EuroVis Workshop on Visual Analytics (2018)
C. Tominski and T. von Landesberger (Editors)

Visual Predictive Analytics using iFuseML

Gunjan Sehgal, Mrinal Rawat, Bindu Gupta, Garima Gupta, Geetika Sharma, Gautam Shroff

TCS Research India

Abstract

Solving a predictive analytics problem involves multiple machine learning tasks in a workflow. Directing such workflows efficiently requires an understanding of data so as to identify and handle missing values and outliers, compute feature correlations and to select appropriate models and hyper-parameters. While traditional machine learning techniques are capable of handling these challenges to a certain extent, visual analysis of data and results at each stage can significantly assist in optimal processing of the workflow. We present iFuseML , a visual interactive framework to support analysts in machine learning workflows via insightful data visualizations as well as natural language interfaces where appropriate. Our platform lets the user intuitively search and explore datasets, join relevant datasets using natural language queries, detect and visualize multidimensional outliers and explore feature relationships using Bayesian coordinated views. We also demonstrate how visualization assists in comparing prediction errors to guide ensemble models so as to generate more accurate predictions. We illustrate our framework using a house price dataset from Kaggle, where using iFuseML simplified the machine learning workflow and helped improve the resulting prediction accuracy.

CCS Concepts

•***Computing methodologies*** → *Machine learning, Visual analytics, Predictive analytics, Model ensembles;*

1. Introduction

Growing volumes and large scale availability of data has focused the attention of enterprises on using machine learning models to make better decisions. Predictive analytics using machine learning enables enterprises to identify profitable opportunities and avoid unknown risks, e.g. financial advisors and property investors need to derive valuable insights from data to identify investment opportunities and to predict suitable time to trade.

However, practitioners often need to drill down and analyze data at every stage of the machine learning workflow to fine tune features and parameters so as to produce accurate predictive models. (i) Data cleaning requires analyzing the quality of data to handle its sparsity. The process involves manually detecting outliers through exploratory visual analysis taking a lot of time and effort. (ii) Correlations help to understand important features and suggest useful feature combinations. (iii) Choosing an appropriate machine learning model with suitable hyper parameters needs training and testing different models. (iv) The workflow requires analyses of model predictions to account for model's faulty behavior. (v) Finally, models are compared and combined in ensembles to improve accuracy.

It is observed that each of these stages presents sufficient scope for the use of appropriate visualizations to assist a user while performing the machine learning workflow. We propose a novel visual interactive framework to support predictive analytics, iFuseML. It is out of the box workflow for common learning tasks minimizing user's effort in carrying out the repetitive jobs. The objective of this platform is to let the user's with different levels of ML expertise derive insights, learn models, validate and improve the predictions for all general learning tasks.

The remainder of this paper is organized as follows: We begin with describing and differentiating iFuseML's contribution from the previous work in Section 2, followed by an overview of our platform in Section 3, describing the methodology and approach behind each feature. We next go on to illustrate our visual machine learning workflow for predictive analytics in context of a House price prediction problem in Section 4 and finally conclude in Section 5.

2. Previous Work

Many real-world data analysis problems are intrinsically hard. The analytical power of ML cannot be fully exploited without effective human involvement and exploratory analysis [SSZ*16]. Tools such as rapid miner and konstanz information miner [rap, BCD*09] allow a user to visually assemble and adapt the analysis flow from standardized building blocks, which are then connected through pipes carrying data or models. These platforms require people to be well skilled in predictive analytics as the system design is not intuitive to let the user remember the pipeline components to follow

DOI: 10.2312/eurova.20181106

next. The user is supposed to gather components from all parts of the interface to create a graph like structure to finally execute. Also, the design of the patform does not provide information upfront to intuitively improve the data quality and model at each stage. Frameworks like TFX and KeyStoneML [BBC*17,SVK*17]mostly focus on scaling the components in the production environment or automatic hyper papameter tuning whereas iFuseML is more of insights driven platform. It contributes to visual exploration during data preparation, model generation and post modeling ensembles to improve model. Novel visualization techniques [SJS*17, TLKST09] used for ensembling classifiers exists, but such tools are not available for regression problems to the best of our knowledge. Also, these tools are not end to end machine learning workflows. Such techniques can be integrated to intelligently ensemble in iFuseML as future work.

