ANZ - TASK 2
Predictive Analytics
(2nd Submission)

Prepared by
Teik Ning Yang (16/6/2019)

Data Set

- This task is based on a synthesised transaction dataset containing 3 months' worth of transactions for 100 hypothetical customers. It contains purchases, recurring transactions, and salary transactions.

- The dataset is designed to simulate realistic transaction behaviours that are observed in ANZ's real transaction data

- There are many Null values in the dataset. After investigation, these null values are corresponded to the transactions. For instance, merchant related attributes are only corresponded to Sale-POS/POS transaction.

```
df.info()
```

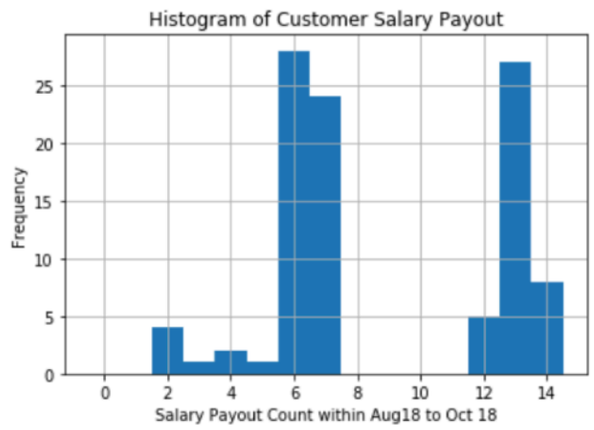
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12043 entries, 0 to 12042
Data columns (total 23 columns):
status                12043 non-null object
card_present_flag     7717 non-null float64
bpay_biller_code      885 non-null object
account               12043 non-null object
currency              12043 non-null object
long_lat              12043 non-null object
txn_description       12043 non-null object
merchant_id           7717 non-null object
merchant_code         883 non-null float64
first_name            12043 non-null object
balance               12043 non-null float64
date                  12043 non-null datetime64[ns]
gender                12043 non-null object
age                   12043 non-null int64
merchant_suburb       7717 non-null object
merchant_state        7717 non-null object
extraction            12043 non-null object
amount                12043 non-null float64
transaction_id        12043 non-null object
country               12043 non-null object
customer_id           12043 non-null object
merchant_long_lat     7717 non-null object
movement              12043 non-null object
dtypes: datetime64[ns](1), float64(4), int64(1), object(17)
memory usage: 2.1+ MB
```

Data Exploration

- Both Charts on the right present the distribution of the salary payment counts and Salary of the customers
- for simplicity, we estimate the Annual Salary based on :
  - Weekly pay ==> 12-14 pay received in the period
  - Fortnightly ==> 6- 7 pay received in the period
  - Monthly ==> 2 - 5 pay received in the period

```
df_S.Pay_Cnt.hist(bins = 15, range = (0,15), align = 'left')
plt.title('Histogram of Customer Salary Payout')
plt.xlabel('Salary Payout Count within Aug18 to Oct 18')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```




```
df_S.Pay.hist(bins = 20, align = 'left', xlabelsize=10, xrot=None, ylabelsize=10, yrot=None,)
plt.title('Histogram of Customer Salary')
plt.xlabel('Salary $')
plt.ylabel('Frequency')
```

```
Text(0, 0.5, 'Frequency')
```


```
df_S['Annual_Sal'] = 0
for i in range(0,len(df_S.Pay_Cnt)):
    if df_S['Pay_Cnt'][i] >=12:
        df_S['Annual_Sal'][i] = df_S['Pay'][i] / 7 *365.25
    elif df_S['Pay_Cnt'][i] <=5:
        df_S['Annual_Sal'][i] = df_S['Pay'][i] * 12
    else:
        df_S['Annual_Sal'][i] = df_S['Pay'][i] / 14 *365.25
df_S.head()
```

Data Exploration

- We observed that the Pay for each customer is constant.some of the customer have missing payment within the period, probably before they are casual staffs?

