, vecino) de la Ciudad Autónoma de Buenos Aires.	00007005/17
PERIODO	texto	Mes-año del llamado	201701
CATEGORIA	texto	Agrupamiento de prestaciones y subcategorías de acuerdo a una misma temática	ALUMBRADO
SUBCATEGORIA	texto	Agrupamiento de prestaciones de acuerdo a un mismo objeto	LIMPIEZA DE EQUIPAMIENTO DE ALUMBRADO
CONCEPTO	texto	Concepto que describe con el mayor nivel de detalle al contacto generado por el vecino	LIMPIEZA DE ARTEFACTO DE ALUMBRADO
TIPO_PRESTACION	texto	Clasificación del contacto de acuerdo a la naturaleza específica de la prestación	SOLICITUD
FECHA_INGRESO	Fecha	Día en el que se generó la prestación	2/1/2017
HORA_INGRESO	hora	Hora de ingreso de la prestación	22:16:40
DOMICILIO_CGPC	texto	Comuna de la prestación	COMUNA 7
DOMICILIO_BARRIO	texto	Barrio de la prestación	PARQUE CHACABUCO
DOMICILIO_CALLE	texto	Calle de la prestación	BALBASTRO
DOMICILIO_ALTURA	entero	Altura de la prestación	1700
DOMICILIO_ESQUINA_PROXIMA	texto	Esquina más próxima de la prestación	
LAT	Coordenada	Código de geocodificación Latitud	-34,6393054
LONG	Coordenada	Código de geocodificación Longitud	-58,44193718
CANAL	texto	Medio por el cual se realiza el contacto	WEB
GENERO	texto	Género del vecino que realiza el contacto	MASCULINO
ESTADO_DEL_CONTACTO	texto	Estado de la prestación	CERRADO

Contenido consolidado 2017 + 1er semestre 2018. Shape: 1.5M filas y 16 columnas.

- Preguntas y motivaciones

¿Cuales son las comunas que más generan Tickets?

¿Cuáles son los rubros o tipos de reclamos más cargados?

¿Cómo se distribuyen en la ciudad?

¿Existe algún patrón entre el tipo y categoría de reclamo con la ubicación de donde se carga?

¿Es posible predecir el canal por el cual una persona carga un reclamo, a partir de cierta información?

¿Podemos obtener accionables a partir del análisis de esta información?

¿Cuáles son los horarios picos para determinados reclamos?

¿Podemos ver los tipos de reclamos por ubicación en un mapa de la ciudad?

- Preguntas y motivaciones

- ¿Vamos a poder responder estas preguntas?

- Por supuesto, que sí...
 - ¿Cómo?
 - La respuesta es...



SECCIONES

● INTRODUCCIÓN

RESUMEN DEL PROYECTO

DESCRIPCIÓN DEL DATASET

PREGUNTAS Y MOTIVACIONES

● EXPLORACIÓN DE DATOS

PRIMERAS VISUALIZACIONES

RANKING DE TICKETS

DETALLE POR CONCEPTO

MAPAS Y ACTION ITEMS

● MODELO PREDICTIVO

PREPROCESSING Y MODELO

BACKGROUND TEÓRICO

DESARROLLO DEL MODELO

RESULTADOS

● CONCLUSIONES Y REFERENCIAS

CONCLUSIONES

REFERENCIAS UTILIZADAS

Comunas...



Comuna 1
Retiro, San Nicolás, Puerto Madero, San Telmo, Montserrat y Constitución



Comuna 2
Recoleta



Comuna 3
Balvanera y San Cristóbal



Comuna 4
La Boca, Barracas, Parque Patricios, y Nueva Pompeya



Comuna 5
Almagro y Boedo



Comuna 6
Caballito



Comuna 7
Flores y Parque Chacabuco



Comuna 8
Villa Soldati, Villa Riachuelo y Villa Lugano



Comuna 9
Liniers, Mataderos y Parque Avellaneda



Comuna 10
Villa Real, Monte Castro, Versalles, Floresta, Vélez Sarsfield y Villa Luro



