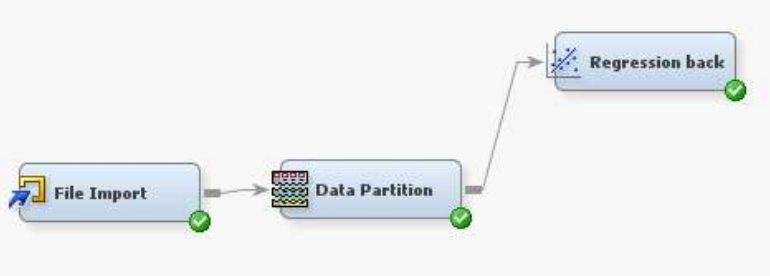
BANKRUPTCY PREDICTION

TABLE OF CONTENTS

Variable Selection

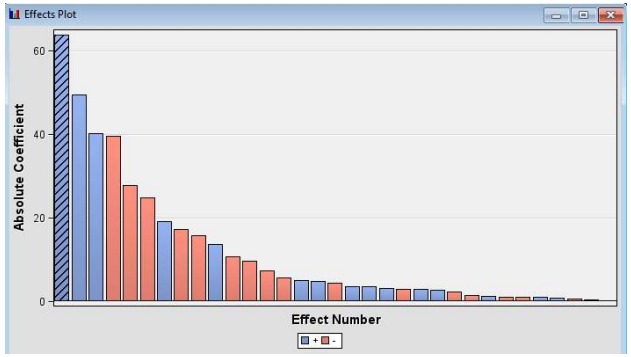
A logistic regression with backward variable selection was run. The variables selected during the process were used in the final model.

Logistic Regression
Data partition: 80% Train, 20% Val, Stratified, Random Seed: 12345.
Regression Node: Backward variable selection. Set selection criterion to misclassification.



Effect	Point Estimate		
Attr1	16.719	Attr35	0.847
Attr10	999.000	Attr36	2.491
Attr11	3.050	Attr38	<0.001
Attr12	0.586	Attr4	30.846
Attr14	<0.001	Attr40	152.944
Attr16	117.398	Attr43	999.000
Attr19	<0.001	Attr46	<0.001
Attr2	999.000	Attr48	15.176
Attr20	0.004	Attr50	0.111
Attr22	0.058	Attr51	0.394
Attr23	999.000	Attr52	<0.001
Attr26	0.014	Attr61	0.391
Attr3	1.957	Attr62	<0.001
Attr32	999.000	Attr63	<0.001
Attr33	31.295	Attr8	1.364
Attr34	23.422	Attr9	0.274

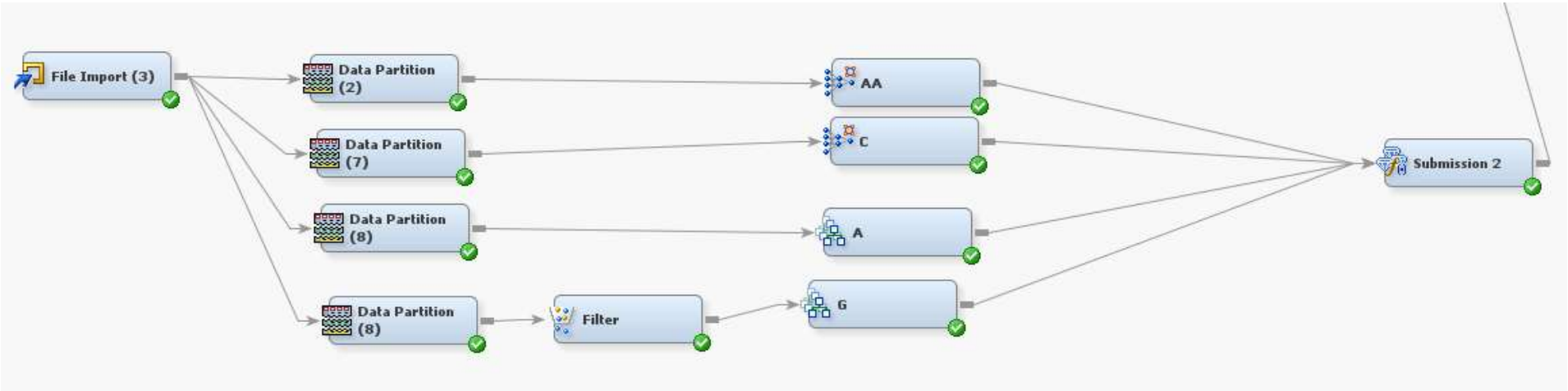
Target	Target Label	Fit Statistics	Statistics Label	Train	Validation
class		_AIC_	Akaike's Inform...	1287.821	.
class		_ASE_	Average Square...	0.018195	0.02096
class		_AVERR_	Average Error F...	0.076383	0.087795
class		_DFE_	Degrees of Free...	7965	.
class		_DFM_	Model Degrees ...	33	.
class		_DFT_	Total Degrees o...	7998	.
class		_DIV_	Divisor for ASE	15996	4004
class		_ERR_	Error Function	1221.821	351.5302
class		_FPE_	Final Prediction ...	0.018346	.
class		_MAX_	Maximum Absol...	0.999965	0.999557
class		_MSE_	Mean Square Er...	0.01827	0.02096
class		_NOBS_	Sum of Frequen...	7998	2002
class		_NW_	Number of Esti...	33	.
class		_RASE_	Root Average S...	0.134889	0.144776
class		_RFPE_	Root Final Predi...	0.135447	.
class		_RMSE_	Root Mean Squ...	0.135168	0.144776
class		_SBC_	Schwarz's Baye...	1518.391	.
class		_SSE_	Sum of Squared...	291.0477	83.92376
class		_SUMW_	Sum of Case W...	15996	4004
class		_MISC_	Misclassificatio...	0.021255	0.023477