- On the right, I extracted some strange observation due to multiple salary payments on same date. There is 7 occurrences for 4 customers

- As mentioned in earlier slides, for simplicity, I grouped them based on count of payments in the period into monthly, fortnightly and weekly payment

Analyzing Salary payment time

```
# Select data that matches Pay/salary transactions
df1 = df.loc[df['txn_description'] =='PAY/SALARY']
# df1['date'] = pd.to_datetime(df['date']) already in datetime
df1 = df1.sort_values(by = ['account', 'date'], ascending = False)

df1['Pay_time'] = df1.groupby(['account'])['date'].apply(lambda x: x.diff())
df1.Pay_time.value_counts()

#for acc in df1.account.values:
#    count = df1.loc[df1['account'] ==acc].account.count()
#    for i in range(0,count):
#    print(acc, count)
```

```
-7 days      478
-14 days     279
0 days         7
-28 days       5
-31 days       5
-30 days       5
-29 days       2
-61 days       1
-32 days       1
Name: Pay_time, dtype: int64
```

‘0’ day gap salary payment

Acc	Txt	Date	Amt
ACC-1523339231	PAY/SALARY	8/20/18	8835.98
ACC-1523339231	PAY/SALARY	9/20/18	8835.98
ACC-1523339231	PAY/SALARY	10/19/18	8835.98
ACC-1523339231	PAY/SALARY	10/19/18	8835.98
---	---	---	---
ACC-2270192619	PAY/SALARY	8/6/18	3026.95
ACC-2270192619	PAY/SALARY	9/6/18	3026.95
ACC-2270192619	PAY/SALARY	10/5/18	3026.95
ACC-2270192619	PAY/SALARY	10/5/18	3026.95
---	---	---	---
ACC-1683215619	PAY/SALARY	8/17/18	2500
ACC-1683215619	PAY/SALARY	8/17/18	2500
ACC-1683215619	PAY/SALARY	9/18/18	2500
ACC-1683215619	PAY/SALARY	9/18/18	2500
ACC-1683215619	PAY/SALARY	10/18/18	2500
ACC-1683215619	PAY/SALARY	10/18/18	2500
---	---	---	---
ACC-354106658	PAY/SALARY	8/15/18	5103.51
ACC-354106658	PAY/SALARY	9/14/18	5103.51
ACC-354106658	PAY/SALARY	9/14/18	5103.51
ACC-354106658	PAY/SALARY	10/15/18	5103.51
ACC-354106658	PAY/SALARY	10/15/18	5103.51

Data Preparation

- The predicting attributes used are : Age, debit transactions (i.e. Pos. Sales-Pos, Payment, Interbank and Phone Bank) and State of the customer.

- Geopy packages is installed to estimate the state of the customer based on Latitude and longitude information

	gender	state	Annual_Sal	POS	SALES-POS	PAYMENT	INTER BANK	PHONE BANK	TtlSpend	age
0	F	WA	54112	1135.01	1267.17	1543.0	1383.0	0.0	5328.18	21
1	F	WA	66395	1023.70	871.69	4094.0	702.0	0.0	6691.39	47
2	F	WA	63935	1758.59	3224.68	1786.0	372.0	0.0	7141.27	25
3	M	WA	57339	2499.76	1459.01	2798.0	0.0	0.0	6756.77	26
4	F	WA	59420	1700.88	2875.66	897.0	1579.0	0.0	7052.54	20

- Feature engineering by adding up all debit transactions to create TtlSpend.

