# Predicciones de Monto de dólares enviados vía Dinero Express-Money Gram

Pedro Vladimir Hernández Serrano
159549

Proyecto final - Aprendizaje Máquina

Ciencia de Datos - Otoño 2015

## 1 Introducción

Haremos uso de algunas herramientas de Aprendizaje Máquina las cuales nos ayudarán durante el estudio, Sklearn para e preprocesamiento de los datos, Matplotlib para la visualización y Graphlab para el ajuste de los modelos la cual es muy útil ya que no se requiere de demasiadas lineas de programación para entrenar un modelo. Todo el estudio se lleva a cabo en lenguaje Python en su versiónn 2.7 .

El presente documento se divide en 4 partes principales, un vistazo general del negocio y de los datos utilizados, depués se comparan y ajustan algunos métodos de regresión, posteriormente se estudian un par de métodos de clasificación, y finalmente se hacen algunos comentarios del experimento.

El ejercicio se basa en una problemática de Grupo Salinas encarga al área de inteligencia de negocios, cuya estratégia es la de abordarle con las técnicas de ciencia de datos que a continuación se mencionan.

# 2 Caso de Estudio

Uno de los nichos más importantes y con mayor proyección de negocio, son las transferencias de dinero nacional y al extranjero a través de ventanilla, sin la necesidad de una cuenta de banco ni montos mínimos. De esto se dió cuenta Grupo Salinas hace un par de años y ha aprovechado dicho modelo de negocio con la creación de Dinero Express y su afiliación con la agencias Money Gram, de este modo y de manera muy sencilla las personas pueden acuidir a una tienda Elektra las cuales cuentan todas con ventanillas de Banco Azteca y hacer transferencias de dinero en el tipo de cambio del país de origen y enviarlo a cualquier parte del mundo, con solo una identificación, una fotografía y el nombre de la persona que hará el retiro correspondiente por una módica comisión.

Los países de origen son donde existen o han existido tiendas Elektra, los cuales son: México, Guatemala, Honduras, EL Salvador, Panamá, Perú Brasil y Argentina (En Argentina cerraron todas las tiendas en 2013).

Debido a que cada país de origen y de destino tienen diferentes tipos de moneda se lleva a cabo lo siguiente: cuando se efectua el movimiento se hace una conversión general a dolares americanos para cada transacción con el tipo de cambio del día del movimiento, del mismo modo se hace la conversión inversa cuando se hace el retiro en la sucursal de destino, de esta manera se hace estándar el tipo de cambio en todo el análisis presentado.

Se considera los periodos desde el 01 de Enero del 2011 hasta el 30 de Junio del 2015 debido a que anteriormente no se contaban con bases de datos consolidadas, y dado que las operaciones del negocio de transferencia de dinero comenzaron en 2009, se hace un corte en Junio ya que existe un umbral de 3 meses en los que se transfieren de forma definitiva las transacciones de las bases de transferencias hacía la bodega de datos del área de inteligencia de negocio.

# 3 Análisis Exploratorio

En el presente estudio se construye una tabla resumen la cual es por si misma una consulta hecha a la base de transferencias procesadas de la bodega de datos en la que considera los montos acumulados de dinero y el conteo de transacciones agrupado por mes, año, país de origen y país destino, el monto promedio de envío es la división del monto entre el número de transacciones por cada tupla.

| MES | ANIO | PAISORIGEN | ORIGEN | PAISDESTINO | DESTINO | TRANSACCIONES | MONTOENVIOMG | AVGENVIO |
|---|---|---|---|---|---|---|---|---|
| 1 | 2011 | 3 | HONDURAS | 6 | EL SALVADOR | 432 | 58534 | 136 |
| 1 | 2011 | 3 | HONDURAS | 5 | PANAMA | 21 | 2688 | 128 |
| 1 | 2011 | 6 | EL SALVADOR | 2 | GUATEMALA | 540 | 255045 | 472 |
| 1 | 2011 | 5 | PANAMA | 5 | PANAMA | 720 | 111305 | 155 |
| 1 | 2011 | 1 | MEXICO | 4 | PERU | 6167 | 1779414 | 289 |
| 1 | 2011 | 1 | MEXICO | 5 | PANAMA | 1020 | 395155 | 387 |
| 1 | 2011 | 7 | ARGENTINA | 7 | ARGENTINA | 103 | 10648 | 103 |
| 1 | 2011 | 5 | PANAMA | 1 | MEXICO | 1460 | 479200 | 328 |
| 1 | 2011 | 4 | PERU | 4 | PERU | 56760 | 6555750 | 116 |
| 1 | 2011 | 7 | ARGENTINA | 2 | GUATEMALA | 3 | 194 | 65 |