Comuna 11
Villa General Mitre, Villa Devoto, Villa del Parque y Villa Santa Rita



Comuna 12
Coghlan, Saavedra, Villa Urquiza y Villa Pueyrredón



Comuna 13
Núñez, Belgrano y Colegiales

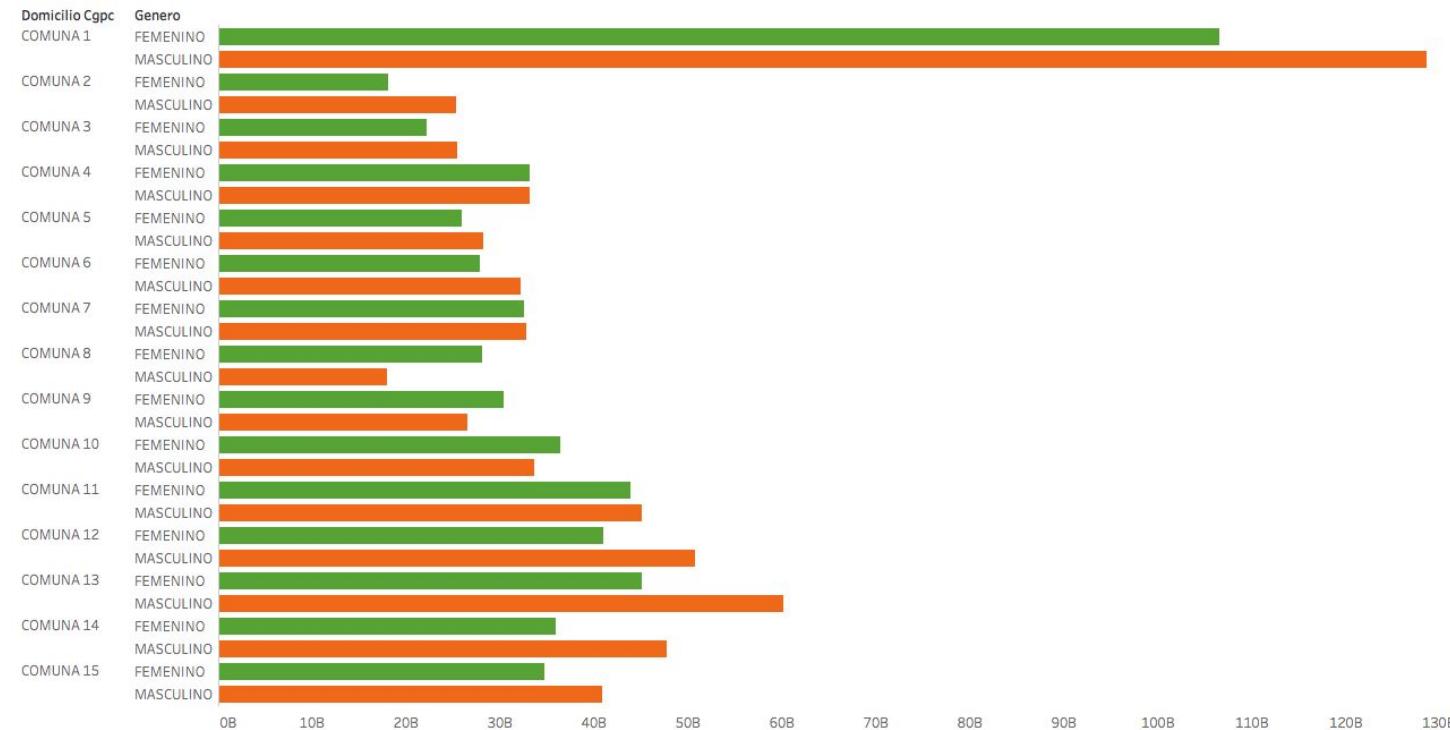


Comuna 14
Palermo

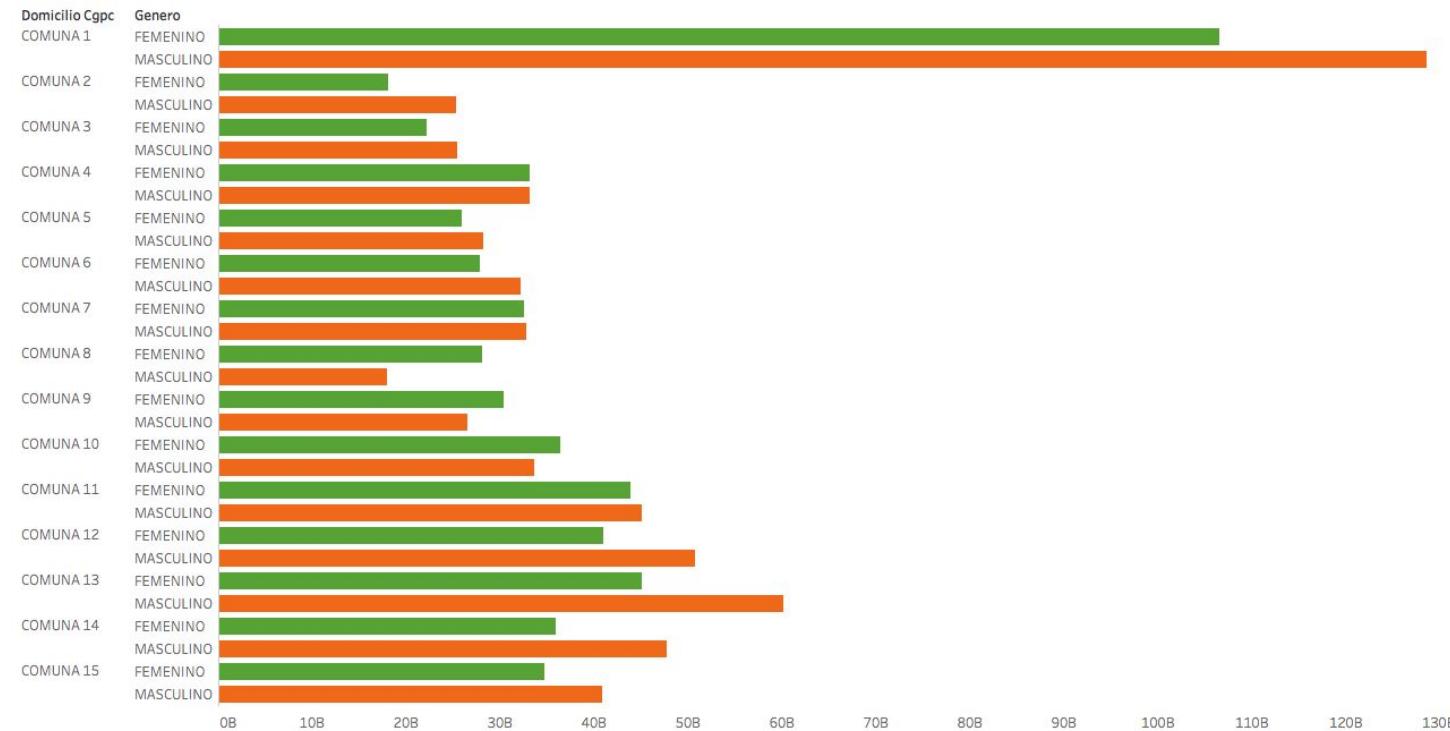


Comuna 15
Chacarita, Villa Crespo, La Paternal, Villa Ortúzar, Agronomía y Parque Chas

