Project 03: Lending Club Loan Data Analysis

By: Shubham Chaudhary, Kirun Haque, Michelle Baginski

Problem Selection

The regression problem statement is to predict the interest rate on a loan with respect to the demographics provided for that customer. The classification problem statement is whether the customer will be likely to fully paid off their loan or charged off.

Data Collection

Our dataset contains data from the Lending Club from a two year time-span. Prior to cleaning the dataset, it contained 144 columns with about 188,181 observations. This dataset provides financial information and a credit profile per each individual who has received a loan.

Link to the original data: https://www.kaggle.com/wendykan/lending-club-loan-data

Data Preparation/Data Cleaning

To prepare our data, we first began by removing all the columns that contained only NaN values or missing values. The next step we took was looking for columns where all cells contained the exact same values, and dropping those from our dataset. We also removed all the NA values from each observation. For each column that had missing values, we calculated the percentage of missing values. For columns that has less than 20% of the observations missing, those values were imputed with the mean. For columns that had more than 20% missing, we removed those columns. Then, we stripped of the last character in the interest_rate column and converted that object type column to a float64 type. We split our dataset into two subsets, one

contained numerical variables and the other contained categorical variables. Once we had the numerical data frame, we took the correlation all of the independent variables and if the correlation value was higher than 80% we removed the column.

To handle outliers, we made box plots to see which numerical variables contained outliers. For outliers that were above Q4, we detected them using the formula Q3 + (1.5*IQR) and used np.where to detect values that were higher than the number calculated by the formula and replaced them. If a column Outliers that were below Q1 were detected using the formula Q1 - (1.5 * IQR). We again used np.where to detect values that were less than the number calculated by the formula and replaced those.

Snippet of the correlation matrix

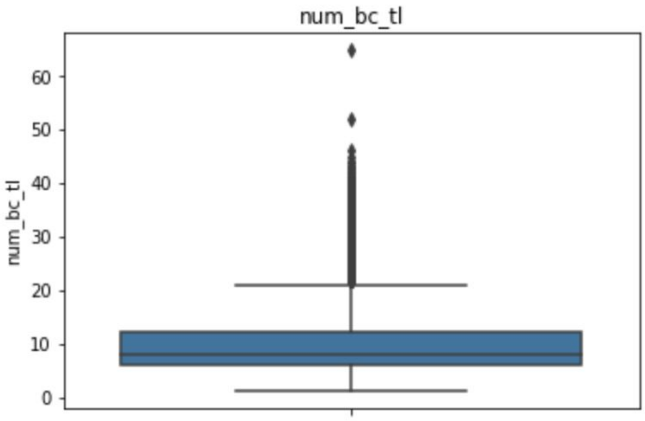
	loan_amnt	funded_amnt	funded_amnt_inv	int_rate
loan_amnt	1.000000	0.999799	0.999663	0.182654
funded_amnt	0.999799	1.000000	0.999874	0.182485
funded_amnt_inv	0.999663	0.999874	1.000000	0.182933
int_rate	0.182654	0.182485	0.182933	1.000000
installment	0.955011	0.955254	0.955211	0.165173
annual_inc	0.368164	0.368151	0.368084	-0.026026
dti	0.044557	0.044572	0.044746	0.147471
delinq_2yrs	0.011184	0.011214	0.011391	0.097230
inq_last_6mths	0.019741	0.019703	0.020091	0.241345
open_acc	0.191571	0.191614	0.191719	0.017359