Figure 1: Tabla "Paises" (Primeras 10 Filas)

Una combinación de variables las cuales podrían explicar una a la otra son el número de transacciones por mes de un país a otro y el monto acumulado de los envíos, lo primero que notamos son las escalas grandes, hablamos de cientos de miles de dolares, se necesitará un escalamiento para tener un mejor ajuste del modelo.

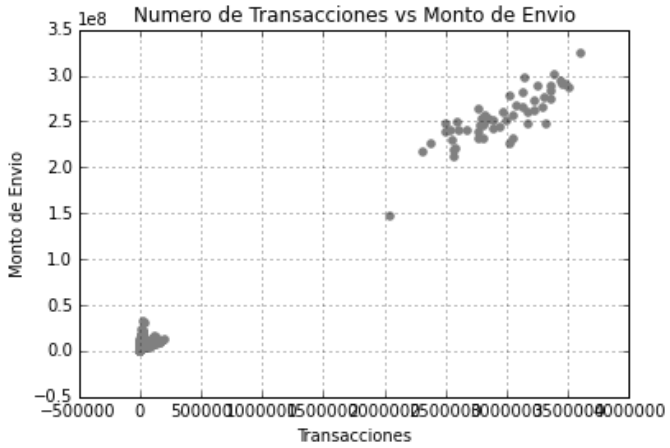



Figure 2: Países

En la figura 3 se presenta un resumen descriptivo de las variables, nos ayuda a darnos cuenta cómo se comportan, las que se presentan como frecuencias se refieren a variables categóricas.

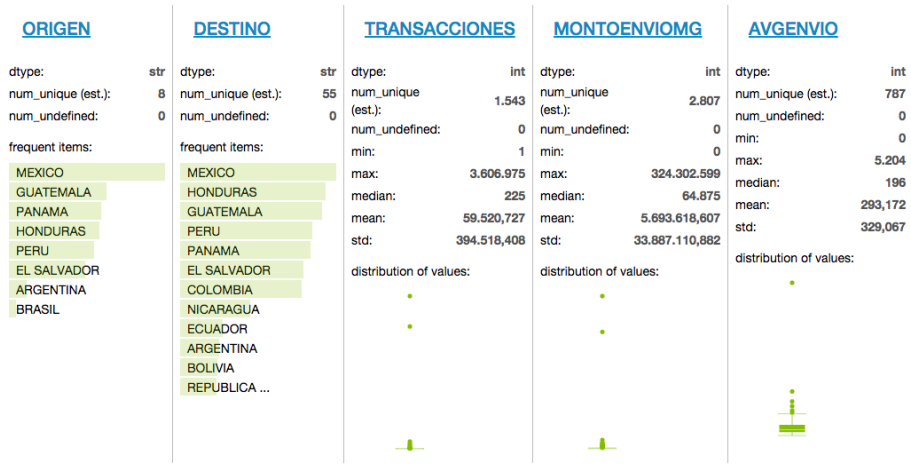



Figure 3: Resumen Variables

Con ayuda del gráfico podemos notar que los países desde donde se hacen envíos de dinero son claramente aquellos donde las tiendas Elektra tienen presencia, como se había mencionado anteriormente.