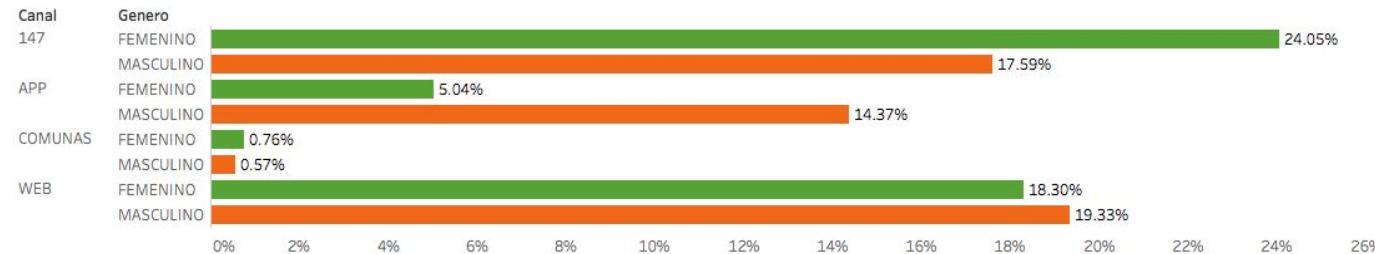
Distribución por comunas y género. La **comuna 1** compuesta por Retiro, San Telmo, Puerto Madero, Montserrat y Constitución presenta el mayor volumen de tickets...



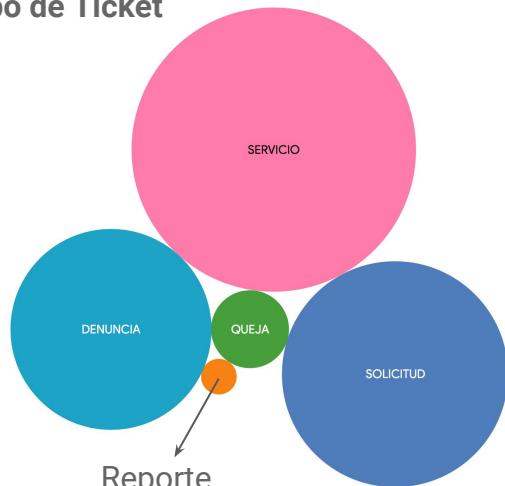
Distribución por comunas y género. La **comuna 1** compuesta por Retiro, San Telmo, Puerto Madero, Montserrat y Constitución presenta el mayor volumen de tickets...



