Capstone Final Report

Predicting Customer Lifetime Value

This project aims to develop a machine learning model to estimate a telecommunications company's customer lifetime value (CLV). CLV is an important metric for identifying high-value customers. If a company knows what leads to high-value customers it can focus marketing and customer retention efforts on customers with a high CLV. This can be difficult when the customer has to make decisions from a list of options like the services offered by a telco. A customer may feel stuck with their services if switching is difficult, and this can lead to churn. A focus on customer's with high CLV enables the businesse to allocate resources more efficiently, improve customer satisfaction, and ultimately increase profitability. Also, predicting CLV allows companies to anticipate future revenue and make informed decisions about pricing, promotions, and service offerings

Data Wrangling

This dataset was sourced from Kaggle.com, it had 7,043 unique rows. Each row has a different customer and accompanying service selections with the telecommunications provider. Originally the dataset had twenty-one columns, which were the services and charges for each customer. The dataset was relatively simple, it did not have timestamps or dates associated with purchases. The transaction type was a monthly account fee and the duration of the business relationship was given as tenure.

Data wrangling for this dataset required only two steps. Convert the 'TotalCharges' column to a numeric data type, and change eleven null values to zero. These eleven rows were customers with tenure values of zero, meaning these customers did not have more than one month of tenure with the company. These changes were made to keep as much data as possible and make model selection easier.

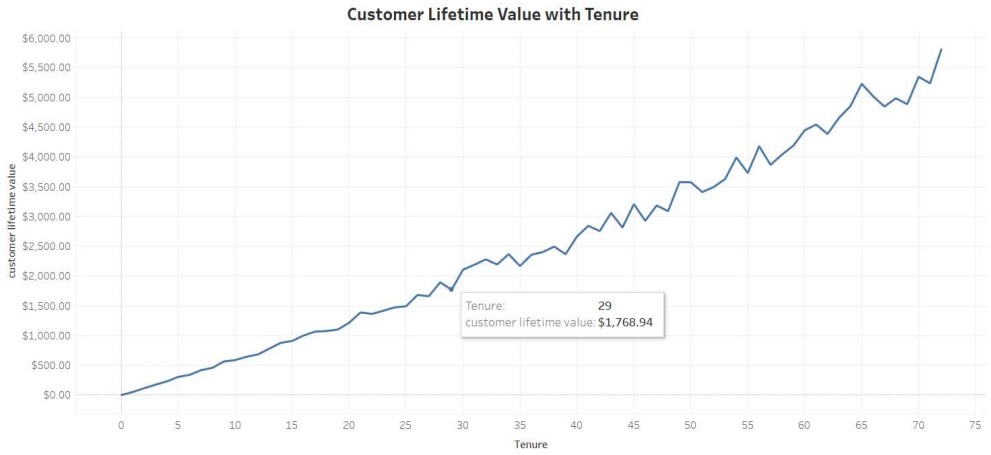
Features

- Gender
- Senior citizen
- Partner
- Dependents
- Tenure
- Phone service
- Multiple phone lines
- Internet service
- Online security
- Online backup
- Device protection
- Tech support
- Streaming TV
- Streaming movies
- Contract
- Paperless billing
- Payment method
- Monthly charges
- Total charges
- Churn
- Tenure group

Feature engineering

This dataset had 21 features before wrangling and preprocessing. Of these features, 18 were categorical with two to four unique values each. This was helpful when it came time to encode them for the machine-learning model.

A new feature was added to further describe the customer's relationship with the company. This new feature gave the customer one of three labels based on their tenure, 'new', 'medium_term', and 'long_term'. New customers had a tenure of less than 24 months, medium-term customers had a tenure of 24 to 48 months, and long-term customers had a tenure greater than 48 months. This feature will segregate customers based on tenure and give the model another data point to link to when predicting CLV.



A line plot showing calculated CLV against customer tenure. There is a trend with higher tenure and higher CLV.

The target feature is the customer's lifetime value (CLV), this was added by calculating the average monthly charges multiplied by the customer's tenure. A CLV feature was added to give the model a true value to test. This number was calculated by averaging the customer's monthly charges and multiplying this by the tenure. A stipulation was added when the customer had churned; their CLV was their total charges, while not-churned customers retained the calculated CLV.

Preprocessing and Modeling

Three models were built to predict the CLV; Gradient boost regressor, adaboost regressor, and a random forest regressor. The Gradient Boost model performed best. A gradient-boosting model is a decision-tree-based model. It combines weak learners sequentially so that each new tree corrects the errors of the previous. The models were scored using the mean squared error method and root mean squared error method.