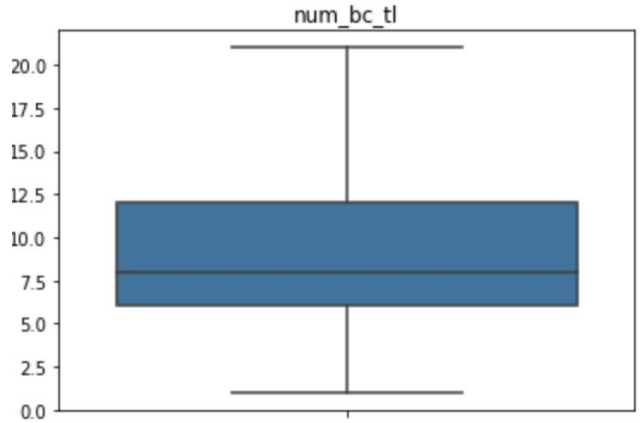
Data Exploration

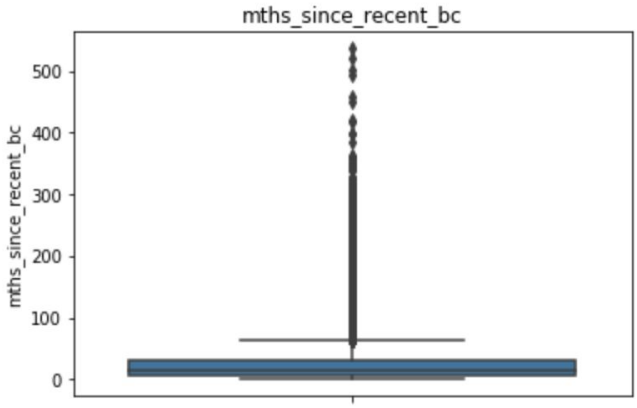
After data preparation, we did Ordinary Least Squares (OLS) Regression method. Our dependent variable was the interest_rate and the rest of the columns were the independent variables. Then whatever variable that had the p-value higher than the significance value which was .05, we removed that variable/column.

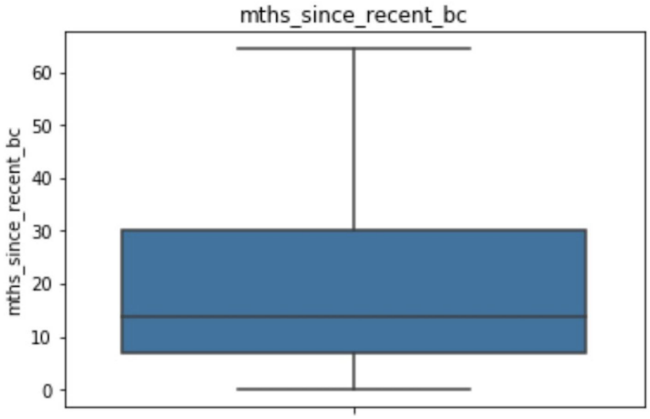
We also used boxplots to understand our dataset better. We found out that are dataset had a lot of outliers once we plotted all of the numerical columns/variables. So we made the

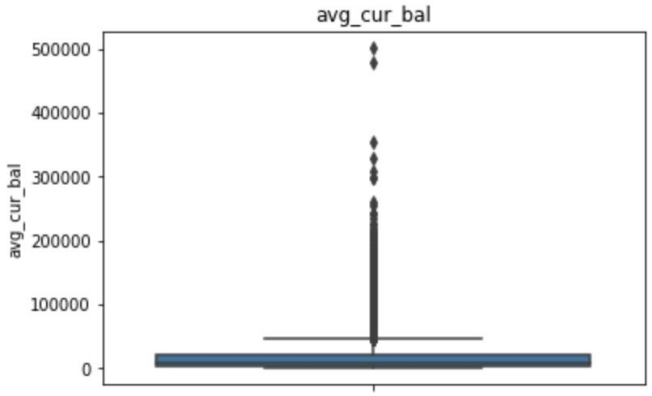
appropriate adjustments to our dataset by removing the outliers. On the following page, we have a few examples of our boxplots from the dataset.

