

---

---

# Customer Retention

By Being Proactive

---

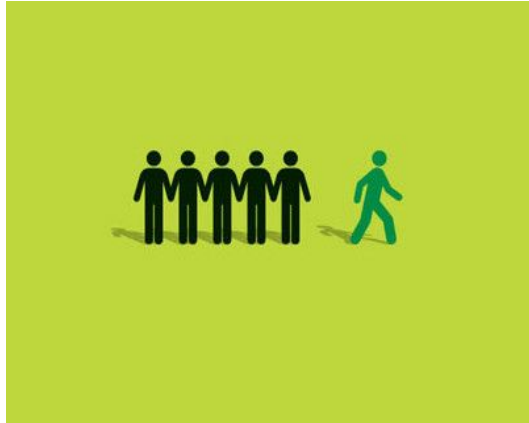
---

A Data Science Project By Aparna Shastry

# Overview

- The Business Problem
- Data Set
- Data Wrangling
- Data Stories
- Predictive Models
- Solutions Proposed
- Scope for Further Work
- References
- Appendix

# The Business Problem



- A Telecom company observes customer “Churn”
- Estimated loss of income: \$140k per month
- Attraction of new ones is more expensive than retention of current ones
- Expectations from this project: Given historical data on loyal and churn customers,
  - Understand relation between churn and certain factors
  - Provide a predictive model that ranks the customers
  - Make recommendations to the business to minimize the revenue loss.

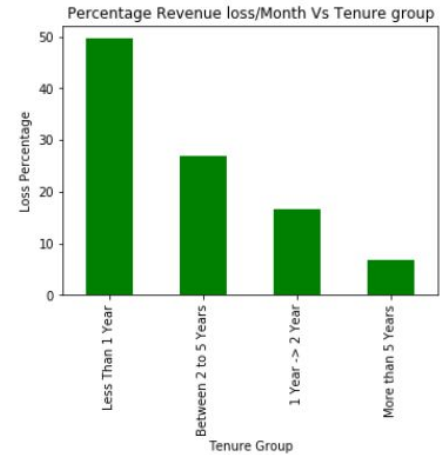
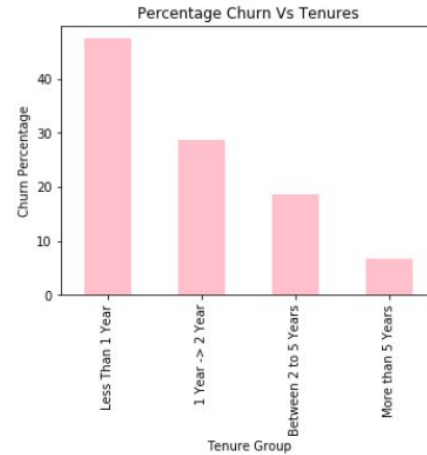
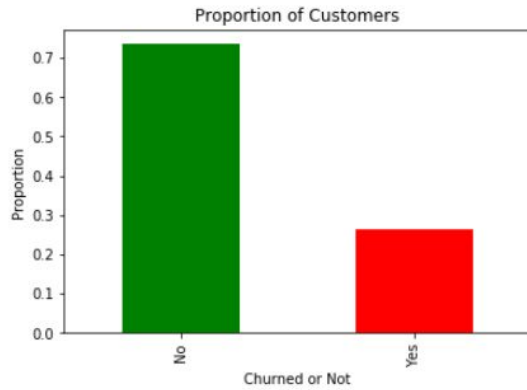
# Data Set

- Dataset link is given in Ref [1] in the last slide.
- 7043 records, 21 columns
- The target variable: Churn : { "Yes", "No" } entries.
- One column has unique customer ID
- 19 Predictor Variables, which are of following types:
- Service specific :
  - Phone: Phone Service, Multiple Lines
  - Internet: Internet Service, Online Security, Online backup, Streaming TV, Streaming Movies, Tech support, Device protection
- Person specific : Gender, Senior Citizen, Partner, Dependents, Tenure (num)
- Money specific: Monthly Charges (num), Total Charges (num), Contract, Paperless billing, Payment Method
- \*(num) above indicates numerical. Rest are all categorical: 2 to 4 categories