Con respecto a los países a donde llegan los envíos, notamos que son 55 diferentes desde 2011 a la fecha de corte, se tiene un promedio de transacciones por mes de país a país de $59,520,727$ un promedio de monto envíado de $5,693,618,607$ $USD$ de manera muy general, el promedio de Monto promedio de envío $293,172$ $USD$ nos da una idea del movimiento de dólares que se hacen de manera mensual de un país a otro operado por MoneyGram-Dinero Express.

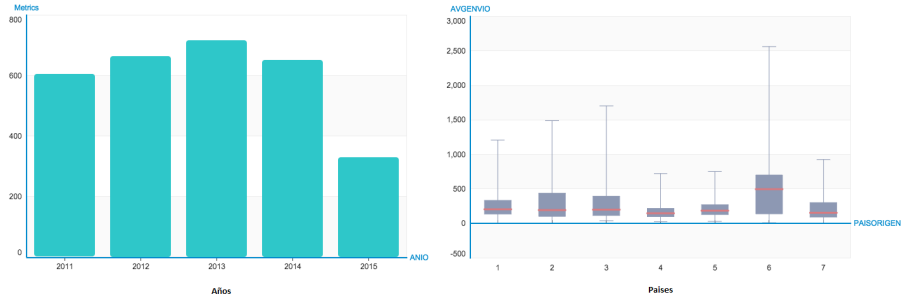



Figure 4: Monto total en envíos (en Millones)

En la figura 4 se observa el monto acumulado de dinero en operaciones de manera anual, como se mencionó antes, en 2015 se ve a la mitad ya que se cuenta con información hasta el primer semestre del año, precisamente uno de los objetivos del presente ejercicio es el de tener modelos adecuados para el pronostico de los meses con los que se cerrará el año.

Del mismo modo en el gráfico de cajas podemos notar con mayor claridad el comportamiento de la variable promedio de envío, se observa que el país con mayor promedio de envío y mayor variabilidad en sus envíos es Panamá (correspondiente a la categoría 6), por lo que las predicciones de envío en aquel país serán las menos certeras, debido al alto grado de varianza, por el momento nos concentramos en México, dado que los tomadores de deciisiones desean proyectar la cobertura de Dinero Express en el México para el cierre del año.

# 4 Modelos y Metodología

## 4.1 Preprocesamiento

Lo primero notamos al analizar los datos, fue el tamaño en que se presentan los montos de envío, el número de transacciones así como el envío promedio, por lo que será necesario un re-escalamiento de dichas variables, se invocó la manera tradicional de escalamiento bajo una distribución normal estandar, los arreglos de cada variable se agregaron como nuevas columnas a los metadatos de la tabla paises para una mejor manupulación, se ilustra en la figura 5.

| MES | ANIO | ORIGEN | DESTINO | TRANS_SCALED | MONTO_SCALED | AVG_SCALED |
|---|---|---|---|---|---|---|
| 1 | 2011 | HONDURAS | EL SALVADOR | -0.149774322923 | -0.166289909643 | -0.16801321973 |
| 1 | 2011 | HONDURAS | PANAMA | -0.150816099408 | -0.167937910861 | -0.168013455808 |
| 1 | 2011 | EL SALVADOR | GUATEMALA | -0.149500571438 | -0.160490920148 | -0.168003304455 |
| 1 | 2011 | PANAMA | PANAMA | -0.149044318963 | -0.164732650891 | -0.168012659045 |
| 1 | 2011 | MEXICO | PERU | -0.135237612119 | -0.115507179714 | -0.168008704739 |
| 1 | 2011 | MEXICO | PANAMA | -0.148283898171 | -0.15635630978 | -0.168005812784 |
| 1 | 2011 | ARGENTINA | ARGENTINA | -0.150608251058 | -0.167703013291 | -0.168014193551 |
| 1 | 2011 | PANAMA | MEXICO | -0.147168614343 | -0.153876163264 | -0.168007553859 |
| 1 | 2011 | PERU | PERU | -0.00699771504058 | 0.0254412775521 | -0.168013809925 |
| 1 | 2011 | ARGENTINA | GUATEMALA | -0.150861724656 | -0.168011508165 | -0.168015314922 |

[2964 rows x 7 columns]
Note: Only the head of the SFrame is printed.