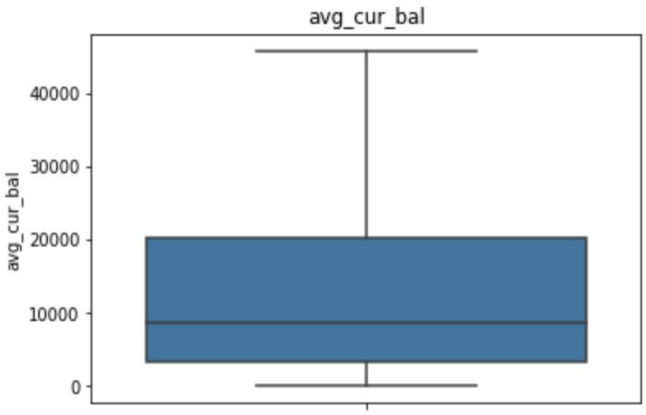


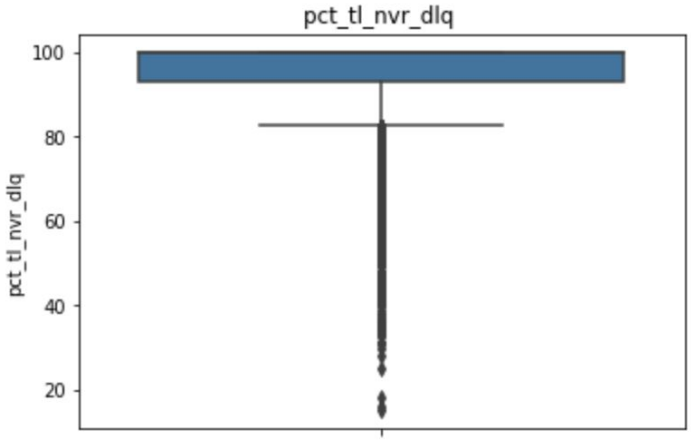


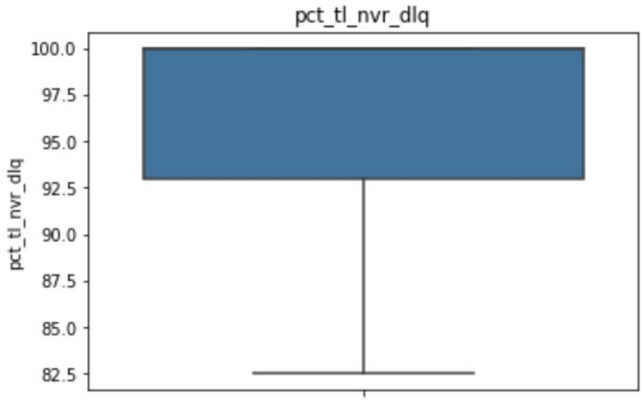




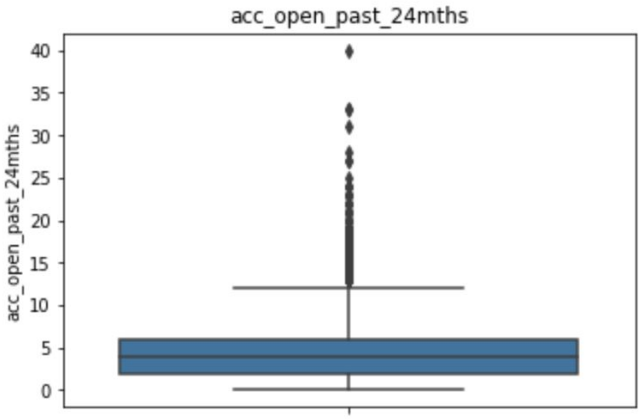


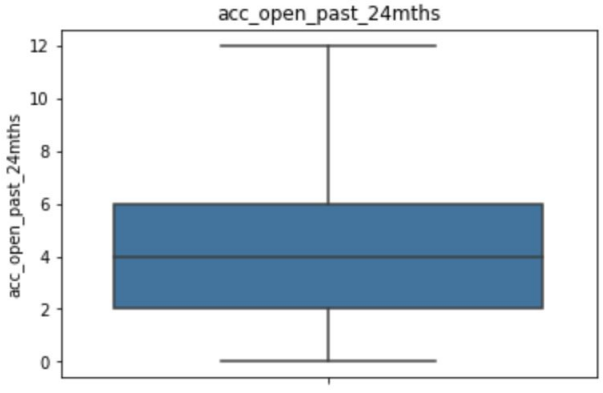






Data Modeling





Regression

To perform regression, we split our dataset into a testing and training set using the holdout method and ran linear regression on three models.

The first model was on the numerical dataset.