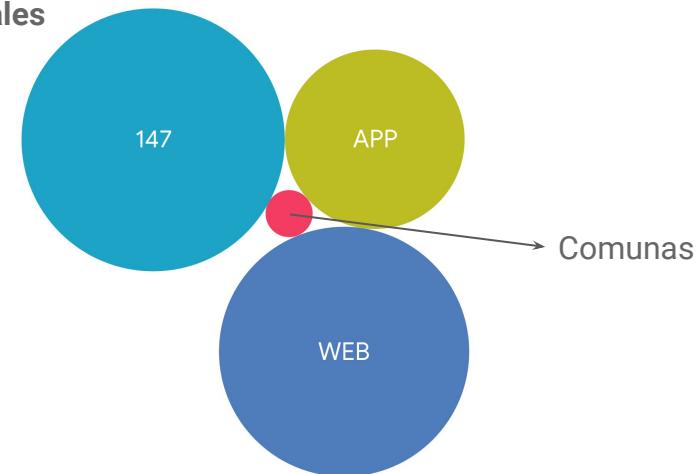
Los dos canales más utilizados en la ciudad son el Web y el 147. Existe una mayor tendencia del público masculino a utilizar la APP (14% vs 5%).



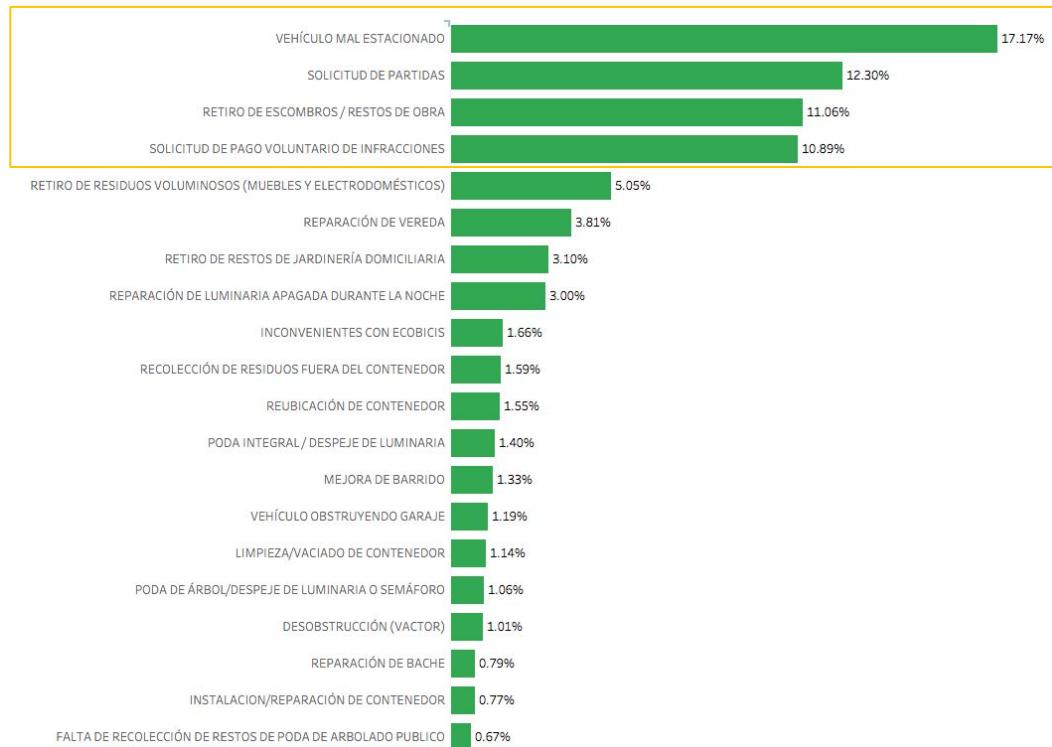
Tipo de Ticket



Canales

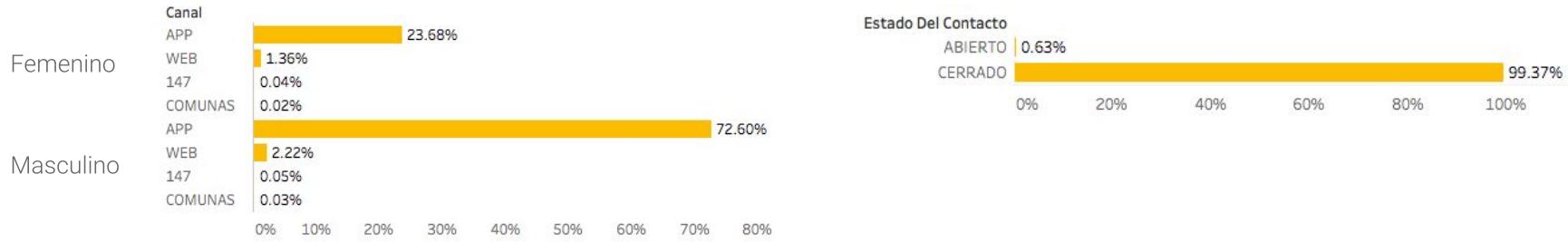


Intensidad de tickets por **Concepto**. El TOP 20 de los motivos de los tickets, componen el 80% del volumen total del último año y medio...