Figure 5: Variables Escaladas

Antes de comenzar con el ajuste del modelo definimos dos grupos en el conjunto de datos con diferentes fines, sklearn nos ayuda a separar de manera aleatoria un sub grupo de entrenamiento y otro de evalucación de los datos originales, esto con el obbjetivo de llevar a cabo el entrenamiento del modelo con el primer grupo, y hacer las evaluaciones de la eficacia del modelo con el segundo grupo, arbitrariamente utilizamos 80% y 20% para cada uno. Resultando entonces $2,369$ registros para el entrenamiento y 595 para pruebas.

## 4.2 Regresión lineal con una variable explicativa

En la figura 6 se muestra el resumen del entrenamiento del modelo, pareciera un ajuste muy bueno, pero posiblemente engañoso, ya que el coeficiente para la variable transacciones es casi 1 y para el intercepto es casi 0, debido a lo anterior se considera después una regresión que considere las demás variables, pero el hecho de agregar variables, resulta en un efecto no proporcional hacía la función objetivo, es por ello que se regularizó el entrenamiento.

## 4.3 Regresión lineal dinámica regularizada considerando 5 variables explicativas

Tomamos entonces las demás variables Origen, Destino, las variables escaladas y la dicotómica "México", se utilizó la técnica de regularización Lasso, la cual se considero un factor de penalización del 1% como se muestra en la figura 6.

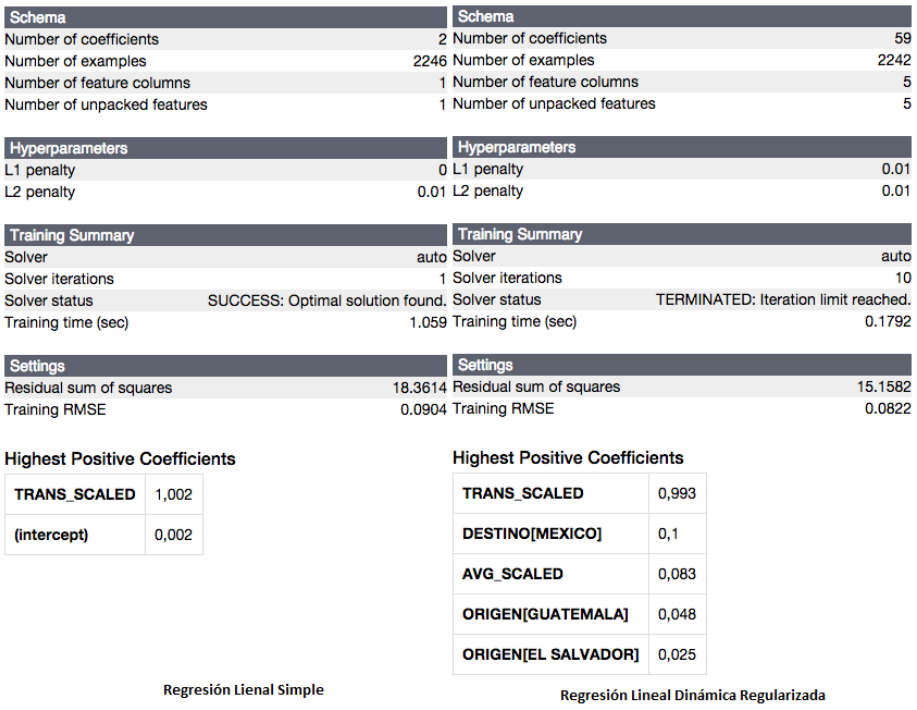



Figure 6: Resumen entrenamiento de la Regresión Lineal

El error del modelo es menor, pero para este punto hace falta una última consideración, y es qué, las variables Origen, Destino y la dicotómica México, son categóricas, por lo que se procedió a utilizar una técnica más de entrenamiento con la cual trabaje muy bien con dicho tipo de variables sin meternos en problemas de la "maldición de la dimensionalidad", se entrenó entonces una regresión utilizando bosques aleatorios.