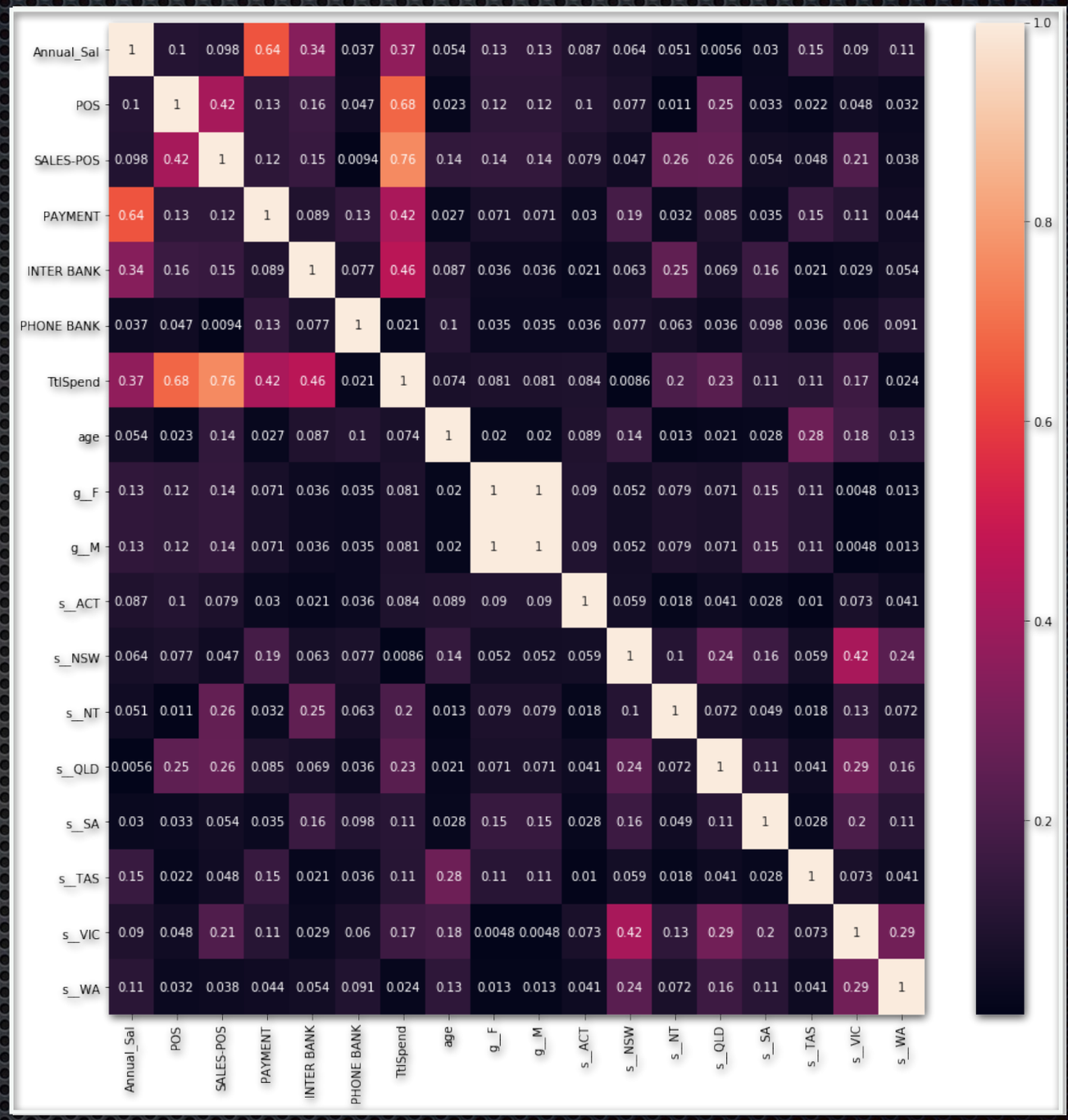
- There are 100 unique customers in this dataset. I dropped one customer as his state is likely oversea which may cause error in the state estimation from the geo coding.