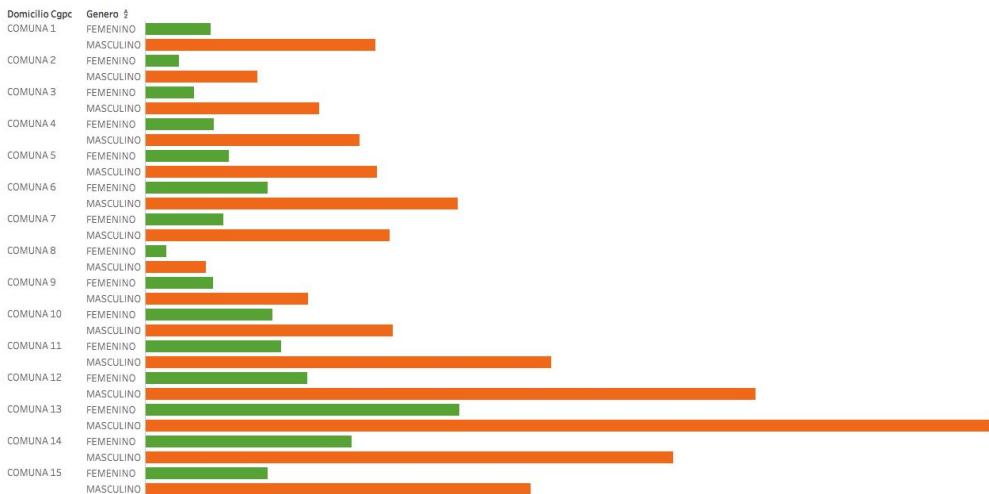
Veamos estas categorías con más detalle...

Vehículo mal estacionado. El pico de tickets se genera a las 11 de la mañana y el canal principal de entrada es vía APP...