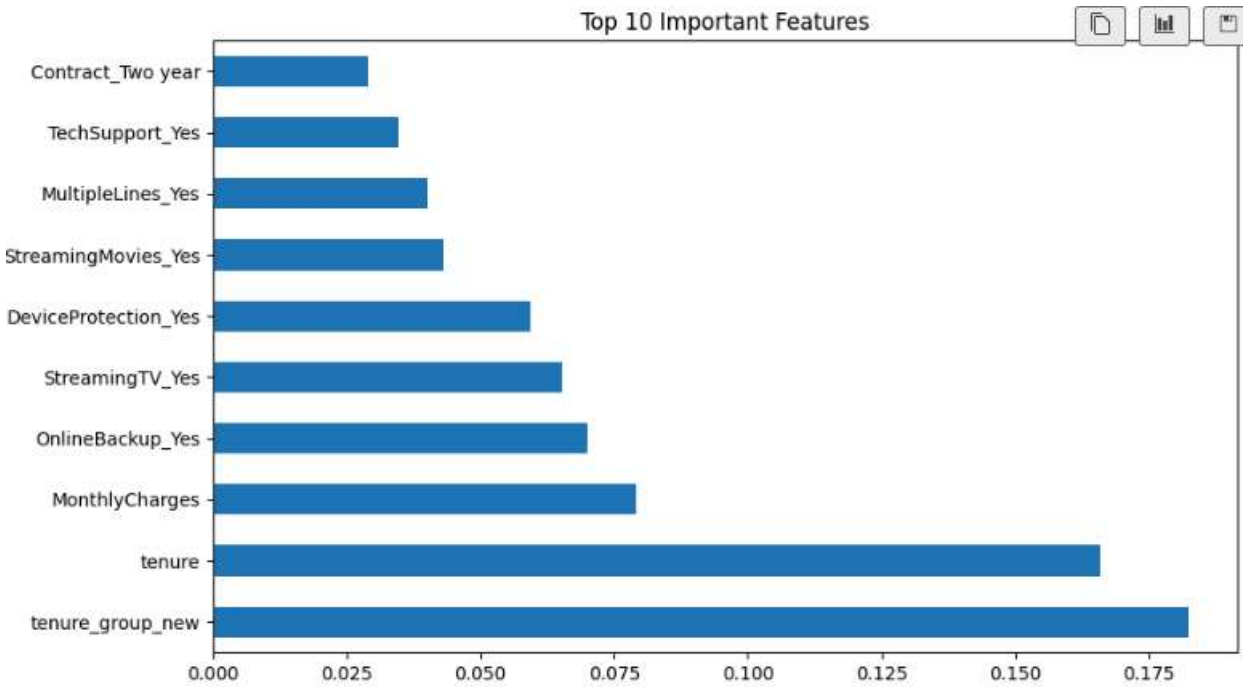
	Gradient Boosting	Random Forest	AdaBoost
Mean Absolute Error	176.82	235.91	213.48
Root Mean Squared Error	257.83	319.89	257.15

A table of each model's scoring results

Hyperparameter tuning

GridSearchCV is the method I chose to implement the hyperparameter tuning. It automates the process of finding the optimal set of hyperparameters for a given model by performing an exhaustive search over a specified parameter grid. This method uses cross-validation splitting the dataset into multiple subsets and trains the models on these to ensure good performance evaluation.

During the modeling phase, these were the features the model picked to be the most important. This means, that to get as close to true values as possible it used these features when finding the most accurate customer lifetime value.



A bar chart showing the ten most important features as designated by the model.