- Dummy Coding carried out for Gender and State

	Annual_Sal	POS	SALES-POS	PAYMENT	INTER BANK	PHONE BANK	TtlSpend	age	g_F	g_M	s_ACT	s_NSW	s_NT	s_QLD	s_SA	s_TAS	s_VIC	s_WA
0	54112	1135.01	1267.17	1543.0	1383.0	0.0	5328.18	21	1	0	0	0	0	0	0	0	0	1
1	66395	1023.70	871.69	4094.0	702.0	0.0	6691.39	47	1	0	0	0	0	0	0	0	0	1
2	63935	1758.59	3224.68	1786.0	372.0	0.0	7141.27	25	1	0	0	0	0	0	0	0	0	1
3	57339	2499.76	1459.01	2798.0	0.0	0.0	6756.77	26	0	1	0	0	0	0	0	0	0	1
4	59420	1700.88	2875.66	897.0	1579.0	0.0	7052.54	20	1	0	0	0	0	0	0	0	0	1

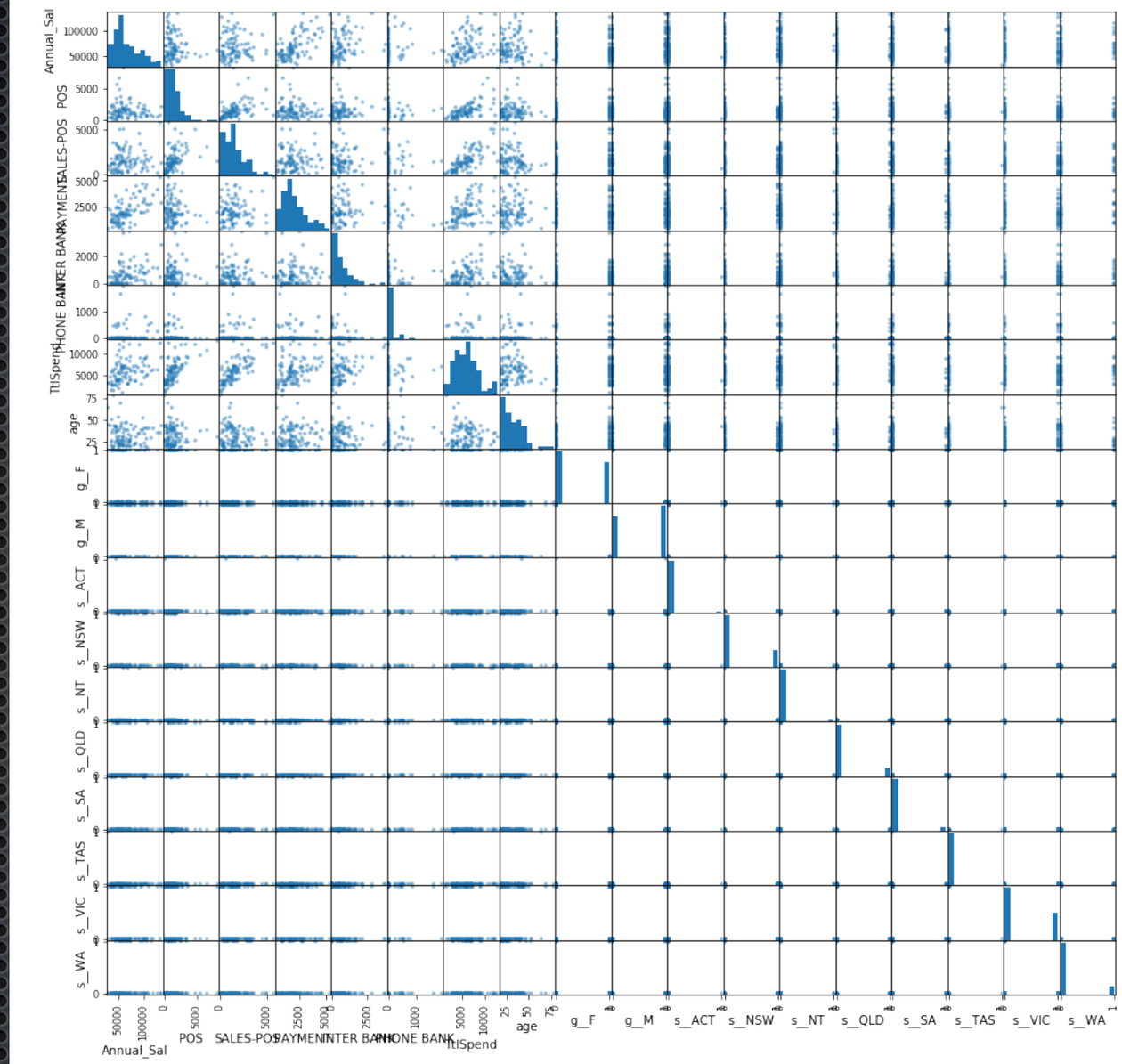
Correlation

- It is observed :
  - Some correlation between Sales Pos and Pos
  - It is expected that TtlSpend has rather significant correlation with all the debit transaction as it is featured engineered based on these transaction
  - Annual Salary has rather significant correlation with Payment



Correlation

- It is observed :
  - Some correlation between Sales Pos and Pos
  - It is expected that TtlSpend has rather significant correlation with all the debit transaction as it is featured engineered based on these transaction
- Annual Salary has rather significant correlation with Payment



Modelling

- It is rather small data (i.e. only 100) to split data model training and test. Hence, use cross validation instead rather than splitting the data into train and test data.
- Standard Scaler had been used to transform the attributes and target
- 3 models had been carried as shown in right:
- Linear Regression : Low R2 (high bias) and High variance
- Polynomial 2 degress : model no significant improvement with 2 degrees
- Lasso regression : seem slight improvement in R2 and variance as compared to Linear regression. However, the variance of R2 is vary large as compared to R2 value.

Linear Regression