OLS Regression Results

Dep. Variable:	int_rate	R-squared:	0.411
Model:	OLS	Adj. R-squared:	0.411
Method:	Least Squares	F-statistic:	4103.
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	0.00
Time:	15:50:07	Log-Likelihood:	-4.9761e+05
No. Observations:	188181	AIC:	9.953e+05
Df Residuals:	188148	BIC:	9.956e+05
Df Model:	32		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	15.2796	0.157	97.379	0.000	14.972	15.587
loan_amnt	0.0001	1.25e-06	98.418	0.000	0.000	0.000
annual_inc	-1.84e-06	1.95e-07	-9.420	0.000	-2.22e-06	-1.46e-06
dti	0.0298	0.001	23.222	0.000	0.027	0.032
delinq_2yrs	0.5857	0.013	46.173	0.000	0.561	0.611
inq_last_6mths	0.7765	0.010	76.788	0.000	0.757	0.796
open_acc	-0.0755	0.003	-24.107	0.000	-0.082	-0.069
pub_rec	0.7662	0.020	39.029	0.000	0.728	0.805
revol_bal	1.292e-06	4.94e-07	2.617	0.009	3.24e-07	2.26e-06
total_acc	-0.0070	0.002	-3.658	0.000	-0.011	-0.003
total_rec_late_fee	0.0215	0.001	21.349	0.000	0.020	0.024
recoveries	0.0006	9.65e-06	62.001	0.000	0.001	0.001
last_pymnt_amnt	7.961e-05	1.56e-06	51.042	0.000	7.66e-05	8.27e-05
tot_coll_amt	5.745e-05	9.98e-06	5.756	0.000	3.79e-05	7.7e-05
acc_open_past_24mths	0.1088	0.005	23.913	0.000	0.100	0.118
avg_cur_bal	-3.857e-05	6.55e-07	-58.900	0.000	-3.99e-05	-3.73e-05
bc_open_to_buy	-4.477e-05	8.68e-07	-51.554	0.000	-4.65e-05	-4.31e-05
bc_util	0.0514	0.000	121.769	0.000	0.051	0.052
mo_sin_old_il_acct	-0.0020	0.000	-10.521	0.000	-0.002	-0.002
mo_sin_old_rev_tl_op	-0.0035	0.000	-32.054	0.000	-0.004	-0.003
mo_sin_rcnt_rev_tl_op	0.0046	0.001	5.728	0.000	0.003	0.006
mo_sin_rcnt_tl	-0.0089	0.001	-7.045	0.000	-0.011	-0.006
mort_acc	-0.1508	0.005	-30.466	0.000	-0.160	-0.141
mths_since_recent_bc	-0.0025	0.000	-7.149	0.000	-0.003	-0.002
mths_since_recent_inq	-0.0619	0.002	-32.650	0.000	-0.066	-0.058
num_accts_ever_120_pd	0.1198	0.011	10.891	0.000	0.098	0.141
num_actv_bc_tl	-0.0873	0.008	-10.734	0.000	-0.103	-0.071
num_actv_rev_tl	0.1780	0.006	30.415	0.000	0.166	0.189
num_bc_tl	-0.0701	0.003	-21.982	0.000	-0.076	-0.064
num_il_tl	-0.0070	0.002	-2.831	0.005	-0.012	-0.002
num_tl_op_past_12m	0.4447	0.008	54.304	0.000	0.429	0.461
pct_tl_nvr_dlq	-0.0575	0.002	-37.437	0.000	-0.060	-0.054
total_il_high_credit_limit	-3.497e-06	3.33e-07	-10.510	0.000	-4.15e-06	-2.85e-06

Omnibus:	4803.670	Durbin-Watson:	1.947
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5550.412
Skew:	0.355	Prob(JB):	0.00
Kurtosis:	3.451	Cond. No.	2.00e+06

After performing regression on only the numerical dataset, the R-squared value is .411, thus, this model explains roughly 40% of the proportion of the variance of the interest_rate variable. This model had no variables that had p-values greater than the .05 significance value so nothing was removed.

The second model was run on the full dataset which we completed using dummy variables since mathematical operations could not be performed on non-numerical values.