# Data Wrangling

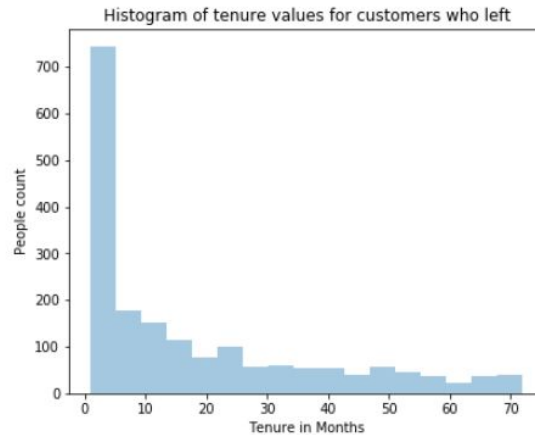
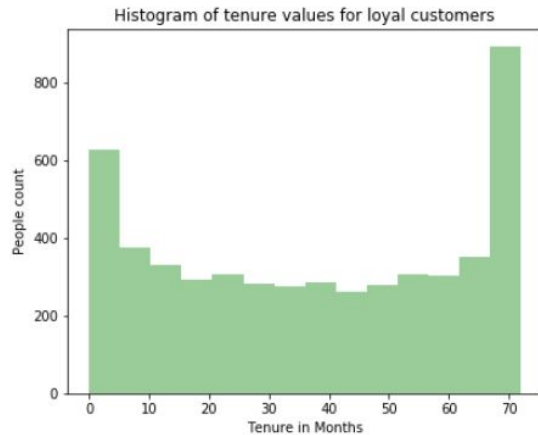
- Only Total Charges has 11 missing entries.
  - All of them are in rows with Churn “No”
  - tenure values are 0 for them
  - No other information is given.
  - Concluded that these should be for customers just registered
- Cleaning:
  - Set Total Charges to 0
  - Converted all other entries from string to float
- Observed :  $\text{Total Charges} \approx \text{tenure} * \text{Monthly Charges}$
- Verified: By fitting a simple linear regression through  $\text{Tenure} * \text{Monthly Charges}$   
[Refer Appendix [Slide 1](#) for statmodels OLS summary]
- Total Charges is hence redundant

# Business Problem visualization



- 26.54% Churn rate overall
- \$139130 loss per month, about 30% of the total income
- Less than a year tenure category has highest churn and result in highest revenue loss
- 2 to 5 year tenure category results in highest revenue loss

# Data Stories (1) : Tenure in Months



- Fig 1: Drop from bin 70 to bin 60 => Huge churn happened about 5.5 years ago.
- Fig 1: Between 10 to 60 months, not much variation => Once they cross a year, they remain loyal
- Fig 2: High count in first bin, drastic drop=> Most in churn group < 5 months tenure