3. iFuseML Overview

iFuseML provides a visual machine learning framework built on top of a data lake architecture [SPP*16]. It generates proactive data profiles facilitating better decision making. The visual interface design allows easy navigation of the pipeline through visual insights to generate accurate model predictions. It consists of three main panels: Dataset Panel to the right showing the various datasets available in the datalake, left panel allowing data upload and category based search. This could be flipped to ML Panel facilitating step by step machine learning workflow and the bottom panel consisting of the model and the data cart. Data and models added to these carts keeps track of the files being used currently. We now describe the features of the iFuseML platform.

3.1. Looking and Searching

An analyst using iFuseML can upload, search and explore the datasets as shown in Figure 1. In iFuseML each dataset is represented visually as a 'data tile' [SPP*16]. We create an inverted-index of the datasets to enable keyword based serach over the datatiles. iFuseML provides multidimensional data visualizations such as motion charts [BC], parallel coordinates and spatial visualizations as cartograms and bubble maps. All visualizations open in a 'Compare View'facilitating exploration using multiple charts.

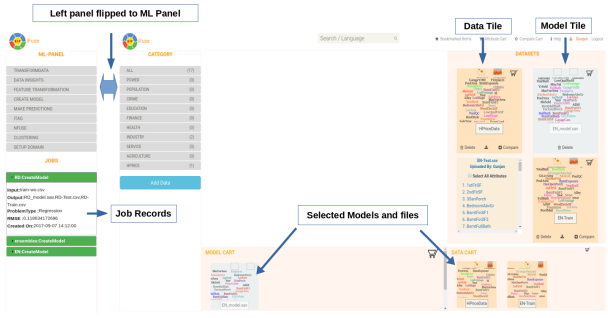

Figure 1: *Main page with category and ML Panels along with data tiles and models added to data and model carts respectively.*

3.2. Linking

User may need to join multiple datasets for analysis. iFuseML provides a NL based component called *nFuse*. It lets users ask questions and perform tasks such as fusion on the datasets.

The data lake architecture lets the user access datasets from different sources to create a new domain. NL is used on top of these domains to extract relevant information. We use semantic web technologies to create the domain ontology (in RDF format) using relational data of the business application. The ontology of the domain describes the terms and their relationships in a 'subject-predicate-object'structure for each of the concepts. This RDF graph is then traversed using the graph traversal functions to get the subject, predicate or object (or a combination of these) to fetch an answer for the given query [GBGA10].

3.3. Finding Insights

The *Data Insights* feature of the ML Panel detects and visualizes top correlated features and multidimensional outliers helpful in understanding samples to be removed [TBR14] and features to be combined in the training data.This is important for getting correct predictions.

Outliers: For every numerical pair of X,Y attributes we perform density based clustering taking Euclidean as distance metric. Cluster quality is determined using silhouette score [Rou87] which returns the mean silhouette coefficient for all the samples between 0 and 1 where values near 0 indicate overlapping clusters. This technique is useful for optimal radius selection. For every sample a count is maintained for the number of times it remained unclassified. This count ranks the exceptional samples and returns top 5 outliers with suitable scatter plots.

Correlations: For every pair of numerical X,Y attributes present in the dataset, iFuseML computes the Pearson's correlation coefficient and returns ten highly correlated features. These correlations are demonstrated using suitable motion chart visualizations. Further, dependencies between categorical attributes are discovered using Bayesian Networks. In iFuseML, user can create these networks using the minimum spanning tree based approach(MST) which can be queried for what-if analysis using bayesian coordinated views as described in section IV.