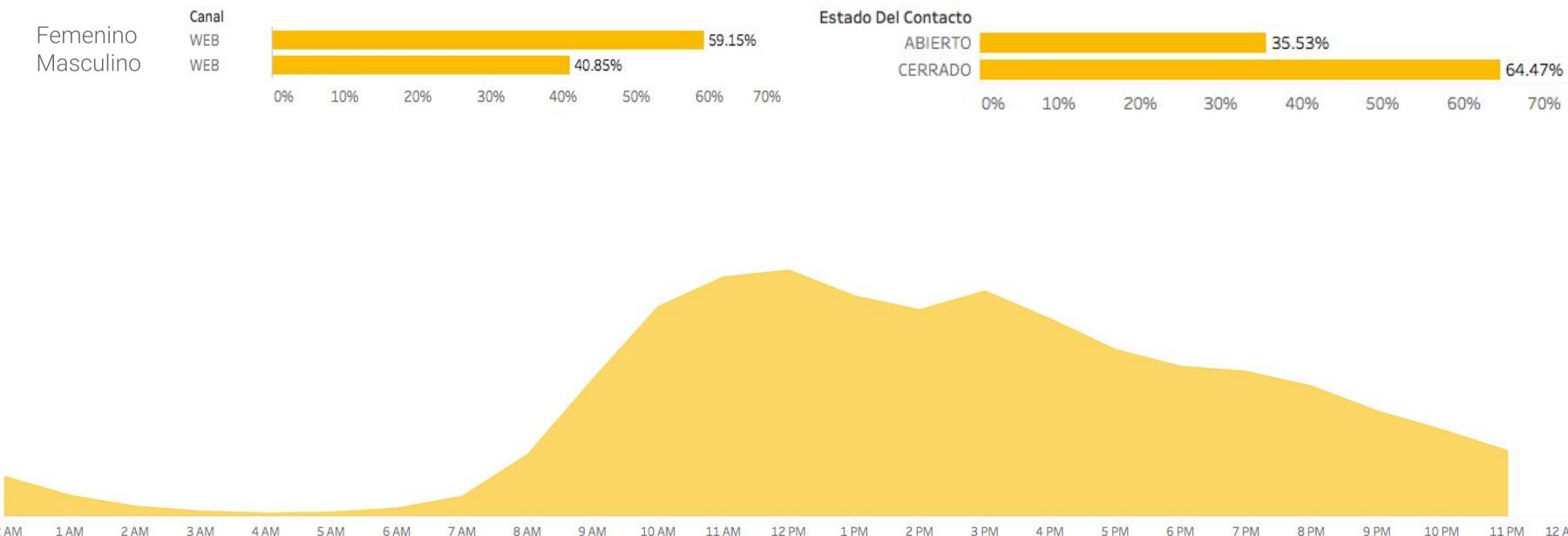


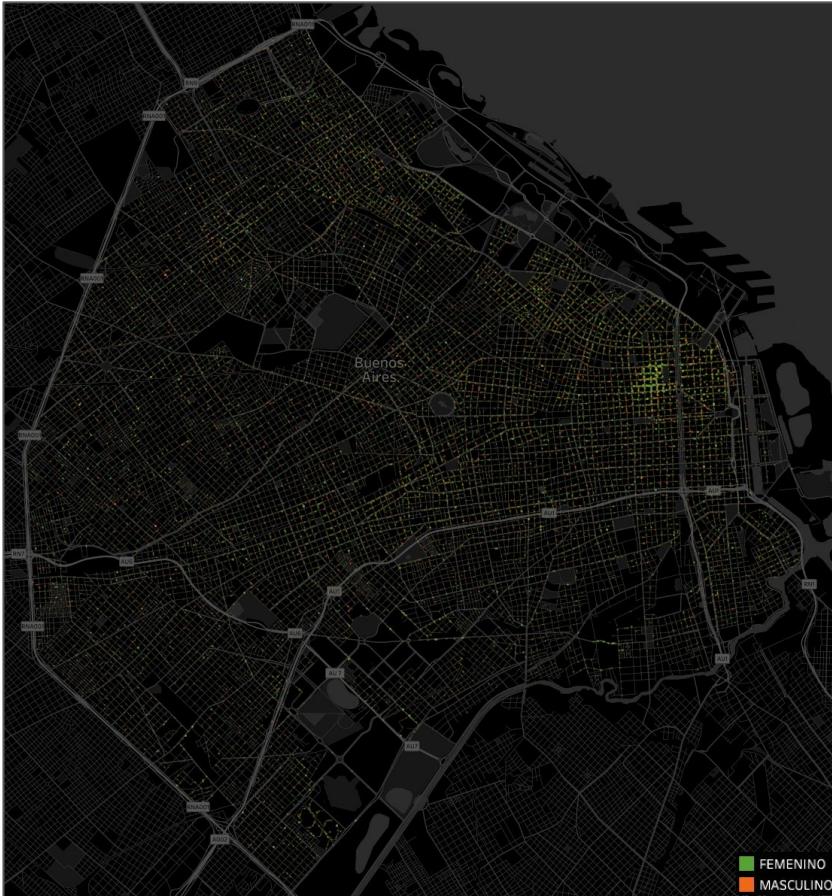
Vehículo mal estacionado. La mayor concentración de Tickets se observa en la zona norte de la ciudad, compuesta por los barrios de Núñez, Belgrano y Colegiales....



★ Action item! Poner estacionamientos en Núñez...

Solicitud de partidas. Se observa que el único canal para este tipo de Ticket es el Web y cuenta con más de un 30% de casos aún abiertos...



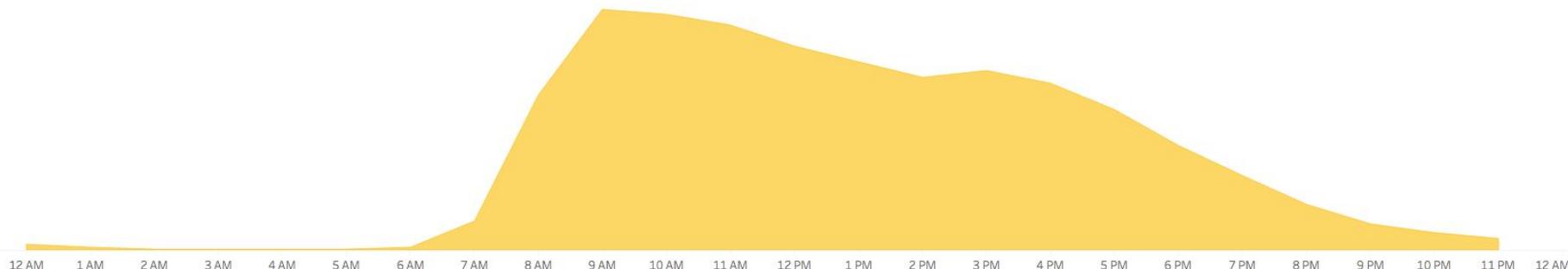
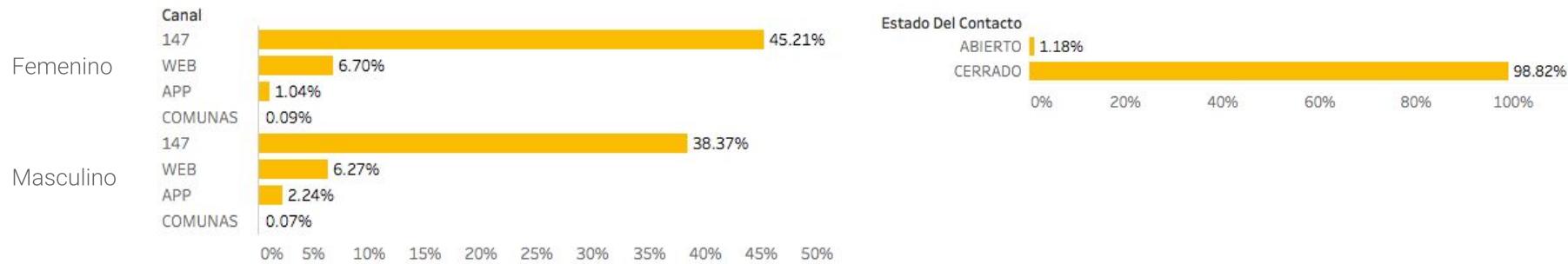


Solicitud de partidas. La mayor concentración de Tickets de observa en la zona de microcentro y Constitución...