# Data Stories (2) Monthly Charges Vs Services

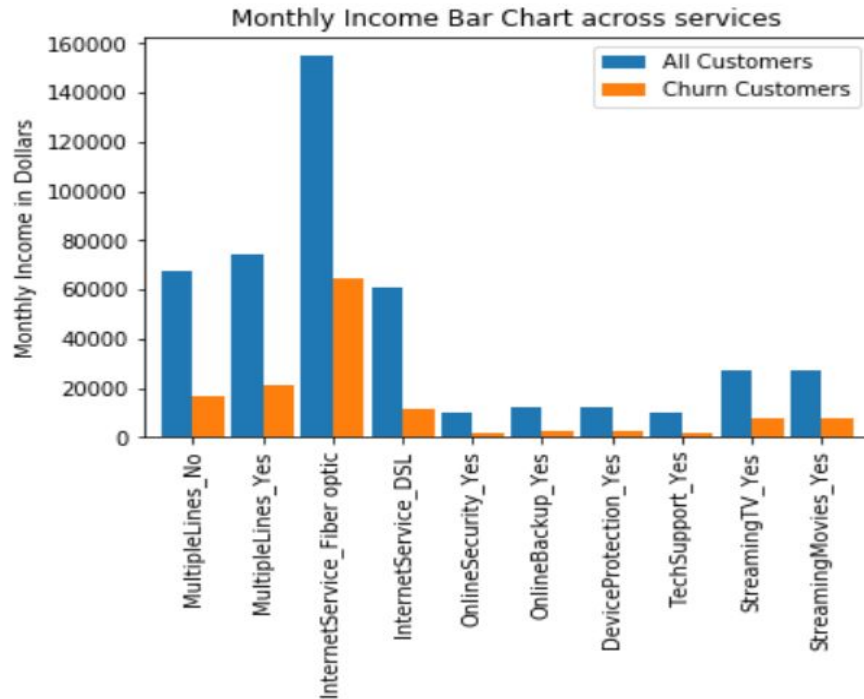
- Rate sheet is not available, hence linear regression was done to know the rate per each service. The curve was a great fit. Intercept 0 indicated no fixed monthly charges

Type of Service	Charges
Phone Line Single	\$20
Phone Line Multiple	\$25
Online Security,Online Backup,Tech Support,Device Protection	\$5 per service
Streaming Movies/ Streaming TV	\$10 per service
Internet DSL	\$25
Internet Fiber Optic	\$50

- See Appendix [Slide 2](#) for validation of regression assumptions

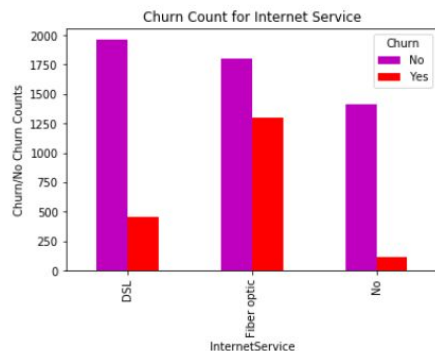
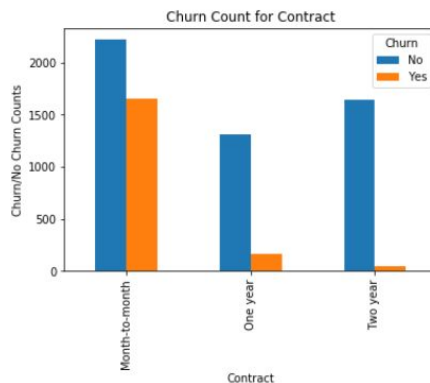
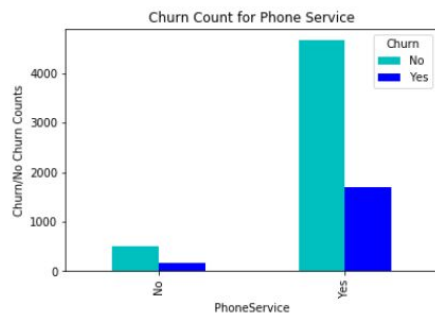


# Monthly Charges Split by Services



- The height of the bar = count of people subscribed \* Charge as per rate table of previous slide
- Blue bars are from the entire 7043 rows
- Orange bars are from the Churn group
- The importance of Fiber Optic Service on the Monthly Income is very clear!
- MultipleLines\_No essentially means Phone Line Single

# Data Stories (3) Categorical Data



```
Mean Churn Across Contract
Month-to-month    0.427097
One year          0.112695
Two year          0.028319
Name: Ch10, dtype: float64
Mean Churn Across PhoneService
No                0.249267
Yes               0.267096
Name: Ch10, dtype: float64
Mean Churn Across InternetService
DSL              0.189591
Fiber optic      0.418928
No               0.074050
Name: Ch10, dtype: float64
```