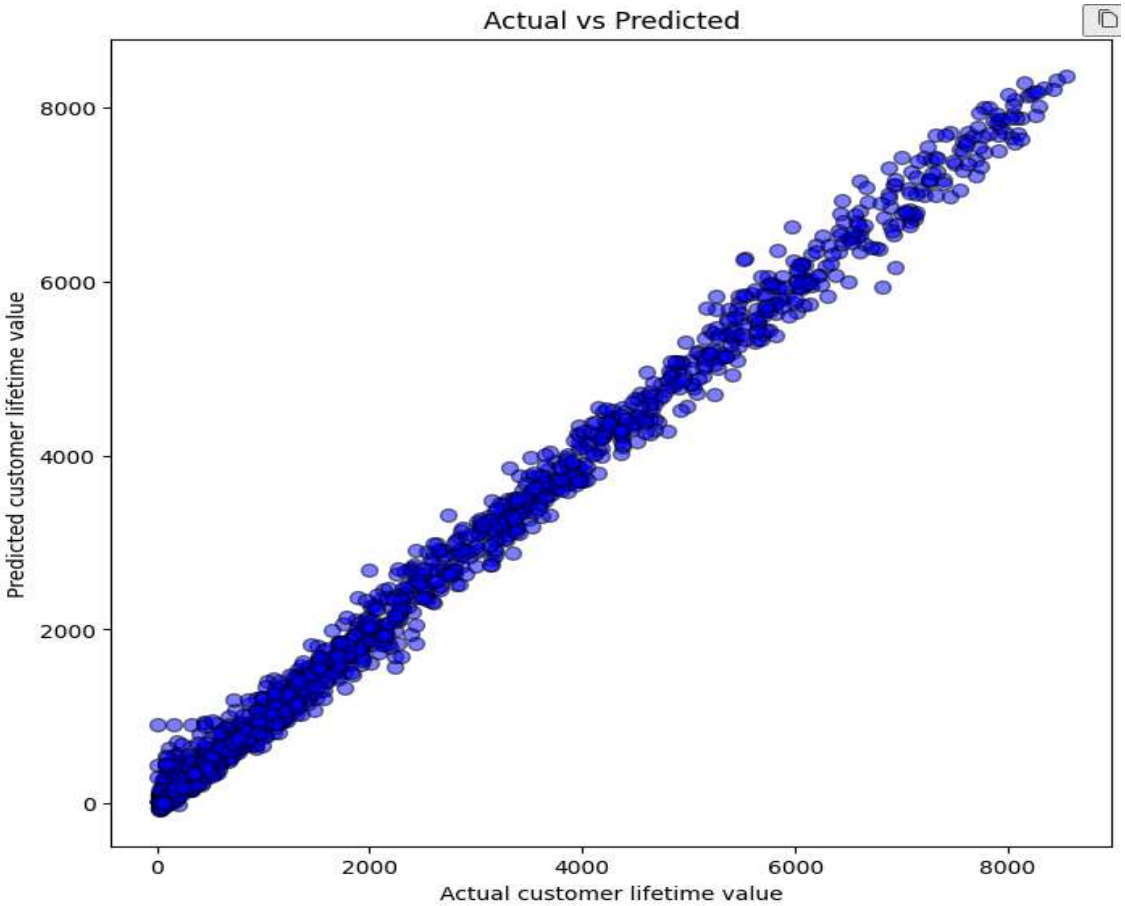
Another way to say this is by using these different options the customer has selected themselves to pay or not to pay for every month, we can most accurately predict their CLV.

Hyperparameter tuning

GridSearchCV is the method I chose to implement the hyperparameter tuning. It automates the process of finding the optimal set of hyperparameters for a given model by performing an exhaustive search over a specified parameter grid. This method uses cross-validation splitting the dataset into multiple subsets and trains the models on these to ensure good performance evaluation.

Predictions

The scatter plot below shows the results of the model's predictions plotted against the actual calculated values. If the model was 100% accurate, the dots would be a straight line along the line 'y=x'. To avoid overfitting the models, each model had relatively similar parameters. I also kept the parameter lists short to keep model training times below 10 minutes.



A scatter plot showing the actual (test) values against the predicted CLV values.

Results

To predict a high CLV for a customer the model used tenure as its most valued feature. The customers with the highest CLV also had the maximum measured tenure of 72 months. The list of features below is also a good indicator of a high CLV. A customer that paid for these service, had higher than average monthly charges and a two year contract was likely to be a high value customer. This list of features is what should be focused on when appealing to customers who are at risk of churn.

- tech support
- streaming movies
- streaming TV
- device protection
- online backup
- contract
- monthly charges

customer_lifetime_value	predicted	difference	tenure	TechSupport_Yes	StreamingMovies_Yes	StreamingTV_Yes	DeviceProtection_Yes	OnlineBackup_Yes	Contract_Two year	MonthlyCharges	Churn_Yes
8460.00	8222.457025	237.542975	72	1	1	1	1	1	1	117.50	0
8334.00	8207.517443	126.482557	72	1	1	1	1	1	1	115.75	0
8211.60	8214.989727	-3.389727	72	1	1	1	1	1	1	114.05	0
8182.80	8073.482987	109.317013	72	1	1	1	1	1	1	113.65	0
8128.80	7821.411273	307.388727	72	1	1	1	1	1	0	112.90	0
8118.00	8060.976889	57.023111	72	1	1	1	1	1	1	112.75	0
8092.80	7649.012200	443.787800	72	0	1	1	1	1	1	112.40	0
8082.00	7968.012513	113.987487	72	1	1	1	1	1	1	112.25	0
8060.50	8090.092273	-29.592273	70	1	1	1	1	1	1	115.15	0
8060.40	7469.993979	590.406021	72	1	1	1	1	1	1	111.95	0

This is a screenshot of a dataframe from the sample taken to predict CLV. It shows the recommendations above are valid. A ‘1’ in the dataframe is a ‘yes’ value and ‘0’ is a ‘no’ value.