Final Project 2 | Data Science

Original Business Problem:

Of the three problems presented during the lightening presentation the one that would potentially provide the most impact to the business is USA Scheme Cost Prediction and Modeling (Potential Project 3). Originally a linear regression was envisioned and this evolved to a random forest decision tree model since costs can never be below zero. This can be the case of linear regression.

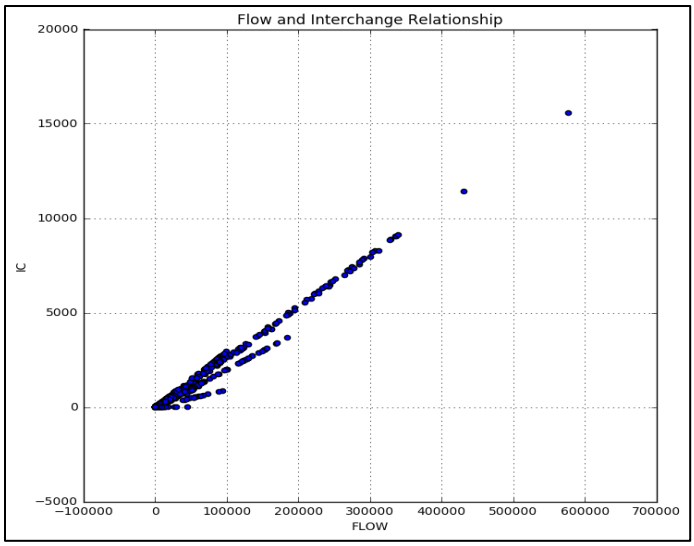
The intention was to develop an application based on business type where we could understand the median costs and use it to make pricing and breakeven threshold decisions.

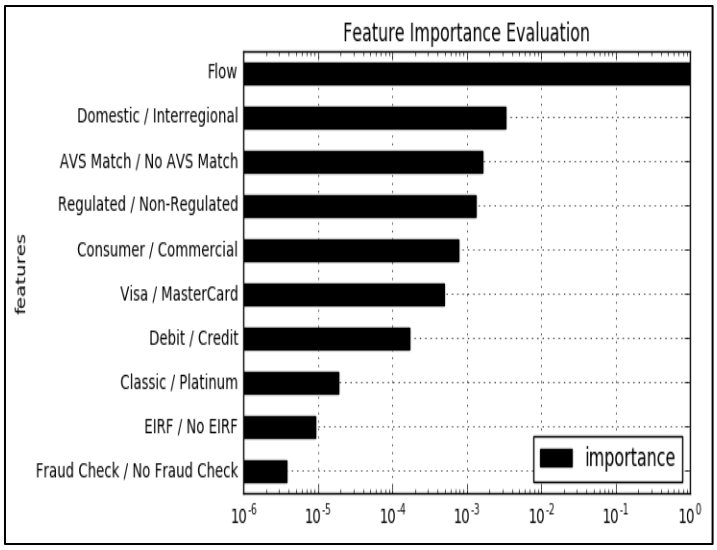
Source:

The information came directly the company's cost databases and features were extrapolated from string data in a few of the columns.

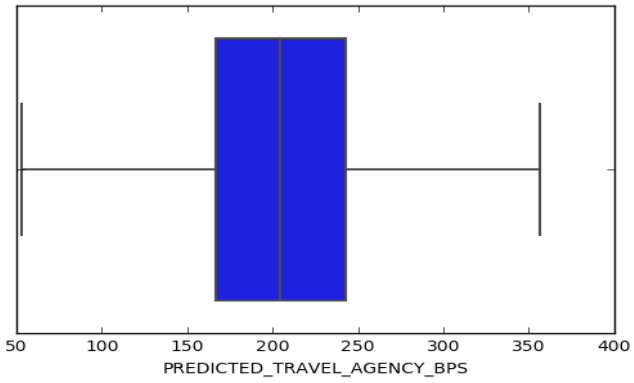
Results:

The random forest decision tree did an excellent job at predicting scheme fees based off a number of features from a year's worth of cost data and accounted for the outliers that skewed the original linear regression. The features chosen could be easily reviewed based on the respective industry/merchant.





The application ultimately did what it was supposed to, identify a median cost for certain industries based on the model from the test datasets.