- The bar charts show counts in each category of these variables
- Mean rate of churn shown
- Plots can show individual variations
- The effect of several combinations, and also specific types of services within internet services can be understood only by modeling

# Predictive Modeling

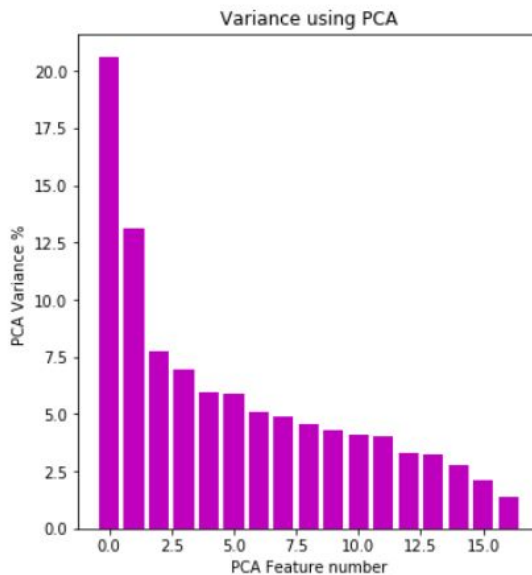
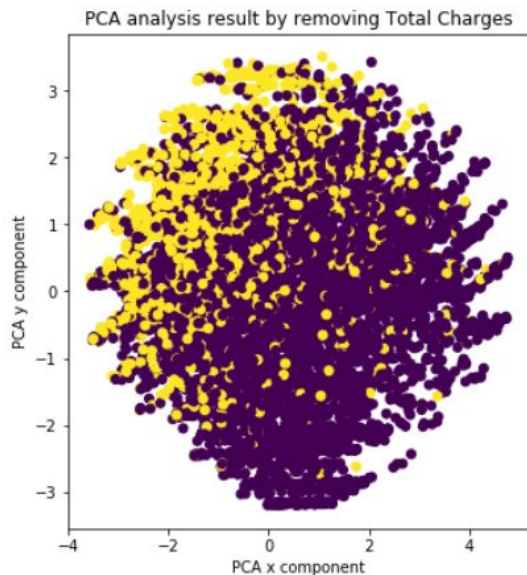
- Acceptance Criteria:
  - Mainly Area Under ROC Curve (AUC)
  - Train and Test Accuracy approx equal or not.
  - “Recall” on the Churn class to be higher priority than maximizing accuracy
- Data Prep/Feature Selection :
  - Drop numerical variable Total Charges, and Monthly Charges
  - Drop the column Phone Service, as it is a subset of Multiple Lines

For logistic regression: (and previously for Monthly Charges Linear Regression)

- Convert to dummies, and drop first, drop originals
- “No internet service” dummy in some services are dropped, correlated with Internet Services (See [Appendix Slide 3](#) for correlations)

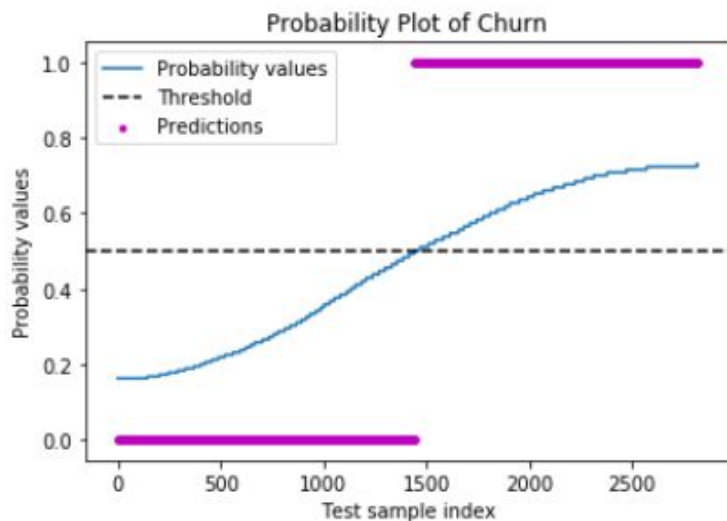
# Principal Component Analysis (PCA)

- First plot: 2-D approx helps in understanding whether classes could be separable
- Second plot: Shows the explained variance. “Elbow” happens at 3. But all of them have good variance.



