Industry: Tele-Comm Sector.

Problem Statement:

Predicting the Job effort Coefficient. The time taken by the technician to install the device i.e. time between customer signature and job start time.

Data description:

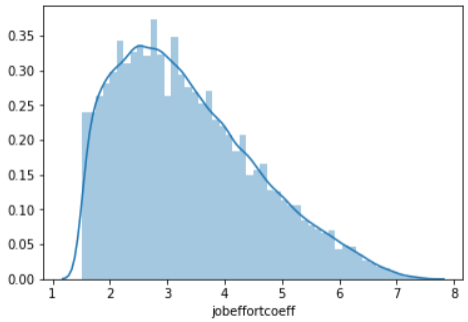
Data has 180647 rows and 38 columns with 5 Continuous variables and 33 Categorical variables.

Exploratory data analysis

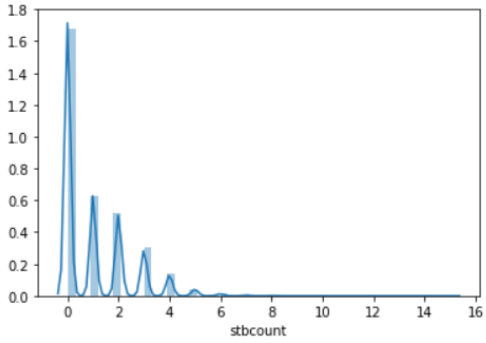
Continuous Variable's analysis

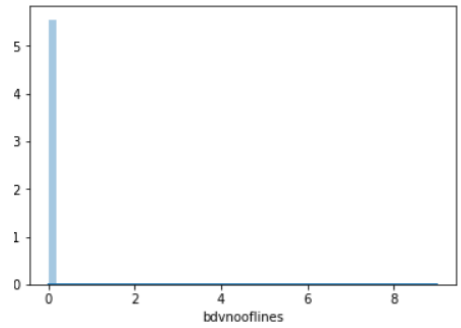
Jobeffortcoeff: Data Distribution – This is the target variable (Data looks right skewed, Applied logarithmic transformation but didn't work)

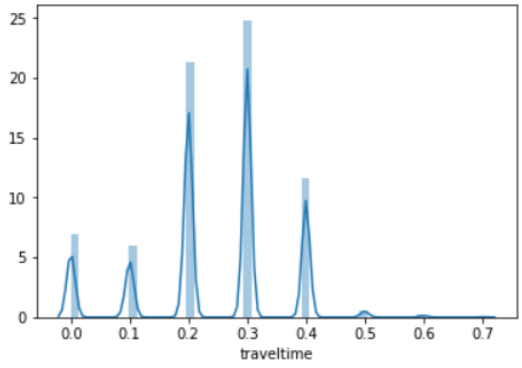
Jobeffortcoeff and diff_cx_onprem both represent same. Jobeffortcoeff is defined in decimal (0.1 = 6 mins) and diff_cx_onprem defined in minutes.



SBT Count, Travel time and BDV number of lines are continuous, but holds discrete values







Correlation matrix between continuous variables

```
# Correlation between the numerical variables
data.corr()
```