## 4.4 Regresión con Bosques Aleatorios

Bajo la misma premisa, se considera de igual modo las variables que el modelo anterior.

| Schema | |
|---|---|
| Number of examples | 2369 |
| Number of feature columns | 5 |
| Number of unpacked features | 5 |

| Settings | |
|---|---|
| Number of trees | 10 |
| Max tree depth | 6 |
| Train RMSE | 0.0617 |
| Validation RMSE | |
| Training time (sec) | 0.0765 |

**Regresión con Bosques Aleatorios**

Figure 7: Resumen entrenamiento de Bosques Aleatorios

Se observa claramente el la figura 7 un menor error, por lo que trataremos este último como el definitivo en este experimento para predecir, en el gráfico 8 se ilustra el ajuste con el conjunto de prueba y posteriormente con toda la tabla.

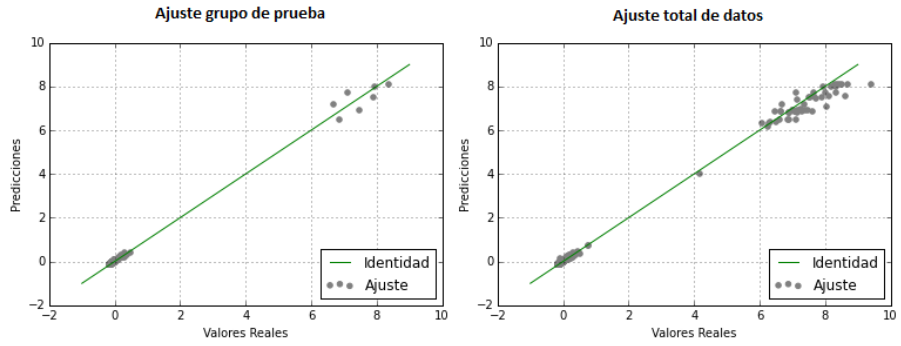



Figure 8: Ajuste de Regresión

Para este punto podemos presumir de un buen modelo, pero como mencionamos anteriormente no será lo suficientemente bueno si no hasta predecir.

Se practicó una estimación de un mes en particular. Definimos el subgrupo de los envios de dinero en el mes de Enero que provienen de Perú (El segundo país con más transacciones) y tienen como destino México, con ayuda de la

| Modelo | rmse |
|---|---|
| **Regresión simple** | 0.092116 |
| **Regresión dinámica** | 0.085489 |
| **Bosques Aleatorios** | 0.052707 |

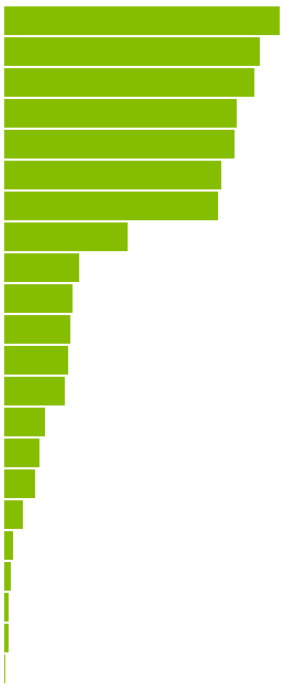
Table 1: Comparativo Error cuadrado medio

función predictora resulta una tabla con los valores reales comparado con las estimaciones y se agregó una columna adicional para las diferencias.
Finalmente invocando una vez más a la función predictora pero con tendencia obtenemos $326,054$ con un error de $+-9,253\ USD$ para el mes de Enero del 2016.