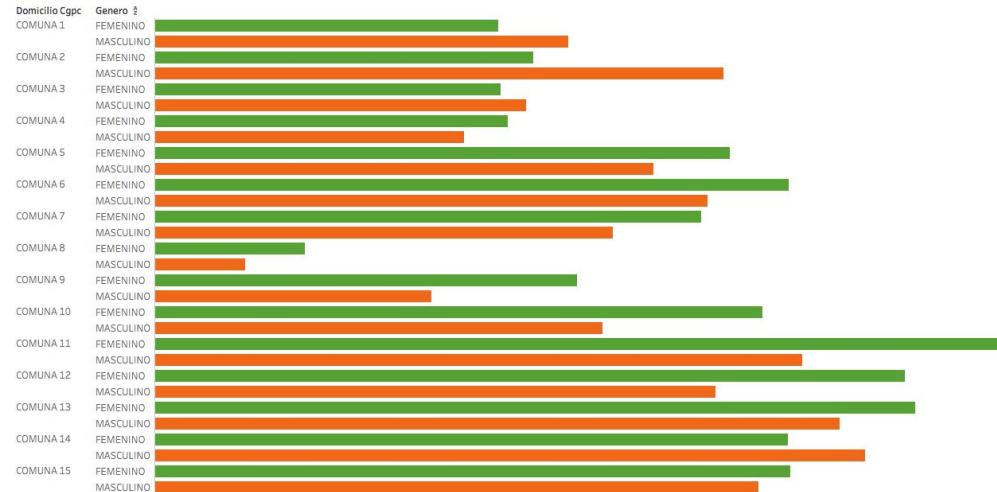
★ **Action item!** Disponibilizar más canales además del Web.

Retiro de escombros. El canal principal para este tipo de Ticket es el 147, su pico se encuentra a las 9 de la mañana y decrece con marcada pendiente a lo largo del día...

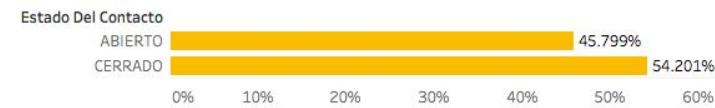
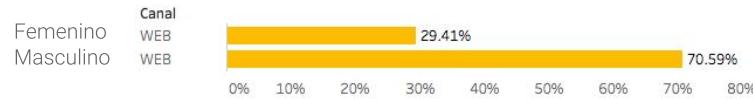




Retiro de escombros. Los tres barrios con más pedidos de recolección de escombros son Palermo, Caballito y Recoleta.



Solicitud de pago voluntario de infracciones. El único canal disponible es el Web, actualmente más del 40% de los tickets se encuentran abiertos...





Solicitud de pago voluntario de infracciones. Los pedidos se encuentran distribuidos de forma uniforme a lo largo de la capital...



★ **Action item!** Disponibilizar más canales además del Web.
Chequear estado de los contactos.

SECCIONES

● INTRODUCCIÓN

RESUMEN DEL PROYECTO

DESCRIPCIÓN DEL DATASET

PREGUNTAS Y MOTIVACIONES

● EXPLORACIÓN DE DATOS

PRIMERAS VISUALIZACIONES

RANKING DE TICKETS

DETALLE POR CONCEPTO

MAPAS Y ACTION ITEMS

● MODELO PREDICTIVO

PREPROCESSING Y MODELO

BACKGROUND TEÓRICO

DESARROLLO DEL MODELO

RESULTADOS

● CONCLUSIONES Y REFERENCIAS

CONCLUSIONES

REFERENCIAS UTILIZADAS

- Preprocessing and feature engineering

Primero eliminamos las columnas con altos valores NaNs. Estas fueron las columnas de altura del domicilio, esquina más cercana y contacto.

Luego eliminamos todas las filas que contenían “DESCONOCIDO” en el campo de Género, con esto eliminamos 15 mil filas.

Había discrepancias entre el dataset del 2017 y el del 2018. Así que al consolidar, curamos la data lo mejor posible. Unificamos acentos, cambiamos nombres, modificamos ubicación de columnas.