OLS Regression Results

Dep. Variable:	int_rate	R-squared:	0.994
Model:	OLS	Adj. R-squared:	0.994
Method:	Least Squares	F-statistic:	3.634e+05
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	0.00
Time:	15:50:55	Log-Likelihood:	-69087.
No. Observations:	188181	AIC:	1.383e+05
Df Residuals:	188097	BIC:	1.392e+05
Df Model:	83		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	9.1528	0.011	803.264	0.000	9.130	9.175
loan_amnt	-1.894e-06	1.37e-07	-13.813	0.000	-2.16e-06	-1.62e-06
annual_inc	8.221e-08	2.01e-08	4.085	0.000	4.28e-08	1.22e-07
dti	0.0020	0.000	14.874	0.000	0.002	0.002
delinq_2yrs	0.0058	0.001	4.458	0.000	0.003	0.008
inq_last_6mths	0.0031	0.001	2.895	0.004	0.001	0.005
open_acc	-0.0007	0.000	-2.074	0.038	-0.001	-3.67e-05
pub_rec	0.0009	0.002	0.467	0.640	-0.003	0.005
revol_bal	-2.863e-08	5.08e-08	-0.564	0.573	-1.28e-07	7.09e-08
total_acc	0.0016	0.000	7.920	0.000	0.001	0.002
total_rec_late_fee	8.397e-05	0.000	0.809	0.419	-0.000	0.000
recoveries	5.516e-06	1.13e-06	4.881	0.000	3.3e-06	7.73e-06
last_pymnt_amnt	1.041e-06	1.67e-07	6.238	0.000	7.14e-07	1.37e-06
tot_coll_amt	7.904e-07	1.02e-06	0.772	0.440	-1.22e-06	2.8e-06
acc_open_past_24mths	-0.0018	0.000	-3.831	0.000	-0.003	-0.001
avg_cur_bal	-2.29e-07	7.07e-08	-3.238	0.001	-3.68e-07	-9.04e-08
bc_open_to_buy	-6.738e-08	9.16e-08	-0.736	0.462	-2.47e-07	1.12e-07
bc_util	0.0005	4.58e-05	11.298	0.000	0.000	0.001
mo_sin_old_il_acct	-1.784e-05	1.98e-05	-0.900	0.368	-5.67e-05	2.1e-05
mo_sin_old_rev_tl_op	-5.725e-05	1.14e-05	-5.018	0.000	-7.96e-05	-3.49e-05
mo_sin_rcnt_rev_tl_op	-1.442e-05	8.2e-05	-0.176	0.860	-0.000	0.000
mo_sin_rcnt_tl	-0.0001	0.000	-0.962	0.336	-0.000	0.000
mort_acc	-0.0029	0.001	-5.492	0.000	-0.004	-0.002
mths_since_recent_bc	-4.408e-05	3.55e-05	-1.243	0.214	-0.000	2.54e-05
mths_since_recent_inq	-0.0011	0.000	-5.766	0.000	-0.002	-0.001
num_accts_ever_120_pd	0.0010	0.001	0.900	0.368	-0.001	0.003
num_actv_bc_tl	0.0040	0.001	4.791	0.000	0.002	0.006
num_actv_rev_tl	-0.0011	0.001	-1.908	0.056	-0.002	3.13e-05