## 4.5 Clasificación

Continuamos con la segunda parte del estudio, y tenemos como premisa que México es el país donde se concentra el mayor movimiento con respecto a los envíos desde otros países, nos interesa saber aquellos puntos que en particular ocurren con poca frecuencia, por lo que estamos interesados en saber si ocurre o no en determinadas temporalidades del año.

| Value | Count | Percent |
|---|---|---|
| MEXICO | 352 | 11,876% |
| HONDURAS | 327 | 11,032% |
| GUATEMALA | 320 | 10,796% |
| PERU | 297 | 10,02% |
| PANAMA | 294 | 9,919% |
| EL SALVADOR | 278 | 9,379% |
| COLOMBIA | 273 | 9,211% |
| NICARAGUA | 158 | 5,331% |
| ECUADOR | 96 | 3,239% |
| ARGENTINA | 87 | 2,935% |
| BOLIVIA | 84 | 2,834% |
| REPUBLICA ... | 81 | 2,733% |
| COSTA RICA | 77 | 2,598% |
| PARAGUAY | 52 | 1,754% |
| BRASIL | 45 | 1,518% |
| URUGUAY | 40 | 1,35% |
| FILIPINAS | 24 | 0,81% |
| RUSIA | 11 | 0,371% |
| ESTADOS UNIDOS | 9 | 0,304% |
| CHINA | 5 | 0,169% |
| ESPA&#65533;A | 5 | 0,169% |
| BENIN | 2 | 0,067% |



| Value | Count | Percent |
|---|---|---|
| ITALIA | 2 | 0,067% |
| ALEMANIA | 2 | 0,067% |
| BELICE | 2 | 0,067% |
| CAMERUN | 2 | 0,067% |
| GRECIA | 2 | 0,067% |
| EGIPTO | 2 | 0,067% |
| TRINIDAD Y TOBAGO | 2 | 0,067% |
| PUERTO RICO | 2 | 0,067% |
| INDIA | 2 | 0,067% |
| CHILE | 2 | 0,067% |
| VENEZUELA | 2 | 0,067% |
| BULGARIA | 2 | 0,067% |
| REINO UNIDO | 2 | 0,067% |
| CROACIA | 1 | 0,034% |
| LESOTO | 1 | 0,034% |
| MACEDONIA | 1 | 0,034% |
| BOSNIA - ... | 1 | 0,034% |
| REPUBLICA CHECA | 1 | 0,034% |
| PAISES BAJOS | 1 | 0,034% |
| BELGICA | 1 | 0,034% |
| GEORGIA | 1 | 0,034% |
| ARMENIA | 1 | 0,034% |

Figure 9: Frecuencia de envíos por país

## 4.6 Regresión Logística

La variable natural que en principio podría explicar el destino del dinero hacía un país dado los datos es el monto promedio, ya que como se vio anteriormente, en México se encruentra balanceado, pero no conocemos el comportamiento de los paises desde donde se hace el movimiento.

Ajustamos un modelo de regresión logistica cuya variable explicativa es el monto promedio de envío y la variable de salida fue "México" (variable dicotómica que indica si llegó o no al menos una transferencia en determinado mes).

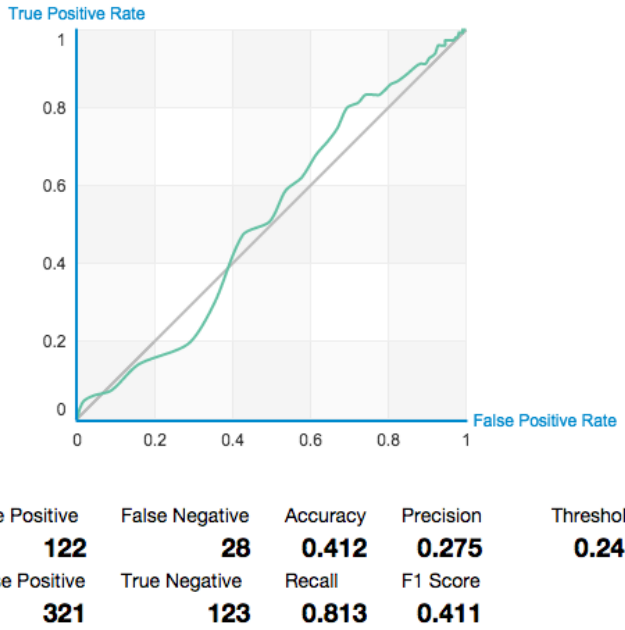
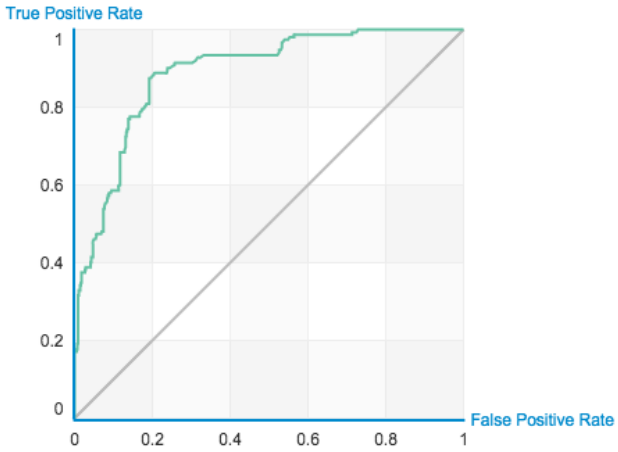