3.4. Creating Models/Making Predictions

These features in iFuseML facilitates creation of machine learning models including random forest [Bre01], elastic net [ZH05], ridge [McD09], lasso [Tib96], naive bayes and xgboost [CG16]. Intuitive visual analysis of the model predictions guides user in model ensemble. The user specifies training file, target variable and train-validation split as an input. A list or a default set of hyper-parameters is used for training these models. iFuseML determines the best set using grid search algorithm and returns the model with the predictions and errors over both training and validation datasets uploaded as resultant data tiles. The record of each create model job is maintained in the jobs section of the ML Panel. Every job record maintains information of the input and output files, problem type (classification or regression) and the error over the validation

dataset as shown in Figure 1. This view allows easy comparison of the models and also lets the user explore the quality of the predictions through a parallel coordinate view. In the next section, we illustrate the above features of iFuseML in context of the House Price Data Set.

4. Applying iFuseML on House Price Dataset

The house price dataset [hou] describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. This had 2930 observations with sale price available only for 1460 samples. The dataset comprised of 34 numerical, 23 categorical and 23 ordinal features. The goal of the challenge was to create a regression model to accurately predict the price of the property given the features.

4.1. Discovering/Transforming Datasets

We upload the dataset from the left panel of iFuse. As our data lacked spatial attributes, motion charts is a suitable visualization for exploration. To visualize data at different levels of aggregation *Transform Data* tab is used.This gives an essence of the data including sparsity. It was discovered that features like alleys, pool quality and lot frontage were missing for large number of samples. For some features like lot frontage the values were actually missing for others like pool quality and garage condition the missing values depicted the absence of pool and garage respectively, in the house. This information helps in haldling missing values within ifuseML.

Next, the *Data Insights* feature was used to discover the outliers and correlations in the dataset. For every pair of numerical attribute (36 in this case including 2 ordinal features present as numerical value in the raw data) density based clustering was performed over 35 factorial scatter plots to identify the top 5 outliers shown in Figure 2. The top outlier (house id) was discovered exceptional in 229 cases validated through visualizations using our interface shown in Figure 3. Similarly, top correlations were computed for every numerical attribute pair discovering that the living area and total square feet of basement area were highly correlated with the sale price of the house. Likewise, the first and second floor square fit were partially correlated with the target. We combined the former and the latter independently to create new features. Since the year of renovation was correlated with the house price, the feature was transformed to differentiate a new dwelling from an old one, thus recommending feature combinations and transformations.

Correlations for categorical attributes were discovered by creating Bayesian Networks from *Create Model* tab. The user specifies the target variable, network attributes along with the network structure to create a bayesian model visualized using bayesian coordinated views as shown in Fig 4. The bar charts on the diognal elements show the probability distributions of the network attributes, while the remaining are the scatter plots from original data. Query on the probability distributions updates the view showing the prior distributions in red and posterior distributions in red. Here, we queried on higher bins of sale price to find that attributes like saletype, sale condition and lot-config highly affected the sale price. Lastly, the data was normalized and the categorical features transformed to one hot encoding and passed on for model creation.

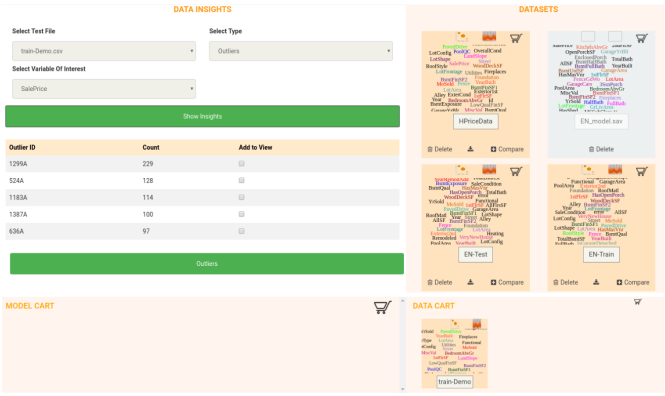

Figure 2: *Data Insights generating top 5 outliers with the counts of x-y scatter plots in which exceptional behavior was encountered.*

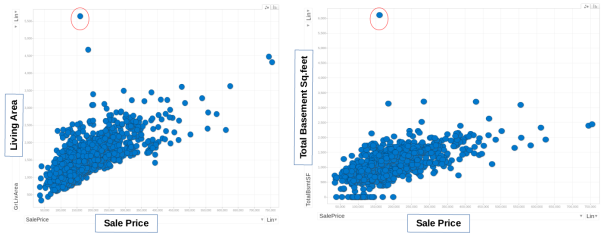

Figure 3: *House identified as the biggest outlier by the Data Insights feature.*

4.2. Training/Comparing Models

We now train different machine learning models using the *Create Model* feature of iFuse. The models performed slightly better when the outliers discovered from iFuseML were removed. Similarly, the features engineered further minimized the errors. The Job records displayed in the bottom part of the ML Panel as shown in Figure 1 shows the accuracies/errors from different algorithms and lets the user compare the results of the various models.