num_bc_tl	-0.0019	0.000	-5.912	0.000	-0.003	-0.001
num_il_tl	-0.0023	0.000	-9.238	0.000	-0.003	-0.002
num_tl_op_past_12m	0.0088	0.001	10.334	0.000	0.007	0.010
pct_tl_nvr_dlq	-0.0004	0.000	-2.240	0.025	-0.001	-4.43e-05
total_il_high_credit_limit	-6.149e-08	3.43e-08	-1.794	0.073	-1.29e-07	5.69e-09
grade_A	-7.3014	0.003	-2301.683	0.000	-7.308	-7.295
grade_B	-3.5852	0.003	-1383.448	0.000	-3.590	-3.580
grade_C	-0.5565	0.003	-220.365	0.000	-0.561	-0.552
grade_D	2.0939	0.003	774.588	0.000	2.089	2.099
grade_E	4.5038	0.003	1392.196	0.000	4.497	4.510
grade_F	6.3899	0.004	1519.392	0.000	6.382	6.398
grade_G	7.6083	0.009	821.128	0.000	7.590	7.626
home_ownership_MORTGAGE	1.7934	0.014	132.113	0.000	1.767	1.820
home_ownership_NONE	1.8838	0.045	41.648	0.000	1.795	1.972
home_ownership_OTHER	1.8860	0.043	43.442	0.000	1.801	1.971
home_ownership_OWN	1.7989	0.014	131.236	0.000	1.772	1.826
home_ownership_RENT	1.7907	0.014	131.985	0.000	1.764	1.817
Q("loan_status_Charged Off")	4.5726	0.006	770.300	0.000	4.561	4.584
Q("loan_status_Fully Paid")	4.5802	0.006	791.847	0.000	4.569	4.592
purpose_car	0.6736	0.008	84.346	0.000	0.658	0.689
purpose_credit_card	0.6911	0.004	189.130	0.000	0.684	0.698
purpose_debt_consolidation	0.6945	0.003	203.161	0.000	0.688	0.701
purpose_home_improvement	0.6986	0.005	153.373	0.000	0.690	0.708
purpose_house	0.7048	0.010	68.459	0.000	0.685	0.725
purpose_major_purchase	0.6928	0.006	111.409	0.000	0.681	0.705
purpose_medical	0.7309	0.009	82.288	0.000	0.714	0.748
purpose_moving	0.7461	0.011	70.839	0.000	0.725	0.767
purpose_other	0.7220	0.005	153.700	0.000	0.713	0.731
purpose_renewable_energy	0.7316	0.029	24.900	0.000	0.674	0.789
purpose_small_business	0.6846	0.007	98.441	0.000	0.671	0.698
purpose_vacation	0.7035	0.011	62.873	0.000	0.682	0.725
purpose_wedding	0.6786	0.009	72.214	0.000	0.660	0.697
sub_grade_A1	-2.8500	0.005	-580.785	0.000	-2.860	-2.840
sub_grade_A2	-2.2376	0.005	-466.453	0.000	-2.247	-2.228
sub_grade_A3	-1.2779	0.004	-285.755	0.000	-1.287	-1.269
sub_grade_A4	-0.9590	0.004	-246.964	0.000	-0.967	-0.951
sub_grade_A5	0.0230	0.004	6.272	0.000	0.016	0.030
sub_grade_B1	-2.6896	0.003	-860.035	0.000	-2.696	-2.683
sub_grade_B2	-1.6498	0.003	-584.933	0.000	-1.655	-1.644
sub_grade_B3	-0.6404	0.003	-244.339	0.000	-0.646	-0.635
sub_grade_B4	0.2915	0.003	108.260	0.000	0.286	0.297
sub_grade_B5	1.1031	0.003	339.663	0.000	1.097	1.109
sub_grade_C1	-1.5235	0.003	-500.682	0.000	-1.529	-1.517
sub_grade_C2	-0.6986	0.003	-225.551	0.000	-0.705	-0.693
sub_grade_C3	-0.1129	0.003	-35.422	0.000	-0.119	-0.107
sub_grade_C4	0.4429	0.003	137.732	0.000	0.437	0.449
sub_grade_C5	1.3356	0.003	397.229	0.000	1.329	1.342
sub_grade_D1	-0.6970	0.004	-182.936	0.000	-0.704	-0.690
sub_grade_D2	-0.0162	0.004	-3.940	0.000	-0.024	-0.008
sub_grade_D3	0.4094	0.004	93.363	0.000	0.401	0.418
sub_grade_D4	0.8444	0.004	192.044	0.000	0.836	0.853
sub_grade_D5	1.5533	0.005	330.716	0.000	1.544	1.562
sub_grade_E1	-0.1881	0.006	-30.101	0.000	-0.200	-0.176
sub_grade_E2	0.3883	0.006	66.524	0.000	0.377	0.400
sub_grade_E3	0.9169	0.006	143.103	0.000	0.904	0.929
sub_grade_E4	1.4574	0.006	224.804	0.000	1.445	1.470
sub_grade_E5	1.9293	0.007	272.247	0.000	1.915	1.943
sub_grade_F1	0.5513	0.008	67.368	0.000	0.535	0.567
sub_grade_F2	0.9400	0.009	107.013	0.000	0.923	0.957
sub_grade_F3	1.3627	0.009	146.091	0.000	1.344	1.381
sub_grade_F4	1.6889	0.010	163.044	0.000	1.669	1.709
sub_grade_F5	1.8470	0.011	161.679	0.000	1.825	1.869
sub_grade_G1	1.2307	0.017	71.702	0.000	1.197	1.264
sub_grade_G2	1.4191	0.020	71.812	0.000	1.380	1.458
sub_grade_G3	1.6319	0.023	71.798	0.000	1.587	1.676
sub_grade_G4	1.6392	0.029	56.233	0.000	1.582	1.696
sub_grade_G5	1.6874	0.032	52.426	0.000	1.624	1.751

