|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| M4 Yearly (N=23,000): OWA | | | | | | | |
| **Method** | **Micro**  **(6538)** | **Industry**  **(3716)** | **Macro**  **(3903)** | **Finance**  **(6519)** | **Demographic (1088)** | **Other (1236)** | **Mean** |
| **RNN Methods** | | | | | | | |
| DeepAR\* | 0.788 | 0.881 | 0.850 | 0.803 | 0.878 | 0.944 | 0.830 |
| DeepState | 0.952 | 0.965 | 0.932 | 0.958 | 0.965 | 0.908 | 0.973 |
| Smyl | 0.744\*\* | 0.833 | 0.796\*\* | 0.756\*\* | 0.811\*\* | 0.819\*\* |  |
| **Benchmark Methods** | | | | | | | |
| Naïve2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |  |
| Comb | 0.809 | 0.934 | 0.922 | 0.848 | 0.867 | 0.894 |  |
| ARIMA | 0.876 | 0.971 | 0.906 | 0.849 | 0.894 | 0.906 |  |
| ETS | 0.891 | 0.961 | 0.898 | 0.882 | 0.884 | 0.930 |  |
| ETSARIMA | 0.859 | 0.941 | 0.879 | 0.842 | 0.862 | 0.897 |  |
| L&K | 0.713 | 0.866 | 0.810 | 0.781 | 0.835 | 0.852 |  |
| \* indicates that DeepAR is used with series-specific dummies.  \*\* Trained on the entire subset of yearly data. Results are subsetted afterwards.  Numbers for DeepAR and DeepState indicate median values for three to ten trials depending on computational complexity.  Total indicates the weighted arithmetic mean. DeepState and DeepAR networks are trained separately on domain subsets of yearly data and may therefore differ compared to results for entire frequency. For local models the weighted average of domain subsets is equal to the frequency average. | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| M4 Quarterly (N=24,000): OWA | | | | | | | |
| **Method** | **Micro**  **(6020)** | **Industry**  **(4637)** | **Macro**  **(5315)** | **Finance**  **(5305)** | **Demographic (1858)** | **Other (865)** | **Mean** |
| **RNN Methods** | | | | | | | |
| DeepAR\* | 0.861 | 0.893 | 0.864 | 0.890 | 0.916 | 0.893 | 0.880 |
| DeepState | 1.028 | 1.046 | 1.203 | 1.029 | 1.106 | 1.011 | 1.076 |
| Smyl | 0.820\*\* | 0.879\*\* | 0.819\*\* | 0.764\*\* | 0.885\*\* | 0.813\*\* |  |
| **Benchmark Methods** | | | | | | | |
| Naïve2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |  |
| Comb | 0.866 | 0.909 | 0.897 | 0.891 | 0.910 | 0.920 |  |
| ARIMA | 0.866 | 0.923 | 0.894 | 0.908 | 0.924 | 0.915 |  |
| ETS | 0.862 | 0.916 | 0.893 | 0.892 | 0.916 | 0.892 |  |
| ETSARIMA | 0.843 | 0.895 | 0.869 | 0.872 | 0.901 | 0.874 |  |
| L&K | 0.882 | 0.907 | 0.907 | 0.897 | 0.901 | 0.923 |  |
| \* indicates that DeepAR is used with series-specific dummies.  \*\* Trained on the entire subset of yearly data. Results are subsetted afterwards.  Numbers for DeepAR and DeepState indicate median values for three to ten trials depending on computational complexity.  Mean indicates the weighted arithmetic mean. DeepState and DeepAR networks are trained separately on domain subsets of yearly data and may therefore differ compared to results for entire frequency. For local models the weighted average of domain subsets is equal to the frequency average. | | | | | | | |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| M4 Monthly (N=48,000): OWA | | | | | | | |
| **Method** | **Micro**  **(10,975)** | **Industry**  **(10,017)** | **Macro**  **(10,016)** | **Finance**  **(10,987)** | **Demographic (5728)** | **Other (277)** | **Mean** |
| **RNN Methods** | | | | | | | |
| DeepAR\* | 0.947 | 0.977 | 0.899 | 0.983 | 0.876 | 0.975 | 0.943 |
| DeepState | 0.945 | 1.057 | 1.027 | 1.122 | 0.949 | 1.020 | 1.040 |
| Smyl | 0.809\*\* | 0.880\*\* | 0.796\*\* | 0.891\*\* | 0.836\*\* | 0.813\*\* |  |
| **Benchmark Methods** | | | | | | | |
| Naïve2 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |  |
| Comb | 0.965 | 0.920 | 0.900 | 0.893 | 0.884 | 0.907 |  |
| ARIMA | 0.923 | 0.924 | 0.886 | 0.916 | 0.891 | 0.885 |  |
| ETS | 0.948 | 0.916 | 0.889 | 0.936 | 0.860 | 0.861 |  |
| ETSARIMA | 0.898 | 0.893 | 0.846 | 0.900 | 0.910 | 0.844 |  |
| L&K | 0.896 | 0.932 | 0.900 | 0.941 | 0.884 | 0.897 |  |
| \* indicates that DeepAR is used with series-specific dummies.  \*\* Trained on the entire subset of yearly data. Results are subsetted afterwards.  Numbers for DeepAR and DeepState indicate median values for three to ten trials depending on computational complexity.  Total indicates the weighted arithmetic mean. DeepState and DeepAR networks are trained separately on domain subsets of yearly data and may therefore differ compared to results for entire frequency. For local models the weighted average of domain subsets is equal to the frequency average. | | | | | | | |