Para entrenar nuestro modelo no consideramos variables numéricas, solo categóricas. Estas son: CATEGORIA, TIPO_PRESTACION, DOMICILIO_CGPC, y GENERO. Tomamos todas las variables categóricas del dataset (no dependientes), que puedan explicar de la mejor manera la variable objetivo (CANAL).

Transformamos a dummies las 4 variables explicativas, quedándonos con **40 features** como input.

El objetivo principal del modelo es predecir qué canal de entrada utilizará el usuario, en función de estas cuatro variables.

- Background teórico del modelo

Realizamos un test split. Siendo X el data frame compuesto por las dummies de nuestras cuatro variables categóricas explicativas. Etiqueta "Y" compuesta por el campo CANAL. Test 90% - Train 10%.

Para entrenar el modelo utilizamos data segmentada de Enero del 2017 (para agilizar el entrenamiento del modelo). Utilizamos StandardScaler().fit(x_train) sobre nuestro set de entrenamiento y escalamos este último como también al set de test (con el fit de train).

Utilizamos KNeighborsClassifier con 2 vecinos, a diferencia de 5 (default). Como función de pesos utilizamos distancia y "k parametros" de 20 a 50, con saltos de a 2.

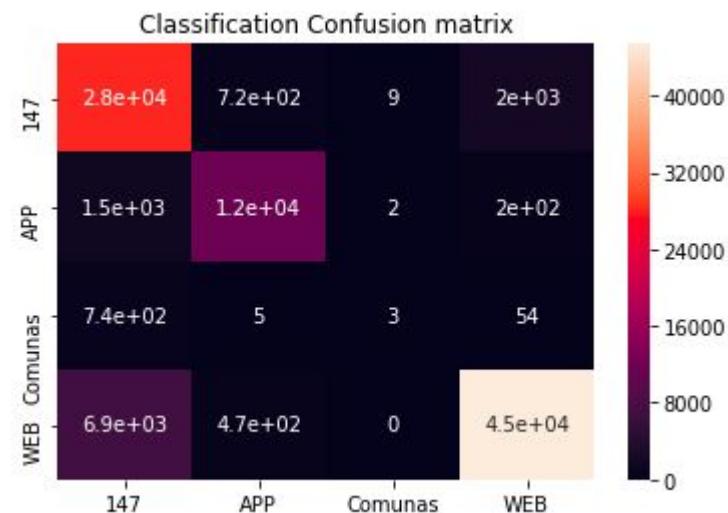
Con el objetivo de mejorar la calidad de nuestro modelo utilizamos GridSearchCV, con los parametros recién mencionados y para evitar overfitting un Cross Validation con 5 folds.

- Resultados test y Enero 2018

El modelo arroja un resultado de **Accuracy: 0.77**. A su vez, un **AUC de 0.84** sobre el set de test.

Realizamos el ejercicio de aplicar este modelo sobre los datos de Enero 2018. Los resultados también fueron buenos: **Accuracy: 0.87**

Matriz de confusión



SECCIONES

● INTRODUCCIÓN

RESUMEN DEL PROYECTO

DESCRIPCIÓN DEL DATASET

PREGUNTAS Y MOTIVACIONES

● EXPLORACIÓN DE DATOS

PRIMERAS VISUALIZACIONES

RANKING DE TICKETS

DETALLE POR CONCEPTO

MAPAS Y ACTION ITEMS

● MODELO PREDICTIVO

PREPROCESSING Y MODELO

BACKGROUND TEÓRICO

DESARROLLO DEL MODELO

RESULTADOS

● CONCLUSIONES Y REFERENCIAS

CONCLUSIONES

REFERENCIAS UTILIZADAS

- Conclusiones

Luego de probar con SVM y KNN, creemos que los resultados alcanzados son los mejores que pudimos obtener. Como principal conclusión pensamos en un caso práctico para nuestro modelo. El objetivo principal es mejorar la experiencia del usuario.

Por ejemplo, si el Gobierno sabe que en la Comuna X, van a existir una gran cantidad de obras para el año siguiente y que el número de tickets sobre retiro de escombros va a aumentar, pueden predecir de qué canal vendrá la mayor cantidad de reclamos. Para de esta forma facilitarle al usuario la carga del ticket a través de un acceso directo desde la APP por ejemplo, o un canal directo por comuna desde el 147.

A su vez pensamos que los accionables provenientes del EDA son de alto valor y que es un dataset con gran potencial para seguir trabajando. El EDA puede ser más extenso pero quisimos abordar en el presente trabajo solamente los puntos más importantes.

- Referencias utilizadas

Stack overflow (**mucho!**)

Scikit learn documentation (**mucho tbm!**)

Tableau documentation

Datascience.stackexchange.com

Towardsdatascience.com

realpython.com

Gracias!

Martina Burone, Blas Leonardo Farias, Ezequiel Talamona.