Omnibus:	275804.446	Durbin-Watson:	1.455
Prob(Omnibus):	0.000	Jarque-Bera (JB):	618172593.074
Skew:	-8.225	Prob(JB):	0.00
Kurtosis:	283.302	Cond. No.	1.18e+16

After performing linear regression on the full dataset, the R-squared value is 0.994, which explains about 99% of the proportion of the variance of the interest_rate variable. The variables that had p-values greater than the significance level of .05 are bc_util, mo_sin_old_il_acct, mths_since_recent_bc, open_acc, revol_bal, pub_rec, total_rec_late_fee, total_coll_amt, mo_sin_recnt_rev_tl_op, mo_sin_rcnt_tl, total_il_high_credit_limit. These variables were removed to avoid over-fitting seen in this model.

```
Training Accuracy:   0.9427503365430738
Testing Accuracy:   0.9432286347084489
```

The third model was built after removing the variables on the basis of p-values.

OLS Regression Results

Dep. Variable:	int_rate	R-squared:	0.943
Model:	OLS	Adj. R-squared:	0.943
Method:	Least Squares	F-statistic:	1.106e+05
Date:	Mon, 02 Dec 2019	Prob (F-statistic):	0.00
Time:	11:07:36	Log-Likelihood:	-2.7829e+05
No. Observations:	188181	AIC:	5.566e+05
Df Residuals:	188152	BIC:	5.569e+05
Df Model:	28		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	22.7929	0.167	136.786	0.000	22.466	23.119
loan_amnt	1.258e-05	3.37e-07	37.357	0.000	1.19e-05	1.32e-05
dti	0.0053	0.000	15.697	0.000	0.005	0.006
inq_last_6mths	0.1266	0.003	41.431	0.000	0.121	0.133
pub_rec	7.927e-14	4.68e-15	16.937	0.000	7.01e-14	8.84e-14
bc_util	0.0080	0.000	71.781	0.000	0.008	0.008
num_bc_tl	-0.0094	0.001	-14.480	0.000	-0.011	-0.008
num_tl_op_past_12m	0.0741	0.002	30.774	0.000	0.069	0.079
grade_A	-15.6714	0.016	-987.476	0.000	-15.703	-15.640
grade_B	-11.6443	0.015	-798.690	0.000	-11.673	-11.616
grade_C	-8.2245	0.014	-575.989	0.000	-8.253	-8.197
grade_D	-5.2061	0.015	-356.039	0.000	-5.235	-5.177
grade_E	-2.3630	0.016	-146.803	0.000	-2.395	-2.331
home_ownership_MORTGAGE	-0.2472	0.164	-1.508	0.132	-0.568	0.074
home_ownership_OTHER	-0.0734	0.227	-0.324	0.746	-0.518	0.371
home_ownership_OWN	-0.1138	0.164	-0.693	0.488	-0.435	0.208
home_ownership_RENT	-0.0910	0.164	-0.555	0.579	-0.412	0.230
Q("loan_status_Charged Off")	0.1208	0.007	17.546	0.000	0.107	0.134
purpose_credit_card	-0.0097	0.025	-0.393	0.694	-0.058	0.039
purpose_debt_consolidation	0.0591	0.024	2.420	0.016	0.011	0.107
purpose_home_improvement	0.0720	0.026	2.733	0.006	0.020	0.124
purpose_house	0.0496	0.040	1.233	0.217	-0.029	0.128
purpose_major_purchase	0.0493	0.030	1.656	0.098	-0.009	0.108
purpose_medical	0.2837	0.036	7.792	0.000	0.212	0.355
purpose_moving	0.2981	0.041	7.287	0.000	0.218	0.378
purpose_other	0.3533	0.027	13.238	0.000	0.301	0.406
purpose_renewable_energy	0.2418	0.099	2.440	0.015	0.048	0.436
purpose_small_business	0.2074	0.032	6.563	0.000	0.145	0.269
purpose_vacation	0.3337	0.043	7.810	0.000	0.250	0.417
purpose_wedding	0.1804	0.038	4.772	0.000	0.106	0.255

Omnibus:	13817.691	Durbin-Watson:	1.929
Prob(Omnibus):	0.000	Jarque-Bera (JB):	53230.348
Skew:	-0.293	Prob(JB):	0.00
Kurtosis:	5.539	Cond. No.	1.15e+16

After performing linear regression on this dataset, the R-squared value is 0.943, which explains about 94% of the proportion of the variance of the interest_rate variable. The variables that had p-values greater than the significance level of .05 are home_ownership_MORTGAGE, home_ownership_OWN, home_ownership_OTHER, home_ownership_RENT, purpose_credit_card, purpose_house, purpose_major_purchase.