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('scalar', StandardScaler(copy=True, with_mean=True, with_std=True)), ('model', LinearRegression(copy_X=True
, fit_intercept=True, n_jobs=None,
         normalize=False))]
----------------------------------------------------
cross_val_scores, R2 are
[-7.61062323e-01 -6.36846705e+27 -2.58554382e-01  5.89806736e-01
  8.40868912e-02]
[0.38252974 0.37691886 0.31258877 0.53880862 0.46454372]
cross_val_score, R2 mean     is  -6.368467048759352e+26
cross_val_score, R2 variance is  3.650163529602029e+54
----------------------------------------------------
RMSE are
[2.20988376e+04 2.01259158e+18 2.60304711e+04 1.59359792e+04
 1.44155353e+04]
[22373.9739563  13723.45748247 15234.45437577 17068.85609574
 22334.10831766]
RMSE"s mean     is  2.0125915784812192e+17
RMSE"s variance is  3.6454723755954916e+35
----------------------------------------------------
```

Polynomial, 2 Degrees

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('scalar', StandardScaler(copy=True, with_mean=True, with_std=True)), ('poly', PolynomialFeatures(degree=2,
include_bias=True, interaction_only=False)), ('model', LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False))]
----------------------------------------------------
cross_val_scores, R2 are
[-7.33263431e+18 -4.15580108e+24 -8.09589232e+24 -1.82007756e+01
 -1.24569033e+01]
[ 2.57585163e-02 -2.23741518e+24 -2.15263825e+01 -2.32298494e+01
 -1.12158957e+24]
cross_val_score, R2 mean     is  -1.561070547912322e+24
cross_val_score, R2 variance is  6.470873236703152e+48
----------------------------------------------------
RMSE are
[4.50933064e+13 5.14121272e+16 6.60204374e+16 1.09029464e+05
 5.52555876e+04]
[2.81040181e+04 2.60054594e+16 8.72096053e+04 1.23719792e+05
 3.23238880e+16]
RMSE"s mean     is  1.7580700522470336e+16
RMSE"s variance is  5.632214350242106e+32
----------------------------------------------------
```