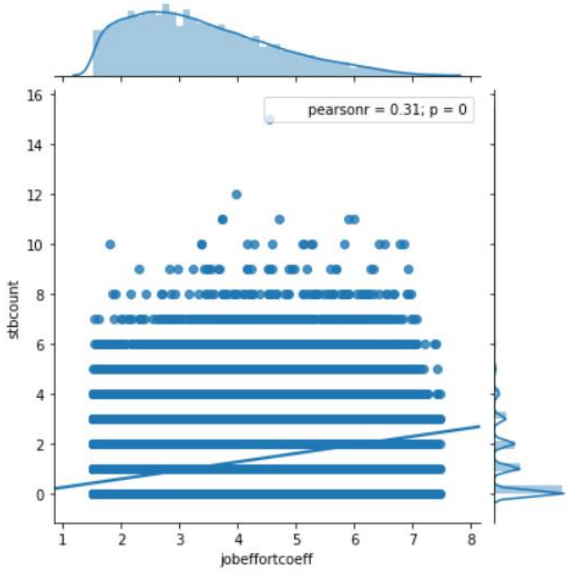
	stbcount	bdvnooflines	traveltime	diff_cx_onprem	jobeffortcoeff
stbcount	1.000000	-0.001358	0.029636	0.308754	0.308758
bdvnooflines	-0.001358	1.000000	-0.001347	0.001738	0.001733
traveltime	0.029636	-0.001347	1.000000	-0.021181	-0.021175
diff_cx_onprem	0.308754	0.001738	-0.021181	1.000000	0.999997
jobeffortcoeff	0.308758	0.001733	-0.021175	0.999997	1.000000

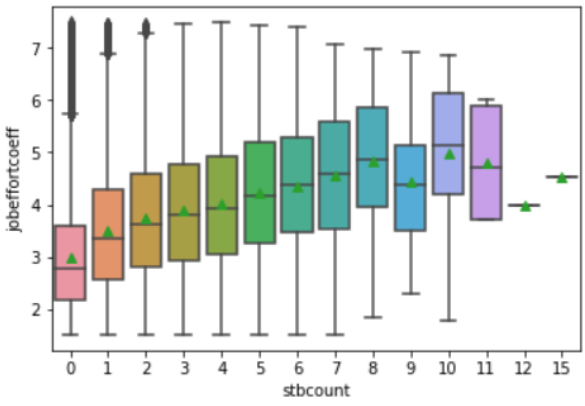
observation:

- stbcount positively correlates with target diff_cx_onprem is highly positively correlated to target variable (just a matter of unit conversion minutes to hours)

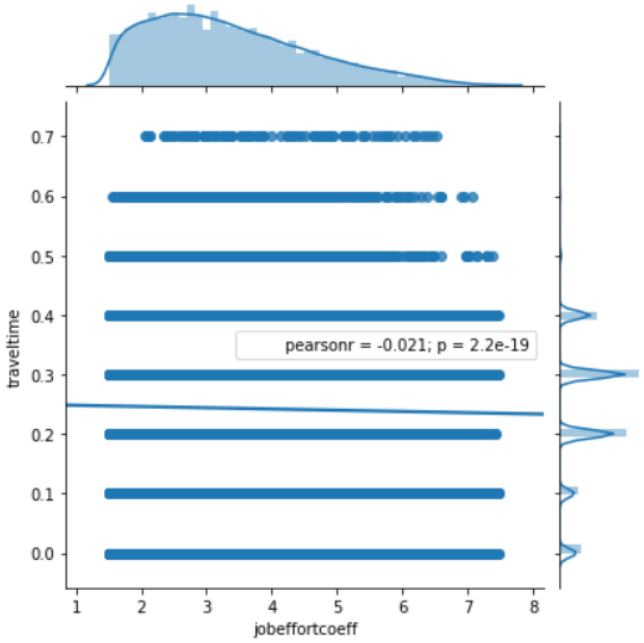
Joint distributions for continuous variables and target variable