Note: Logistic Regression with Forward and Stepwise Variable Selection were also considered

Model Selection and Performance

Two HP neural networks nodes and two gradient boosting nodes were run, and then the average of them was used for the final predictions. The “HP neural networks” and the “Gradient Boosting” nodes were used for this, and an ensemble node was utilized to average the results.



(Individual Model Assessment)

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Valid: Roc Index	Valid: Average Squared Error ▼
	Boost	Boost	A	class		0.021987	0.865	0.020331
	Boost7	Boost7	G	class		0.021708	0.884	0.020092
	HPNNA3	HPNNA3	C	class		0.014563	0.941	0.011697
Y	HPNNA5	HPNNA5	AA	class		0.011425	0.929	0.011123

Model Selection and Performance

(Ensemble Node Model Assessment)

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Valid: Roc Index	Valid: Average Squared Error ▼
Y	Ensmbl2	Ensmbl2	Submissio...	class		0.01371	0.982	0.010182

HPNeuralNet 1

Data partition: 65% Train, 35% Val, Stratified, Random Seed: 1111111. HP **Neural Networks Node:** Create validation (yes), no input standardization, two layers with 10 nodes, maximum iterations (300).

HPNeuralNet 2

Data partition: 65% Train, 35% Val, Stratified, Random Seed: 555555. **HP Neural Networks Node:** Create validation (yes), no input standardization, two layers with 10 nodes, maximum iterations (300).

Gradient Boosting 1

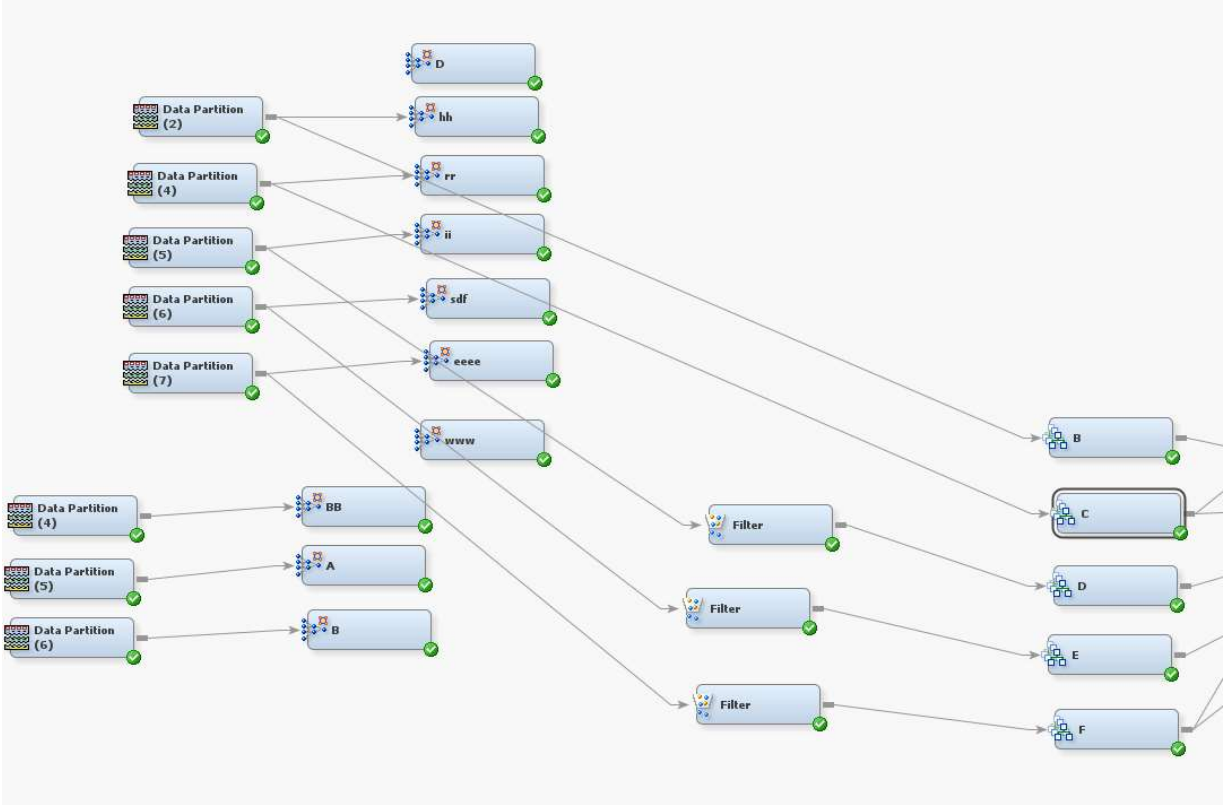
Data partition: 65% Train, 35% Val, Stratified, Random Seed: 66666. **Gradient Boosting Node:** N iterations (100), Seed(56456), Shrinkage (0.05), Train Proportion (80), Max branch (2), Max depth (3), Variable selection (Yes).