Lasso Regression

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('scalar', StandardScaler(copy=True, with_mean=True, with_std=True)), ('poly', PolynomialFeatures(degree=2,
include_bias=True, interaction_only=False)), ('model', Lasso(alpha=2868.9574444580844, copy_X=True, fit_intercept=Tru
e,
   max_iter=1000, normalize=False, positive=False, precompute=False,
   random_state=None, selection='cyclic', tol=0.0001, warm_start=False))]
----------------------------------------------------
cross_val_scores, R2 are
[-0.4624155  -0.04977279 -0.62355976  0.51040947  0.09528671]
[0.23429819 0.29330387 0.29272746 0.67790176 0.54800974]
cross_val_score, R2 mean     is  0.1516189161443005
cross_val_score, R2 variance is  0.16313242886462467
----------------------------------------------------
RMSE are
[20138.05204132 25839.60407031 29565.15599878 17410.08919806
 14327.12738192]
[24915.2323873  14615.29530899 15452.97104676 14264.54743575
 20519.70664332]
RMSE"s mean     is  19704.778151250583
RMSE"s variance is  27062534.309228815
----------------------------------------------------
```

Feature Selection

- Feature selection with Lasso eliminated most attributes and remained 5 predictors.
- The remained attributes are Payment, Inter bank, Phone bank, Pos and Ttl Spend.
- Based on Lasso feature selection, those attributes such as age, gender and others are no significant predictors.

```
alpha = None
reg = LassoCV(eps=0.001, n_alphas=100,cv=5)
reg.fit(x, y)

cv_mse_mean = np.mean(reg.mse_path_, axis=0)
cv_rmse_mean = np.sqrt(cv_mse_mean)
cv_mse_var = np.var(reg.mse_path_, axis=0)
Best_Alpha = reg.alpha_
print("alphas: %s" % alpha)
print("CV RMSE: %s" % cv_rmse_mean)
print("CV MSE VAR %s" % cv_mse_var )
print("Best alpha using built-in LassoCV: %f" % Best_Alpha)
```

```
alphas: None
CV RMSE: [20128.22889369 19389.81750649 20594.73552723 15390.86222071
 19837.09191452]
CV MSE VAR [1.84847724e+15 8.96880506e+15 1.87167300e+15 1.76650717e+15
 3.30979980e+16]
Best alpha using built-in LassoCV: 228368.143182
```

```
coef = pd.Series(reg.coef_, index = x.columns)
print("Lasso picked " + str(sum(coef != 0)) + " variables and eliminated the other "
      +  str(sum(coef == 0)) + " variables")
```

```
Lasso picked 5 variables and eliminated the other 11 variables
```

PAYMENT	13.418750
INTER BANK	9.296358
PHONE BANK	8.200798
POS	1.430636
TtlSpend	0.249085

Final Modelling

- Models after feature selections are shown on the right.
- 3 models had been carried as shown in right:
  - Linear Regression :
    - R2 (Ave): from -6.37 to 0.35
    - R2 (Var) : from 3.65 to 0.0
  - Polynomial 2 degress :
    - R2 (Ave): from -1.56 to -0.05
    - R2 (Var) : from 6.47to 0.336
  - Polynomial 2 degress (with Ridge):
    - R2 (Ave): from 0.306
    - R2 (Var) : from 0.04

Linear Regression

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('scalar', StandardScaler(copy=True, with_mean=True, with_std=True)), ('model', LinearRegression(copy_X=True
, fit_intercept=True, n_jobs=None,
         normalize=False))]
----------------------------------------------------
cross_val_scores, R2 are
[0.0961465  0.12620884 0.20338617 0.65773651 0.10800842]
[0.37047188 0.31226731 0.42484834 0.65455803 0.56529608]
cross_val_score, R2 mean     is  0.35189280642470094
cross_val_score, R2 variance is  0.043579868030366474
----------------------------------------------------
RMSE are
[15831.82901468 23574.47451114 20709.51635884 14556.76198691
 14226.03964716]
[22591.37555584 14417.8689642  13935.09018364 14772.4107962
 20123.49399333]
RMSE"s mean     is  17473.886101195796
RMSE"s variance is  13181035.968599224
----------------------------------------------------
```