The results of the different machine learning models obtained from the job records are shown in the Table I.

Table 1: *Comparing Different Models*

Model Name	Error in log RMSE on validation data	
	With outliers	Without Outliers
Random Forest	0.141	0.1385
Extra Trees	0.144	0.1411
Elastic Net	0.123	0.1189
Lasso	0.123	0.1188
Ridge	0.130	0.1186
Xgboost	0.131	0.1289

The predictions and the error on the train as well as the validation datasets generated by the create model job can be visualized using

Figure 4: *Querying on higher error bins leads to large changes in distribution for sale condition and basement exposure*

the parallel coordinate view as shown in Figure 5. This lets the user differentiate the badly predicted samples from the rest.

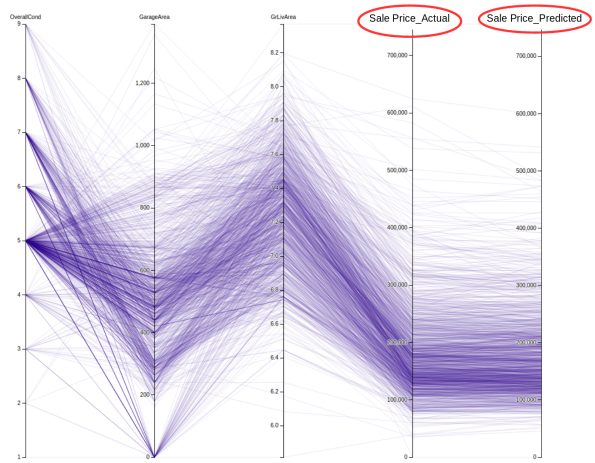

Figure 5: *Actual Vs predicted house price using elastic net model.*

4.3. Ensembling a set of Models

Ensembling different ML models results in better predictions. Diversification introduced through ensemble generates a robust and accurate model reducing chances of variable predictions. Results obtained by training indivisual models are analyzed visually in iFuseML to infer models to ensemble. As predictions of different models are present as different datasets, we join these using the NL component *nFuse* as described in Section II. We add the datasets of indivisual model predictions to the Data Cart and create a new "House Price Domain". We then query this domain to return the errors for all model predictions across each sample. The resultant file is analysed for different pair of model errors as shown in Figure 6(a). On visualizing the absolute error for the top two performing models (lasso and ridge) we discover two sets of non-overlapping samples (Set 1- lower ridge errors and Set 2- lower lasso errors). Similarly two unique sets were identified for ridge and elastic net. This visual analytics approach discovered the complementary behaviour of the two models to preserve both for the ensemble, which otherwise looked interchangeable. Despite average performance, xgboost predicted a set of samples more accurately than any other model as shown in Figure 6(b). The other two models(random forest and extra trees) were discarded.

The Create Model job is then used to ensemble the predictions of the 4 models (ridge, lasso, xgboost, elastic net). The feature uses the predictions from the different models over the train as well as validation dataset to ensemble using xgboost. This approach resulted in log RMSE of 0.1167 on the validation dataset and 0.1169 on test data (results obtained on uploading predictions on Kaggle). We further validated iFuseML motivated 4 model ensemble approach by comparing the results with 3 model (Lasso, Ridge and elastic net)-0.1172 and 2 model ensemble (ridge and elastic net)-0.1170. Hence, iFuseML derived insights proved beneficial in predictive analytics.