Gradient Boosting 2

Data partition: 65% Train, 35% Val, Stratified, Random Seed: 3636555. **Filter Node:** Interval Variable -> Default Filtering Method -> Extreme percentiles. **Gradient Boosting Node:** N iterations (100), Seed(4485), Shrinkage (0.05), Train Proportion (80), Max branch (2), Max depth (3), Variable selection (Yes)

Note: A total number of 12 Neural Networks Nodes and 7 Gradient Boosting Nodes were run before picking the final four model for the submission.

Model Selection and Performance



(Nodes that were not used for the final submission)

Kaggle and Final Results

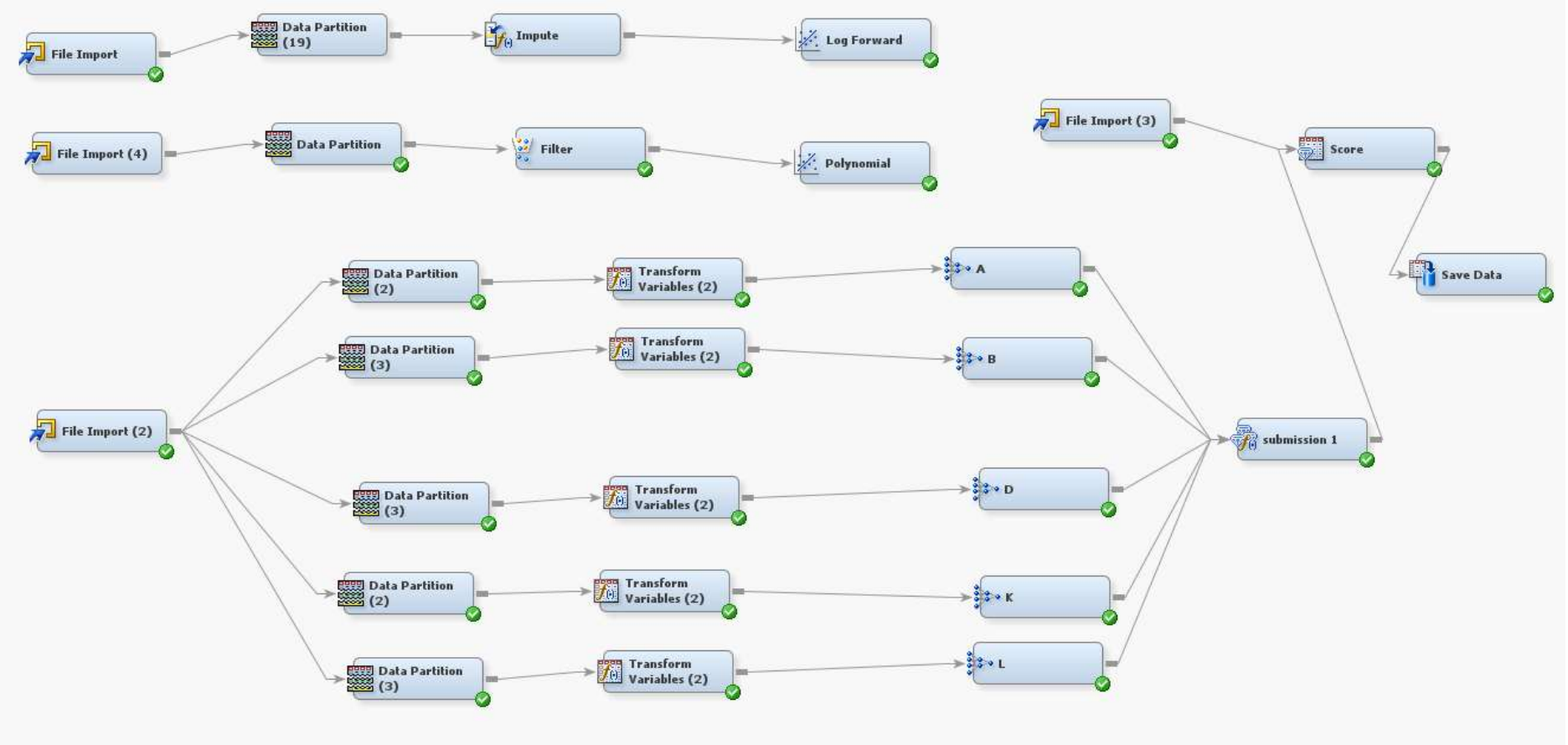
(FINAL SUBMISSION)