SBT Count versus Job effort



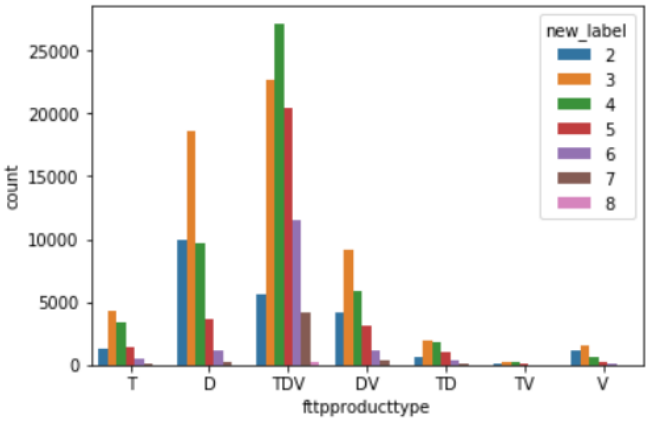


Job effort Vs Travel time

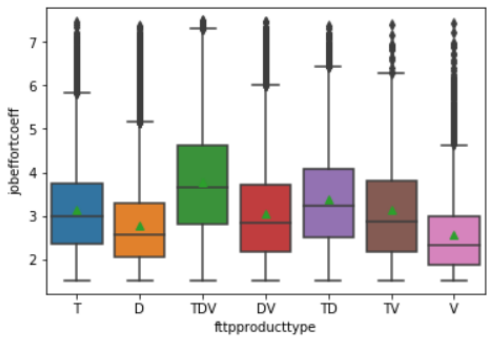


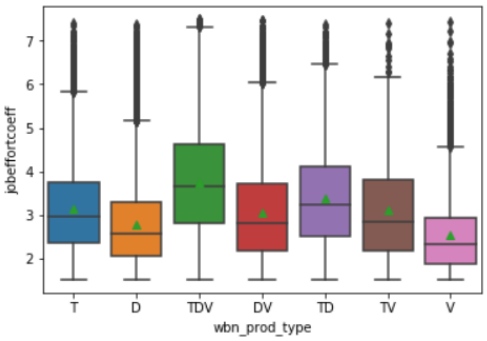
Categorical data Analysis

Label counts in each FTTP product type.



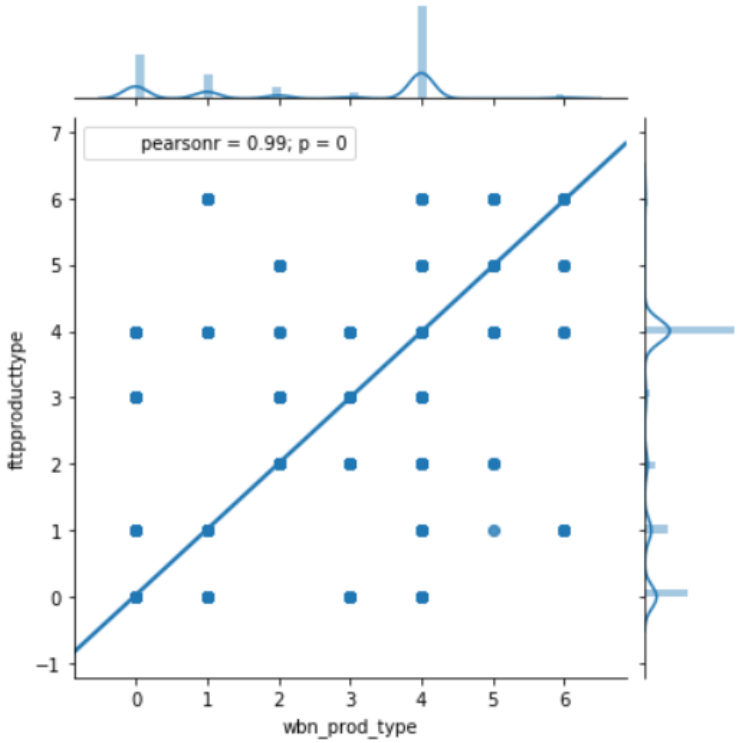
Box plot between categorical variables and target variables