Polynomial, 2 Degrees

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('scalar', StandardScaler(copy=True, with_mean=True, with_std=True)), ('poly', PolynomialFeatures(degree=2,
include_bias=True, interaction_only=False)), ('model', Lasso(alpha=1068.886407401765, copy_X=True, fit_intercept=True
, max_iter=1000,
   normalize=False, positive=False, precompute=False, random_state=None,
   selection='cyclic', tol=0.0001, warm_start=False))]
----------------------------------------------------
cross_val_scores, R2 are
[-1.33133703 -0.12766835 -0.90561916  0.4481784  -0.14982766]
[0.26328038 0.25522634 0.21755278 0.57986144 0.20518435]
cross_val_score, R2 mean     is  -0.05451685116730519
cross_val_score, R2 variance is  0.3362621048446405
----------------------------------------------------
RMSE are
[25426.38379646 26781.12972091 32030.52751873 18483.48417761
 16151.7659466 ]
[24439.15723134 15003.87346305 16253.47095535 16291.46338738
 27210.70650687]
RMSE"s mean     is  21807.19627043155
RMSE"s variance is  32906350.506341815
----------------------------------------------------
```

Polynomial 2nd order Regulated by Ridge

```
MODEL SUMMARY
Model use :  Pipeline
Steps : [('poly', PolynomialFeatures(degree=2, include_bias=True, interaction_only=False)), ('model', RidgeCV(alphas=
array([             1,             10,            100,           1000,
                10000,         100000,        1000000,       10000000,
            100000000,     1000000000,    10000000000,   100000000000,
         1000000000000, 10000000000000]),
    cv=None, fit_intercept=True, gcv_mode=None, normalize=False,
    scoring=None, store_cv_values=True))]
----------------------------------------------------
cross_val_scores, R2 are
[ 0.12064326 -0.03440266  0.28576549  0.67181795  0.0940931 ]
[0.47078768 0.42694577 0.51402113 0.25788234 0.24862989]
cross_val_score, R2 mean     is  0.3056183959741531
cross_val_score, R2 variance is  0.04200286734014723
----------------------------------------------------
RMSE are
[15615.81354813 25649.74270503 19609.49680326 14254.16889385
 14336.57534645]
[20713.33750007 13161.00637266 12809.35568278 21652.10912425
 26456.57189905]
RMSE"s mean     is  18425.817787553002
RMSE"s variance is  23500446.891818874
----------------------------------------------------
```

Modeling - Decision Tree

- Decision Tree Regressor tried to overfit the data. It make the model more worst as R2 average become negative. RMS Error is no better than linear regression

```
MODEL SUMMARY
Model use :  DecisionTreeRegressor
Steps : no steps
------------------------------------------------
cross_val_scores, R2 are
[-0.29874703 -0.57845505  0.01989005  0.85674521 -0.24366869]
[ 0.07032819 -0.55164703 -2.6327805  -0.04131862  0.10486363]
cross_val_score, R2 mean     is  -0.3294789849016675
cross_val_score, R2 variance is  0.7385575658979794
------------------------------------------------
RMSE are
[18977.73386893 31685.04348585 22971.18685441  9417.57465593
 16797.93853721]
[27453.64024497 21656.47189641 35021.77651262 25648.12621226
 28876.94291106]
RMSE"s mean     is  23850.64351796479
RMSE"s variance is  51503545.61785303
------------------------------------------------
```

Summary

- The attributes selected are not able to produce a reasonably good model to predict the annual salary of the customers. From the R2 value, it seems that the predictors (after feature selection) can only explain 35% of the total error. \
- The models are mostly high bias and high variance and thus not good model for this predictive analysis.