	Private Score	Public Score
FinalSubmission2_2HP2Grad.csv 5 days ago by Ricardo Nunez Magana add submission details	0.96668	0.97420

(Public and private scores were fairly similar)

Submission and Description	Private Score	Public Score
bankruptcy_sample_submission.csv 4 days ago by sarnav chauhan add submission details	0.96719	0.97419
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97526	0.97998
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.96868	0.97903
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.96868	0.97903
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97635	0.97981
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97738	0.97837
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97265	0.98087
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	Error	Error
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97070	0.97889
fdfsf.csv 5 days ago by Ricardo Nunez Magana add submission details	0.97670	0.97708

Other Models Considered



Other Models Considered

Five neural network nodes were run, and then the average of them was used for the final predictions. The “neural networks” node was used for this, and an ensemble node was utilized to average th

	Private Score	Public Score
FinalSubmission1NeuralNetAvg5.csv 5 days ago by Ricardo Nunez Magana add submission details	0.94481	0.96592

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Misclassification Rate	Valid: Average Squared Error	Valid: Roc Index
Y	Neural6	Neural6	B	class		0.009663	0.009741	0.945
	Neural5	Neural5	D	class		0.010996	0.009718	0.944
	Neural12	Neural12	K	class		0.013995	0.012405	0.951
	Neural4	Neural4	A	class		0.014662	0.012736	0.938
	Neural14	Neural14	L	class		0.016661	0.014836	0.94

(Model Assessment)

Conclusion

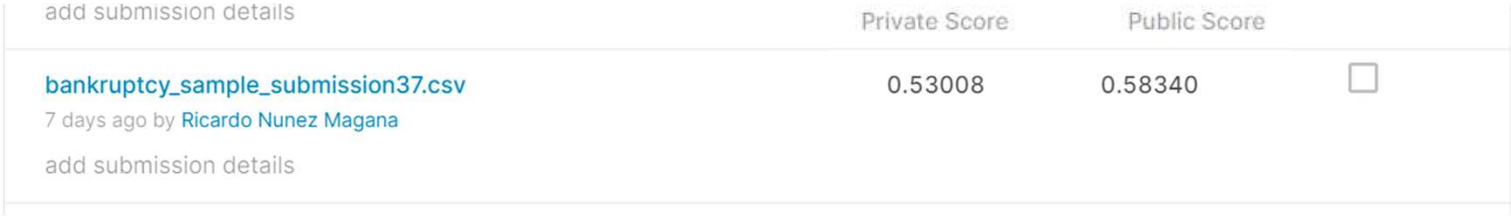
- Using **different data partition** nodes with **unique random seeds** helped in building more robust models in the end. From our experience, that prevented big differences in public and private scores. It also diminished the effects of overfitting models when they were **ensembled**.

-The models that were built only using gradient boosting had a better performance on the private leaderboard.

	Private Score	Public Score	
bankruptcy_sample_submission.csv 6 days ago by Keerthana Nemili add submission details	0.94948	0.92951	☐
bankruptcy_sample_submission104best4gradient.csv 6 days ago by Ricardo Nunez Magana add submission details	0.92608	0.89637	☐
bankruptcy_sample_submission103 best3gradient.csv 6 days ago by Ricardo Nunez Magana add submission details	0.92609	0.89624	☐

Conclusion

-Particularly, for this dataset, Random Forest models did not perform well, even after experimenting with different parameters and settings.



- Overall, the final model was the best performer and had the least FNR amongst models that were considered.