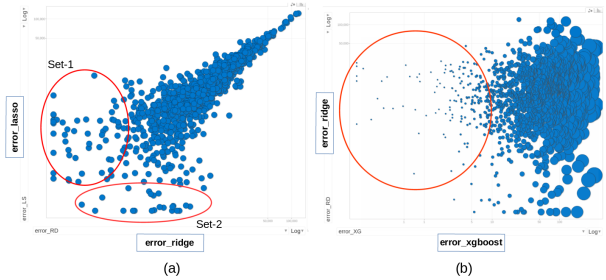

Figure 6: *(a)Absolute errors in predictions for lasso and ridge (b) Absolute errors in predictions for ridge and xgboost, size being mapped to the xgboost error.*

5. Conclusion

We have described iFuseML: a tool which enables and supports machine learning workflows together with appropriate visulizations on data and the results of each stage of the process. In particular we have demonstrated the visual exploration of data, natural language based joins, finding feature relationships and multidimensional outliers, discovering features responsible for faulty model behavior and finally training and ensembling models. All these features were illustrated in context of House Price Dataset. While our framework does not support saving the workflow adopted for a dataset and apply it to other similar datasets automatically, we envisage this as a future direction for enhancing the platform to incorporate AutoML features allowing analysts to quickly apply pre-built workflows and visualizations for greater efficiency.

References

[BBC*17] BAYLOR D., BRECK E., CHENG H.-T., FIEDEL N., FOO C. Y., HAQUE Z., HAYKAL S., ISPIR M., JAIN V., KOC L., ET AL.: Tfx: A tensorflow-based production-scale machine learning platform. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), ACM, pp. 1387–1395. 2

[BC] BATTISTA V., CHENG E.: Motion charts: Telling stories with statistics. 2

[BCD*09] BERTHOLD M. R., CEBRON N., DILL F., GABRIEL T. R., KÖTTER T., MEINL T., OHL P., THIEL K., WISWEDEL B.: Knime-the konstanz information miner: version 2.0 and beyond. *AcM SIGKDD explorations Newsletter 11*, 1 (2009), 26–31. 1

[Bre01] BREIMAN L.: Random forests. *Machine learning 45*, 1 (2001), 5–32. 2

[CG16] CHEN T., GUESTRIN C.: Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), ACM, pp. 785–794. 2

[GBGA10] GOYAL S., BHAT S., GULATI S., ANANTARAM C.: Ontology-driven approach to obtain semantically valid chunks for nl-enabled business applications, 01 2010. 2

[hou] https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data. 3

[McD09] MCDONALD G. C.: Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics 1*, 1 (2009), 93–100. 2

[rap] https://rapidminer.com/. 1

[Rou87] ROUSSEEUW P. J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65. 2

[SJS*17] SCHNEIDER B., JÄCKLE D., STOFFEL F., DIEHL A., FUCHS J., KEIM D.: Visual integration of data and model space in ensemble learning. *arXiv preprint arXiv:1710.07322* (2017). 2

[SPP*16] SINGH K., PANERI K., PANDEY A., GUPTA G., SHARMA G., AGARWAL P., SHROFF G.: Visual bayesian fusion to navigate a data lake. In *FUSION 2016* (2016). 2

[SSZ*16] SACHA D., SEDLMAIR M., ZHANG L., LEE J. A., WEISKOPF D., NORTH S., KEIM D.: Human-centered machine learning through interactive visualization. ESANN. 1

[SVK*17] SPARKS E. R., VENKATARAMAN S., KAFTAN T., FRANKLIN M. J., RECHT B.: Keystoneml: Optimizing pipelines for large-scale advanced analytics. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on* (2017), IEEE, pp. 535–546. 2

[TBR14] TALLÓN-BALLESTEROS A. J., RIQUELME J. C.: Deleting or keeping outliers for classifier training? In *Nature and Biologically Inspired Computing (NaBIC), 2014 Sixth World Congress on* (2014), IEEE, pp. 281–286. 2

[Tib96] TIBSHIRANI R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288. 2

[TLKST09] TALBOT J., LEE B., KAPOOR A., S. TAN D.: Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers, 04 2009. 2

[ZH05] ZOU H., HASTIE T.: Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 2 (2005), 301–320. 2