Implement the Random Forest Regression

In [31]:
```
regr = RandomForestRegressor(n_estimators=10)
print regr.fit(trainX,trainY)

feat_importance_org=regr.feature_importances_
yhat=pd.DataFrame(regr.predict(testX)) #Predict test test set costs using the random forest regression
y=pd.DataFrame(testY) #
```

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features='auto', max_leaf_nodes=None,
           min_impurity_split=1e-07, min_samples_leaf=1,
           min_samples_split=2, min_weight_fraction_leaf=0.0,
           n_estimators=10, n_jobs=1, oob_score=False, random_state=None,
           verbose=0, warm_start=False)
```

Evaluate the Random Forest Regression Score Against the Test Dataset

In [32]:
```
print regr.score(testX,testY)
```

```
0.997840394287
```

Business Problem Evolution:

After the original business problem was proposed to management (CSO) another idea was presented. It is can random forest regression use patterns in the data features to identify instances where costs can be optimized.

Updated Business Project:

Project Problem and Hypothesis

- **What's the project about? What problem are you solving?**
  - Currently our company has no way of identifying “downgrades.” Said downgrades are instances where costs are moved to a more expensive tier of costs because certain reporting requirements are not being met. There are thousands of instances of downgrades in the data but at the moment no one team has ventured far enough to link cost behavior at a transactional level, extrapolate the features and use machine learning to identify downgrade behavioral patterns for the entire portfolio. This is due to the fact that many of the cost datasets provided by our partner banks are very limited.
  - The problem attempting to be solved is can the downgrades be predicted using a large dataset from one partner that provides a quality dataset and applied to the entire portfolio.
- **Where does this seem to reside as a machine learning problem? Are you predicting some continuous number, or predicting a binary value?**
  - Since our cost data that we can link to the backend transactional data identifies downgrades by name, each transaction can be identified as a downgrade or not downgrade.

- **What kind of impact do you think it could have?**
  - The potential impact is tremendous. If these downgrades can be predicted using the trained model then offending merchants can be identified and encouraged to meet the requirements just based on our transactional data. There will be no need to assume how much cost the organization is incurring in the patchy cost information provided by the majority of the acquiring partners. Consequently if the offending merchants change their behavior costs will drop potentially saving the company 100,000s of dollars.
- **What do you think will have the most impact in predicting the value you are interested in solving for**
  - Determining the features will be key particularly from the cost information that can be tied to the backend transactional information.

Datasets

- **Description of data set available**
  - Numerous joined live SQL Neteeza databases will be used

```
SELECT
    t1.STATUSID
    ,t1.AUTHORISATIONDATETIME,t1.DATEAUTHORISATION,t1.STATUSDATE,t1.RECEIVEDDATE,t1.PAYMENTDATE
    ,t1.MERCHANTID,t1.ORDERID,t1.CREDITCARDCOMPANY
    ,t1.CREDITDEBITINDICATOR,t1.CVVINDICATOR,t1.CVVRESULT,t1.CVVSERVICEINDICATOR,t1.AVSINDICATOR,t1.AVSRESULT,t1.FRAUDCODE,t1.FRAUDINDICATOR,t1.FRAUDRESULT
    ,t1.TVSRESULT, t1.EVSRESULT, t1.ZVSRESULT, t1.SVSRESULT, t1.NVSRESULT,t1.STTINDICATOR
    ,t1.MERCHANTREFERENCE,t1.MERCHANTNAME,t1.PAYMENTREFERENCE --ADDED IN THE MERCHANT COUNTRY FROM ACCOUNT VALIDATION TABLE
    ,t1.SUB_VERTICAL
    ,t1.CUSTOMERID,t1.ISSUER_COUNTRY_CODE,t1.MERCHANT_COUNTRY_CODE
    ,t1.AUTHORISEDCURRENCYCODE,t1.AUTHORISEDAMOUNT,t1.AMOUNT1,t1.AMOUNT2 --FLOW AND CURRENCY
    ,t1.PROVIDERNAME, t1.PROVIDERDESCRIPTION,t1.PAYMENTPROCESSOR,t1.SERVICEPROVIDERID,t1.PAYMENTMETHODID,t1.PAYMENTPRODUCTID,t2.DESCRIPTION,t3.PAYMENTPRODUCTNAME
    ,t4.INTERCHANGE_RATE,t4.AVS_RESPONSE_MESSAGE,t1.IIN, t5.INTERCHANGE_RATE, t4.VANTIV_FLOW,t5.IC_DESC,

CASE WHEN t1.CVVINDICATOR = 1 then 'Checked'
WHEN t1.CVVINDICATOR <> 1 then 'Not Checked'
END AS "CVV Check",

CASE WHEN t1.CVVRESULT = 'M' THEN 'Match'
WHEN t1.CVVINDICATOR = 'N' THEN 'NoMatch'
WHEN t1.CVVRESULT not in ('N','M') THEN 'Unknown'
END AS "CVV Result",

