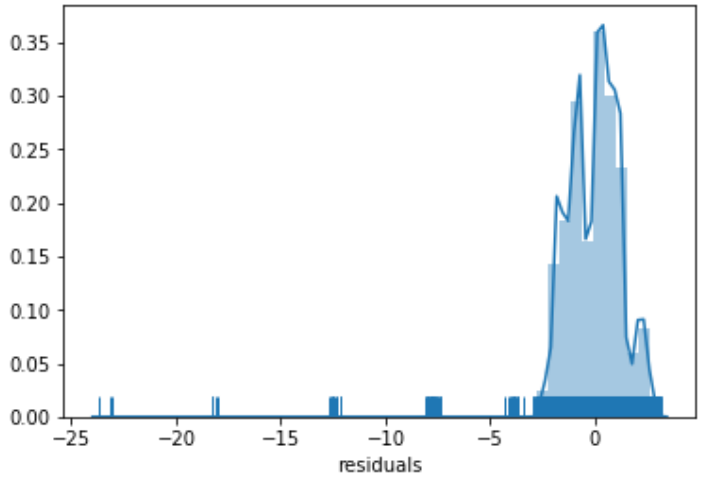
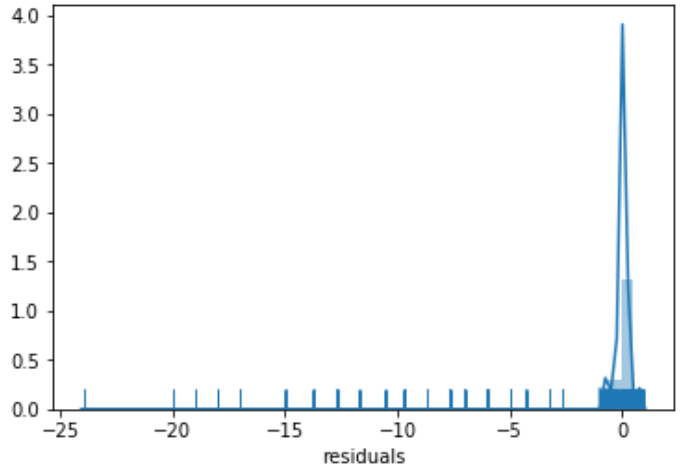
Task 3 – Conclusion and Finalising Best Fit Model

1. **Comparision and conclusion of Residuals distribution.**

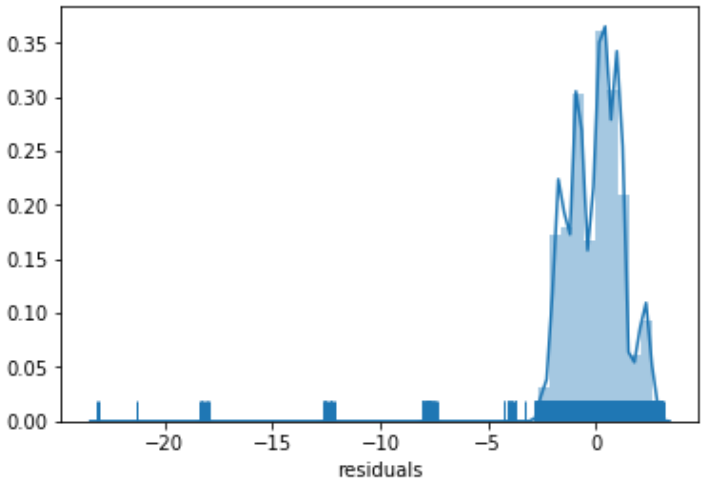
a. **Model1 – Without SUB_GRADE – Multiple Linear Regression**



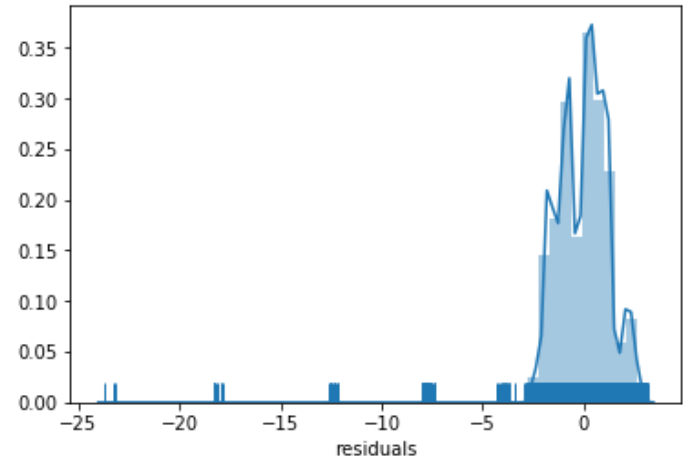
b. **Model1 – Without ADDR_STATE – Multiple Linear Regression**



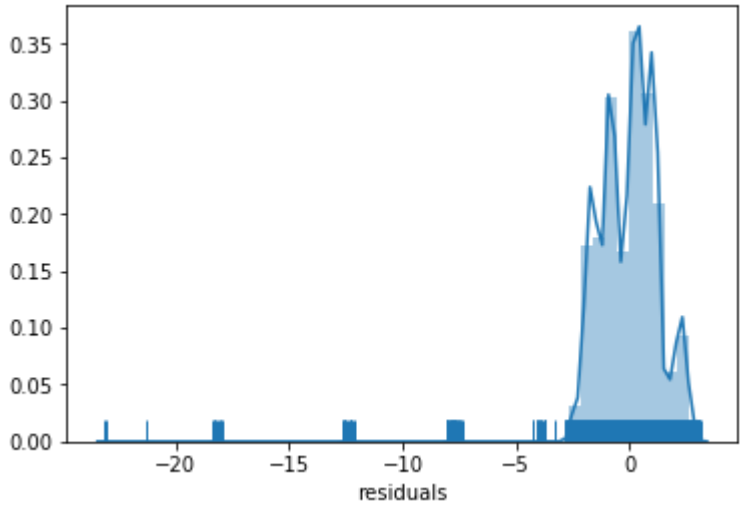
c. **Model2 – Without SUB_GRADE – Gradient Boost Tree Regression**



d. **Model3 – Without SUB_GRADE – Multiple Linear Regression – Polynomial Degree = 2**



e. Model4 – Model 3 With PARAM GRID SEARCH – 12 variations



2. Comparision of R^2 (coefficient of determination).

a. Model1 – Without SUB_GRADE – Multiple Linear Regression - **94.78 %**
b. Model1 – Without ADDR_STATE – Multiple Linear Regression - **99.43 %**
c. Model2 – Without SUB_GRADE – Gradient Boost Tree Regression - **94.86 %**
d. Model3 – Without SUB_GRADE – Multiple Linear Regression – Polynomial Degree = 2 – **94.82 %**
e. Model4 – Model 3 With PARAM GRID SEARCH – 12 variations – **94.65 %**

1. Rug chart distribution shows better model fit if 'addr_state' is not considered.
2. Rug chart shows big variation between different states.
3. R^2 value shows Model1 wihout 'addr_state' feature to be the best.
4. GBT regressor happens to be the best without 'SUB_GRADE' feature.
5. Also the features 'annual_inc' and 'loan_amnt' were transformed with LOG(ln) functions which makes the model prediction accuracy much better.