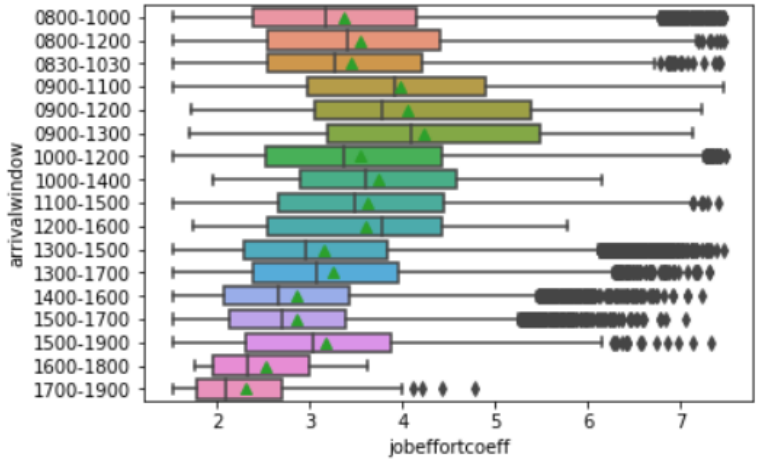


Obesrvation:

- FTTP product type and WBN_PROD_TYPE have same distribution. Only one can be included.
- Median value of jobeffortcoeff for TDV category is highest followed by TD, TV, T



Arrival window vs Jobeffortcoeff



Observation:

Seems like Delay is at peak during office hours that is at 9 am in the morning. slowly delay decreases.

Note: Similar graphs are present in the IPYNB file.

Data Preparation

Columns having null values:

1. premisetype (32081 Unknown)
2. wbn_prod_type (1 None value)
3. onttype (4 UNK)
4. stbcount (116 X)
5. dispatchreason (11544 None)
6. winbackocn (163431 None)

Replaced null values with mode. Ran the model with and without replacing Null values.
Checked for outliers. But didn't remove the outliers because of insufficient domain knowledge.

Encoding categorical variables:

Have done two types of encoding

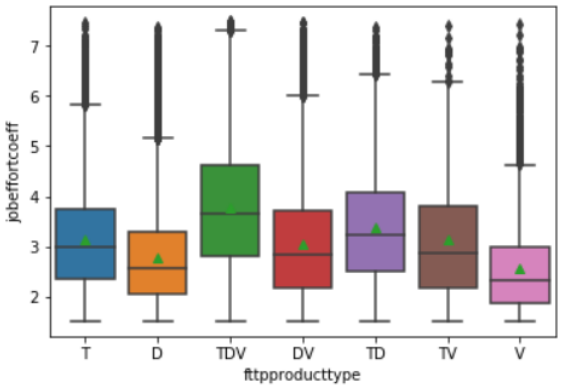
1. Numerical Encoding
2. One Hot Encoding.

Ran the model on both kind of encoded data.
Observation: Model with numerical encoded data performed slightly better.

Redundant features:

Anova test: Statistical test between categorical (Independent variable) and continuous (Dependent variable).
We conducted this test to remove all redundant variables (Non-significant in determining the target) before feeding into the model.
Conducted analysis with in both ways (Including and excluding redundant variables)
I will explain this analysis for two variables (one for significant and one for non-significant variable)

1. Variable: fttpproducttype
   Let's define NULL and Alternate hypothesis
   NULL Hypothesis:
   There is nothing going on between the variables, there is no relationship between the two variables HTTP product type and Job effort Coef. In other words, it does not matter on what types of line you add (Video,data,Television) to accurately to predict the job effort taken to install the device, the mean FTTP product type for all the different channels are same. In mathematical terms
   Mean(T) = Mean(V) = Mean(TV) = Mean(TD)
   Alternate Hypothesis:
   There is something going on between the predictor and target variable, or there is a relationship between the two. In other words type of channel installing affects the time taken to install
   Mean(T) =! Mean(V) =! Mean(TV) =! Mean(TD)



From the above plot, it is clearly visible that the mean (triangular shape in green color) of the group with **V** category of FTTP product type does not overlap with another group means.

Let's see the results.