Classification

Our data set was highly imbalanced because 16% of the observations were charged off and 84% were fully paid, thus we did randomized undersampling. Afterwards, we ended up with a dataset where 50% of the observations were charged off and 50% were fully paid.

We performed the chi-squared test to get the correlation score between dependent variable (loan_status) and all categorical variables, such as home_ownership, term, grade, etc. On the basis of the highest scores, we picked 4 variables that are sub_grade, home_ownership, purpose, term. Using the select k-best method and F_classified method, we picked the 10 best numeric independent variables.

These are the scores for the Chi-squared test: (array([3781.408647 , 382.04154523, 203.78652813, 1489.04599785, 16610.29620058, 304.43478261]), array([0.00000000e+00, 4.47185651e-85, 3.11576294e-46, 0.00000000e+00, 0.00000000e+00, 3.56126919e-68])).

We created three models. One is the random forest with the complete dataset using default parameters. The training accuracy we found for the first model is 84.9% and the test accuracy was 77%, with the F1-score of 77%.

```
=== Confusion Matrix ===
[[6920 2036]
 [2051 6796]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.77      0.77      0.77      8956
           1       0.77      0.77      0.77      8847

    accuracy                           0.77     17803
   macro avg       0.77      0.77      0.77     17803
weighted avg       0.77      0.77      0.77     17803



=== All AUC Scores ===
[0.85322277 0.84096401 0.84318727 0.85340546 0.85268657 0.8391085
 0.85140864 0.84644554 0.86618221 0.84750924]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.8494120213790396
```

The second model is random forest with selected variables using default parameters. When using the selected columns we got the training accuracy to be 85.7% and the testing accuracy to be 77%, however the F1-score remained the same at 77%.

```
=== Confusion Matrix ===
[[6832 2124]
 [1885 6962]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.78      0.76      0.77      8956
           1       0.77      0.79      0.78      8847

    accuracy                           0.77     17803
   macro avg       0.77      0.77      0.77     17803
weighted avg       0.78      0.77      0.77     17803



=== All AUC Scores ===
[0.85980089 0.85142295 0.85345841 0.85283016 0.86135818 0.85134683
 0.85828984 0.85888228 0.87036524 0.85863841]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.8576393182686324
```

The third model is random forest using selected variables with specific parameters. The training accuracy for this model was found to be 88.21%, the testing accuracy was 80%, and the F1-Score at 80%.

```
=== Confusion Matrix ===
[[6389 2567]
 [ 906 7941]]


=== Classification Report ===
              precision    recall  f1-score   support

           0       0.88      0.71      0.79      8956
           1       0.76      0.90      0.82      8847

    accuracy                           0.80     17803
   macro avg       0.82      0.81      0.80     17803
weighted avg       0.82      0.80      0.80     17803



=== All AUC Scores ===
[0.88321193 0.876815   0.87533758 0.8775723  0.88519355 0.87789234
 0.88218542 0.88230857 0.89716215 0.88406711]


=== Mean AUC Score ===
Mean AUC Score - Random Forest:  0.8821745955069854
```

Conclusion

We were able to conclude that the best linear regression model was able to define 94.3% of the variance in the dataset. We didn't go for the second linear regression model with R-Sqaured value to be 99.4% because that model was over-fitting. The final variables for regression were loan_amnt , dti , inq_last_6mths , pub_rec, bc_util , num_bc_tl, num_tl_op_past_12m , grade, home_ownership, loan_status, purpose.

The best classification model had 88.21% training accuracy and the test accuracy to be 80% and the F1-score at 80%. Using these models we will be able to predict the interest rate on a

loan for Lending Club's users and we can also predict the borrowers that will default. The final variables for classification are loan_amnt, int_rate, annual_inc, dti, revol_bal, last_pymnt, total_rev_hi_lim, avg_cur_bal, bc_open_to_buy, bc_util, home_ownership, purpose, term, sub_grade.