# Logistic Regression with Tenure alone

- Fitted a simple logistic regression of Tenure to find out the relation
- Between 9 and 60 months, probability curve is almost linear
- Threshold Tenure = 27 months. Samples above black dotted lines correspond to  $< 27$



Coefficient:-0.037, Intercept:0.997

Train Set Accuracy :63.67%

Dev Set Accuracy 64.16%

Report:

	precision	recall	f1-score	support
0	0.87	0.60	0.71	2070
1	0.40	0.74	0.52	748
avg / total	0.74	0.64	0.66	2818

# Logistic Regression with many

- Trained using 60% of the records
- Membership from both classes in same proportion as original set
- Class weight Balancing done
- C should be kept very high to prevent regularization

How good is this model?

- Can accurately classify 74 to 75% overall
- Among every 100 churn predictions, 51 could be truly churn customers
- For every 100 churn customers, it could identify 80 correctly
- For any new data, this is expected to predict about 41% as Churn (i.e.  $0.8 * 0.51$ )
- AUC is 0.84, indicating good fit.

Confusion Matrix	Pred 0	Pred 1
True 0	1497	573
True 1	146	602

Report:

	precision	recall	f1-score	support
0	0.91	0.72	0.81	2070
1	0.51	0.80	0.63	748
avg / total	0.81	0.74	0.76	2818

# Decision Tree Classifier

- Most or categorical features, hence tried decision trees.
- First three conditions same as previous slide
- `min_weight_fraction_leaf = 0.03`

How good is this model?

- Training accuracy is always 75.5+%
- 53% of Churn predictions would be true
- For every 100 churn customers, this would identify 75 correctly
- AUC score is 0.83
- Recall is not as good as logistic regression!

Confusion Matrix	Pred 0	Pred 1
True 0	1577	493
True 1	186	562

Report:

	precision	recall	f1-score	support
0	0.89	0.76	0.82	2070
1	0.53	0.75	0.62	748
avg / total	0.80	0.76	0.77	2818

# Random Forest Classifier

- Sometimes ensemble methods could give better fits than simple classifiers. Hence tried.
- First three points same as log reg slide
- Number of estimators: 25
- 'Entropy' index
- `min_weight_fraction_leaf = 0.03`
- 4 features were used at each tree.

<b>Confusion Matrix</b>	Pred 0	Pred 1
True 0	1456	614
True 1	149	599

How good is this model?

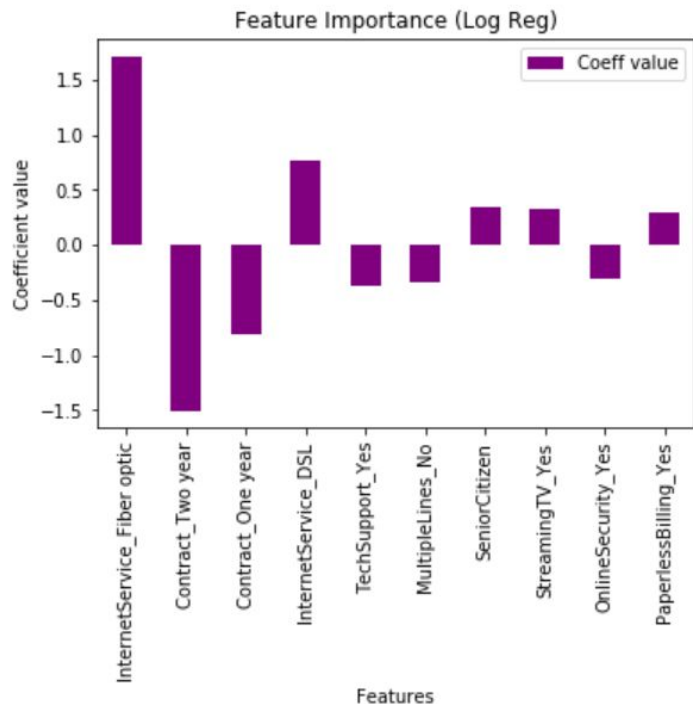
- Overall accuracy slightly lesser at 73-74%
- AUC is 0.83
- Logistic Regression wins!