Figure 10: Bondad de ajuste, una variable

Para probar la bondad de ajuste se calculó una matriz de confusión y a partir de las probabilidades se construyó una curva ROC, aunque no se ve muy bien el modelo, el Accuracy de menos del 50% debido al altísimo índice de falsos positivos, es decir que debió indicar 0 pero predijo 1, lo anterior podemos de hecho notarlo en la forma de la curva, nos disponemos entonces de buscar un ajuste con las demás variables; DESTINO, TRANSACCIONES, MONTOENVIOMG, AVGENVIO y entrenamos.

La curva se ve mucho mejor, pero cabe mencionar en este punto qué los falsos positivos son los que más dañan a la empresa debido al tipo de cambio y a al documentación que se pide, ya que se requiere mayor número de credenciales en el lugar de origen, no así del destino. Probemos entonces un método diferente.




| True Positive | False Negative | Accuracy | Precision | Threshold |
|---|---|---|---|---|
| **59** | **93** | **0.813** | **0.766** | **0.501** |
| **False Positive** | **True Negative** | **Recall** | **F1 Score** | |
| **18** | **425** | **0.388** | **0.515** | |

Figure 11: Bondad de ajuste, cuatro variables

## 4.7 Máquina de soporte vectorial

Utilizamos las mismas variables que en el modelo anterior, entrenamos el modelo como se muestra en la figura 12, Posterior al análisis de los modelos anteriores, hacemos un comparativo. El último modelo lo podemos elegir como definitivo, se presenta entonces el umbra de decisión y unas predicciones.

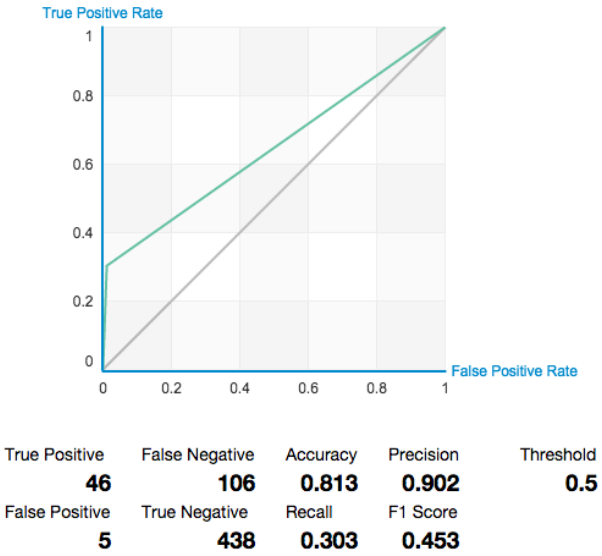



Figure 12: Bondad de ajuste, SVM

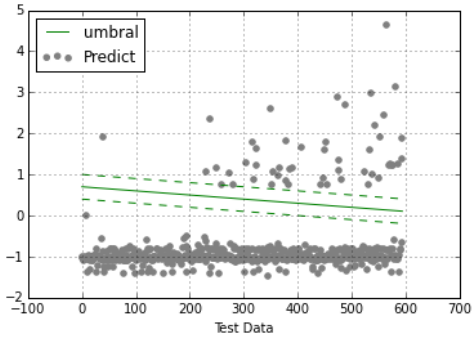



Figure 13: Umbral de Decisión SVM

Análogamente al modelo de regresión lineal disponemos del sub grupo de mes de Enero como país de destino México, y esta vez generalizamos a todos los países, ya que lo que queremos observar es si llega hacía México desde alguno de los países no recurrentes con la finalidad de encontrar patrones o algún tema relacionado con la económia en particular.