OLS Regression Results

Dep. Variable:	jobeffortcoeff	R-squared:	0.127
Model:	OLS	Adj. R-squared:	0.127
Method:	Least Squares	F-statistic:	4390.
Date:	Wed, 27 Feb 2019	Prob (F-statistic):	0.00
Time:	17:34:35	Log-Likelihood:	-2.7902e+05
No. Observations:	180647	AIC:	5.581e+05
Df Residuals:	180640	BIC:	5.581e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2.7931	0.005	512.947	0.000	2.782	2.804
fttpproducttype[T.DV]	0.2736	0.009	29.969	0.000	0.256	0.292
fttpproducttype[T.T]	0.3562	0.012	29.559	0.000	0.333	0.380
fttpproducttype[T.TD]	0.5989	0.016	38.371	0.000	0.568	0.629
fttpproducttype[T.TDV]	0.9722	0.007	147.130	0.000	0.959	0.985
fttpproducttype[T.TV]	0.3470	0.042	8.348	0.000	0.266	0.428
fttpproducttype[T.V]	-0.2177	0.019	-11.172	0.000	-0.256	-0.179

Omnibus:	9479.104	Durbin-Watson:	1.965
Prob(Omnibus):	0.000	Jarque-Bera (JB):	11066.912
Skew:	0.606	Prob(JB):	0.00
Kurtosis:	2.989	Cond. No.	18.0

As you can see we have sufficient evidence to reject the NULL hypothesis.

F statistic value is high.

P value is low.

Hence, we can say that job effort coefficient depends on the FTTP product type.

2. Similar test ran on **hfwsindicator** variable

Dep. Variable:	jobeffortcoeff	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	105.7
Date:	Wed, 27 Feb 2019	Prob (F-statistic):	8.87e-25
Time:	17:34:38	Log-Likelihood:	-2.9126e+05
No. Observations:	180647	AIC:	5.825e+05
Df Residuals:	180645	BIC:	5.826e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.3674	0.003	1161.301	0.000	3.362	3.373
hfwsindicator[T.Y]	-0.1702	0.017	-10.279	0.000	-0.203	-0.138

Omnibus:	11273.589	Durbin-Watson:	1.820
Prob(Omnibus):	0.000	Jarque-Bera (JB):	13313.723
Skew:	0.657	Prob(JB):	0.00
Kurtosis:	2.797	Cond. No.	5.80

Low F value. So HFWS indicator doesn't add any value to the model.

Similar test ran on all the variables, and you can find it in IPYNB file.
We rejected totally 8 variables from the dataset.

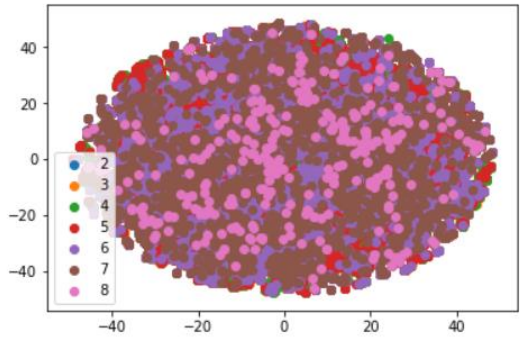
1. hfwsindicator
2. bdvind
3. swapontind
4. onttype
5. premisetype
6. migrateorderind
7. ismigrate
8. isnt
9. bdvnooflines

Data Modelling

- We have tried to solve the problem in three ways. Ran regression analysis on the continuous target variable. Achieved accuracy around 35%
  Algorithms used: Linear Regression

- Converted the problem into Classification by creating target window (0-1,2-3,3-4,4-5 etc..)
  Ran classification analysis on the target class
  Algorithms used: Logistic regression, Random forest, KNN and GBDT (Gradient boosting decision tree algorithm)
  Accuracies:
  Logistic: 39%
  Random forest: 36%
  GBDT: 42%
  KNN: 39%
  Clearly GBDT won.

- Tried to convert the problem into unsupervised learning. Plotted TNSE graph to check if there is any clustering possibility.



observation : tsne plot does not show any clusters for the data