Report:

	precision	recall	f1-score	support
0	0.91	0.70	0.79	2070
1	0.49	0.80	0.61	748
avg / total	0.80	0.73	0.74	2818



# Feature Importance



Top 10 Features with their influence shown

- Tenure is second last among 21. (not shown).
- PCA indicated tenure is most important, but tenure is related to the rest, although not linearly.
- Interpretation:
  - Features with negative coefficients are favorable.

# Solutions Proposed

- A predictive model is given that ranks customers based on their probability of churn and the revenue that they bring.
- Use this model to prioritize whose concerns to be addressed first. Sometimes it might be case by case basis.
- Take the following actions immediately:
  - Try striking a longer contract with new customers: two year or one year in that order of preference.
  - Leverage the time to improve the quality of services, of the high cost ones like Fiber optic.
  - Improve on the Technical support on all services like streaming, phone connection and internet.
  - Be up-to-date with current technology.
  - Collect customer feedback and act on it immediately to prevent new customer churn
- Next: It will be helpful to understand why churn started 5.5 years ago. Give more historical data to the data scientist for analysis.

# Potential Money Savings!

- Assuming the business makes an attempt to convince all the customers identified as Churn (likelihood  $\geq 0.5$ ),
- The income loss prevented using the given dataset: \$115,480 / month if all of them change their mind.
- The above statement has not accounted for the money spent in efforts to retain them, because there is not enough information. Hence the net income will be less than \$115,480.

# Scope for Further Work

- Collect more data through surveys, analyze them using NLP techniques.
- Collect more historical data on customer churn. Especially it might be useful to see what happened 5+ years ago as seen from tenure plot.

# References/Links

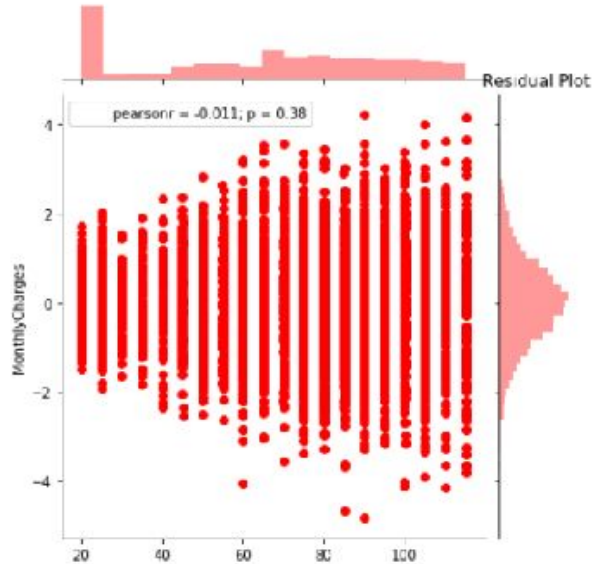
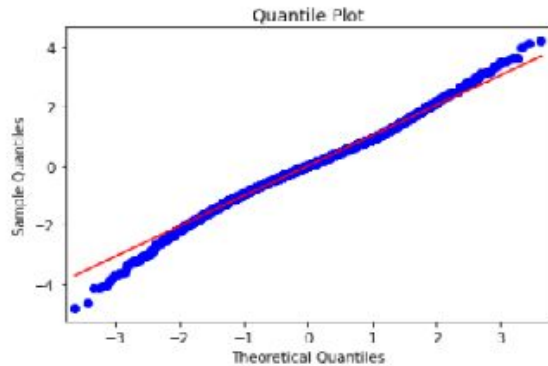
1. [IBM page](#) has many, out of that this [Telecom](#) dataset
2. [Introduction to Statistical Learning](#) by Gareth James et. al
3. [Predicting Customer Churn using R](#) by Susan Li
4. [My Code](#)