| ANIO | ORIGEN | DESTINO | MEXICO | PREDICT |
|---|---|---|---|---|
| 2011 | PANAMA | MEXICO | 0 | 0 |
| 2011 | ARGENTINA | MEXICO | 0 | 0 |
| 2011 | GUATEMALA | MEXICO | 0 | 0 |
| 2011 | HONDURAS | MEXICO | 0 | 0 |
| 2011 | EL SALVADOR | MEXICO | 0 | 0 |
| 2011 | PERU | MEXICO | 0 | 0 |
| 2011 | MEXICO | MEXICO | 1 | 1 |
| 2012 | PERU | MEXICO | 0 | 0 |
| 2012 | EL SALVADOR | MEXICO | 0 | 0 |
| 2012 | MEXICO | MEXICO | 1 | 1 |

[33 rows x 5 columns]
Note: Only the head of the SFrame is printed.

Figure 14: Envíos hacía México

| MES | ORIGEN | MONTO_ESTIMADO |
|---|---|---|
| 12 | GUATEMALA | 14673870.8649 |
| 12 | ARGENTINA | 10110434.388 |
| 12 | BRASIL | 168154.565291 |
| 12 | PERU | 11600895.8995 |
| 12 | PANAMA | 1507863.577 |
| 12 | MEXICO | 142902682.922 |
| 12 | HONDURAS | 12938441.1907 |
| 12 | EL SALVADOR | 4020103.969 |

[8 rows x 3 columns]

Figure 15: Predicción Cierre 2015

# 5 Discusión

El problema anterior tiene mucho material para extenderse, así como para continuar con nuevas pruebas y modelos, se eligieron los anteriores modelos debido a su naturaleza intuitiva con respecto a la utilización de las variables y a su algorítmo matemático.

Pero de cualquier modo se podría probar métodos adicionales y posiblemente integrarlos con técnicas como aprendizaje en ensamble, en clasificación por ejemplo podemos explorar los demás paises desde donde se hacen transferencias de dinero, notamos casos particualares como panamá por ejemplo que tiene el mayor promedio de envío, esto nos indica que menos personas envian más dinero, y a primera instancia podría parecer natural ya que la moneda utilizada en ese país es el dolar (misma con la que se hace el estudio), si se llegara a demostrar que la diferencia del envío promedio con los demás paises es directamente proporcional a la tasa de poder adquisitivo entonces tiene sentido el promedio alto mencionado, pero si no es así entonces podría significar un tema de fuga de capital o de lavado de dinero. Más aún, el presente estudio se puede replicar a un nivel de desagregación menor, como estados, ciudad o incluso hasta sucursal o persona lo cual significarían estadísticas con información directa para la toma de desciciones. Únicamente que se requeriría mayor poder de cómputo así como conocimiento avanzado de técnicas de big data.

Uno de los problemas encontrados durante el estudio es que el modelo se construye con los paises existentes, es decir, aparece un país en la base de datos si se efectua un envío al mismo, y el modelo entrena sobre los que hay, por lo que no podríamos predecir el movimiento de dolares a aquellos paises a los que no se aparecen en el histórico. Aquí existe entonces un área de oportunidad ya que se puede modelar un algoritmo adaptativo en el que combinado con técnicas de clusterización, se definan similitudes con paises a los que se han hecho envios de dinero y proponga paises nuevos para los años siguientes a partir de su similitud. Un buen tema para Aprendizaje Máquina 2 :)

Por lo pronto nos aventuramos únicamente a proyectar la actividad del cierre del ejercicio fiscal 2015 como se mostró en la figura 15