CASE WHEN t1.AVSRESULT in ('Z','A','W','D') THEN 'Partial'
WHEN t1.AVSRESULT in ('X','Y','M','F','P') THEN 'Full'
WHEN t1.AVSRESULT in ('N') THEN 'None'
WHEN t1.AVSRESULT in ('U','G','R','0','S') THEN 'Unsupp/Inconc'
END AS "AVSRESULT"

FROM
    (
        SELECT
        pa.STATUSID
        ,co.AUTHORISATIONDATETIME,co.DATEAUTHORISATION,pa.STATUSDATE,pa.RECEIVEDDATE,pa.PAYMENTDATE
        ,co.MERCHANTID,co.ORDERID,co.CREDITCARDCOMPANY
        ,co.CREDITDEBITINDICATOR,co.CVVINDICATOR,co.CVVRESULT,co.CVVSERVICEINDICATOR,co.AVSINDICATOR,co.AVSRESULT,co.FRAUDCODE,co.FRAUDINDICATOR,co.FRAUDRESULT
        ,co.TVSRESULT, co.EVSRESULT, co.ZVSRESULT, co.SVSRESULT, co.NVSRESULT,co.STTINDICATOR
        ,co.MERCHANTREFERENCE,m.MERCHANTNAME,m.COUNTRYCODE AS MERCHANT_COUNTRY_CODE, pa.PAYMENTREFERENCE --ADDED IN THE MERCHANT COUNTRY FROM ACCOUNT VALIDATION TABLE
        ,mm. SUB_VERTICAL
        ,co.CUSTOMERID,co.COUNTRYCODE AS ISSUER_COUNTRY_CODE
        ,co.AUTHORISEDCURRENCYCODE,co.AUTHORISEDAMOUNT,co.AMOUNT AS AMOUNT1,pa.AMOUNT AS AMOUNT2 --FLOW AND CURRENCY
        ,sp.PROVIDERNAME, sp.PROVIDERDESCRIPTION,co.PAYMENTPROCESSOR,co.SERVICEPROVIDERID,pa.PAYMENTMETHODID,pa.PAYMENTPRODUCTID,co.IIN
        FROM EPS.PCO_CREDITCARDONLINE co

        LEFT JOIN EPS.GPM_SERVICEPROVIDER sp
        ON co.SERVICEPROVIDERID=sp.SERVICEPROVIDERID
```

Domain knowledge

- **What experience do you already have around this area?**
  - The experience is advanced. Familiar with SQL , Python, Random Forest, and understand the business application of each column and potential data feature.
- **What other research efforts exist?**
  - No as mentioned above the CSO wants this explored and our merchant solutions team confirmed that this project has not yet been explored due to a lack of local expertise.

Project Concerns

- **What questions do you have about your project? What are you not sure you quite yet understand? (The more honest you are about this, the easier your instructors can help).**
  - None at the moment.
- **What are the assumptions and caveats to the problem?**
  - We're using only one large dataset from one acquiring bank to assume behavior for the entire portfolio but this is dangerous because we're making assumptions that merchants under different acquirers are behaving the same.
- **What's the cost of your model being wrong? (What's the benefit of your model being right?)**
  - The cost of the data being wrong is that merchants that don't really have downgrades could be flagged and notified and if their validations contradict the model then business confidence in the project could potentially falter. Furthermore, the impact calculations derived from this classification project could inaccurately predict potential savings that aren't really there.
- **Is any of the data incorrect? Could it be incorrect?**
  - Yes. Binning data is occasionally flagging consumer and commercial transactions incorrectly. This however can be corrected during the data munging portion of the process.

Outcomes

- **What do you expect the output to look like?**
  - The output will most likely flag the information where it's predicted to be a downgrade. The downgrade offenders will be grouped and their impact to the portfolio will be assessed. The data will be divided based on merchant and their specific industry
- **What does your target audience expect the output to look like?**
  - An analysis of the data divided based on merchant and their specific industry
- **What gain do you expect from your most important feature on its own?**
  - I don't have one most important feature, the project will seriously be a combination of features affecting the result.
- **How complicated does your model have to be?**
  - The most complicated portion of the model will be feature creation and selection. There will be many features but they're all imperative.
- **How successful does your project have to be in order to be considered a "success"?**
  - If the model can successful predict against merchants for the one acquirer being used at 95% then an assumption can be made that the model can be expanded to merchants under other acquires.
- **What will you do if the project is a bust (this happens! but it shouldn't here)?**
  - Get feedback and try again but confident the model will succeed.