# Appendix Slide 1: Goodness of fit for Total Charges

## OLS Regression Results

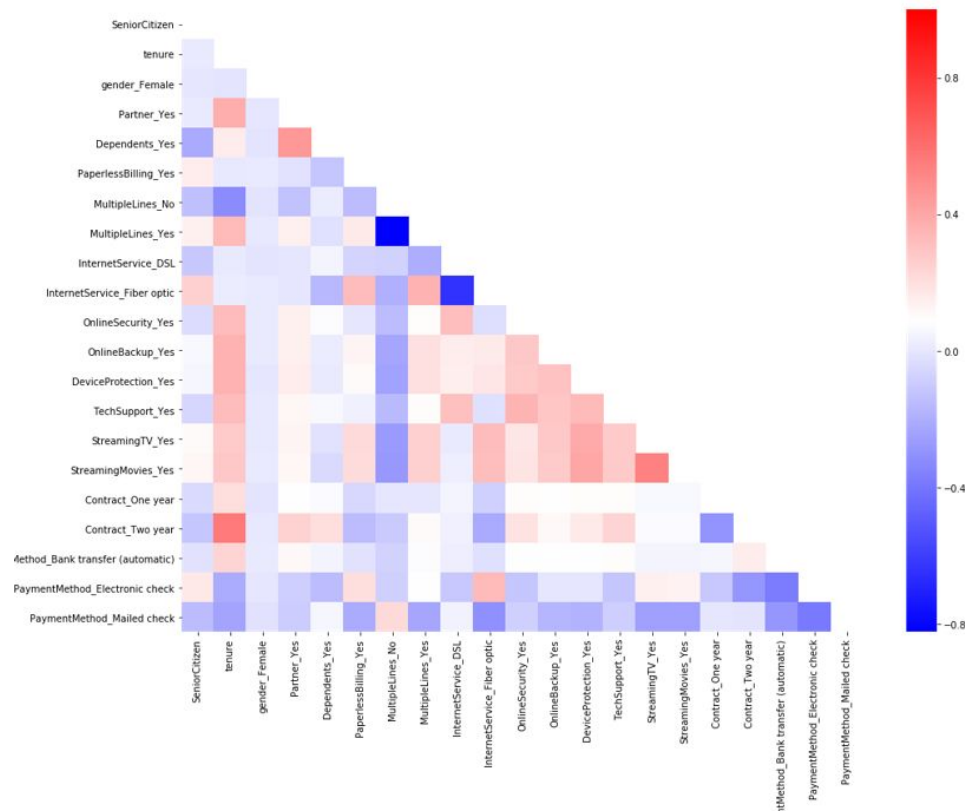
Dep. Variable:	TotalCharges	R-squared:	0.999			
Model:	OLS	Adj. R-squared:	0.999			
Method:	Least Squares	F-statistic:	8.006e+06			
Date:	Mon, 12 Feb 2018	Prob (F-statistic):	0.00			
Time:	13:31:50	Log-Likelihood:	-39627.			
No. Observations:	7043	AIC:	7.926e+04			
Df Residuals:	7041	BIC:	7.927e+04			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-0.9229	1.136	-0.812	0.417	-3.150	1.304
Temp	1.0005	0.000	2829.477	0.000	1.000	1.001
Omnibus:	538.795	Durbin-Watson:	2.055			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3067.278			
Skew:	-0.034	Prob(JB):	0.00			
Kurtosis:	6.232	Cond. No.	4.56e+03			

# Slide 2: Quantile and Residual Plots of Monthly Charges



1. Assumptions verified as seen by plots
2.  $R^2 = 0.999$
3. MSE of the fit = 1.05
4. Percentage of outliers = 3.96

# Slide 3 Correlation plot for Log Reg inputs



There are hardly any highly correlated features